

Chapter 5

Deep Learning

| | | |
|-------|--|-----|
| 5.1 | Supervised learning | 239 |
| 5.1.1 | Nearest neighbors | 241 |
| 5.1.2 | Bayesian classification | 243 |
| 5.1.3 | Logistic regression | 248 |
| 5.1.4 | Support vector machines | 250 |
| 5.1.5 | Decision trees and forests | 254 |
| 5.2 | Unsupervised learning | 257 |
| 5.2.1 | Clustering | 257 |
| 5.2.2 | K-means and Gaussians mixture models | 259 |
| 5.2.3 | Principal component analysis | 262 |
| 5.2.4 | Manifold learning | 265 |
| 5.2.5 | Semi-supervised learning | 266 |
| 5.3 | Deep neural networks | 268 |
| 5.3.1 | Weights and layers | 270 |
| 5.3.2 | Activation functions | 272 |
| 5.3.3 | Regularization and normalization | 274 |
| 5.3.4 | Loss functions | 280 |
| 5.3.5 | Backpropagation | 284 |
| 5.3.6 | Training and optimization | 287 |
| 5.4 | Convolutional neural networks | 291 |
| 5.4.1 | Pooling and unpooling | 295 |
| 5.4.2 | <i>Application:</i> Digit classification | 298 |
| 5.4.3 | Network architectures | 299 |
| 5.4.4 | Model zoos | 304 |
| 5.4.5 | Visualizing weights and activations | 307 |
| 5.4.6 | Adversarial examples | 311 |
| 5.4.7 | Self-supervised learning | 312 |
| 5.5 | More complex models | 317 |
| 5.5.1 | Three-dimensional CNNs | 317 |
| 5.5.2 | Recurrent neural networks | 321 |
| 5.5.3 | Transformers | 322 |
| 5.5.4 | Generative models | 328 |
| 5.6 | Additional reading | 336 |
| 5.7 | Exercises | 337 |

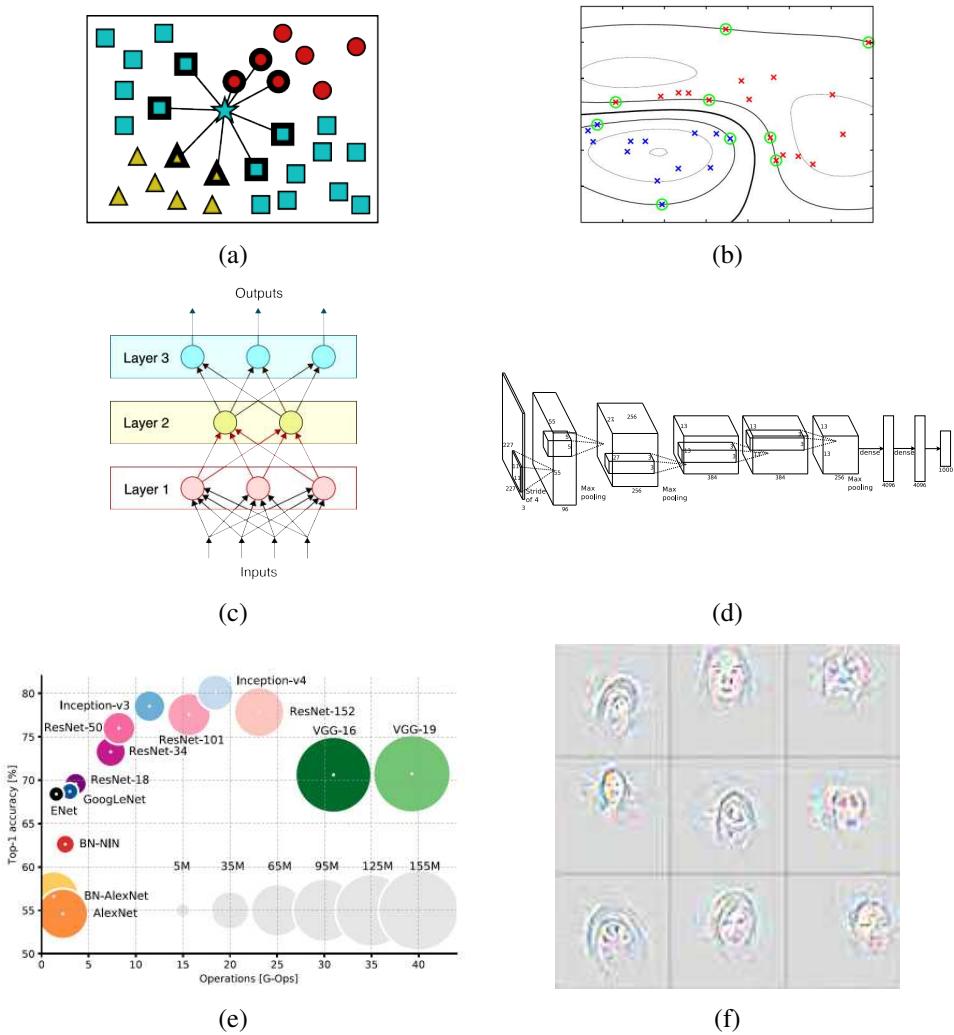


Figure 5.1 Machine learning and deep neural networks: (a) nearest neighbor classification © Glassner (2018); (b) Gaussian kernel support vector machine (Bishop 2006) © 2006 Springer; (c) a simple three-layer network © Glassner (2018); (d) the SuperVision deep neural network, courtesy of Matt Deitke after (Krizhevsky, Sutskever, and Hinton 2012); (e) network accuracy vs. size and operation counts (Canziani, Culurciello, and Paszke 2017) © 2017 IEEE; (f) visualizing network features (Zeiler and Fergus 2014) © 2014 Springer.

Machine learning techniques have always played an important and often central role in the development of computer vision algorithms. Computer vision in the 1970s grew out of the fields of artificial intelligence, digital image processing, and pattern recognition (now called machine learning), and one of the premier journals in our field (*IEEE Transactions on Pattern Analysis and Machine Intelligence*) still bears testament to this heritage.

The image processing, scattered data interpolation, variational energy minimization, and graphical model techniques introduced in the previous two chapters have been essential tools in computer vision over the last five decades. While elements of machine learning and pattern recognition have also been widely used, e.g., for fine-tuning algorithm parameters, they really came into their own with the availability of large-scale labeled image datasets, such as ImageNet (Deng, Dong *et al.* 2009; Russakovsky, Deng *et al.* 2015), COCO (Lin, Maire *et al.* 2014), and LVIS (Gupta, Dollár, and Girshick 2019). Currently, deep neural networks are the most popular and widely used machine learning models in computer vision, not just for semantic classification and segmentation, but even for lower-level tasks such as image enhancement, motion estimation, and depth recovery (Bengio, LeCun, and Hinton 2021).

Figure 5.2 shows the main distinctions between traditional computer vision techniques, in which all of the processing stages were designed by hand, machine learning algorithms, in which hand-crafted features were passed on to a machine learning stage, and deep networks, in which all of the algorithm components, including mid-level representations, are learned directly from the training data.

We begin this chapter with an overview of classical machine learning approaches, such as nearest neighbors, logistic regression, support vector machines, and decision forests. This is a broad and deep subject, and we only provide a brief summary of the main popular approaches. More details on these techniques can be found in textbooks on this subject, which include Bishop (2006), Hastie, Tibshirani, and Friedman (2009), Murphy (2012), Criminisi and Shotton (2013), and Deisenroth, Faisal, and Ong (2020).

The machine learning part of the chapter focuses mostly on *supervised learning* for *classification* tasks, in which we are given a collection of inputs $\{\mathbf{x}_i\}$, which may be features derived from input images, paired with their corresponding class labels (or targets) $\{t_i\}$, which come from a set of classes $\{C_k\}$. Most of the techniques described for supervised classification can easily be extended to *regression*, i.e., associating inputs $\{\mathbf{x}_i\}$ with real-valued scalar or vector outputs $\{\mathbf{y}_i\}$, which we have already studied in Section 4.1. We also look at some examples of *unsupervised learning* (Section 5.2), where there are no labels or outputs, as well as *semi-supervised learning*, in which labels or targets are only provided for a subset of the samples.

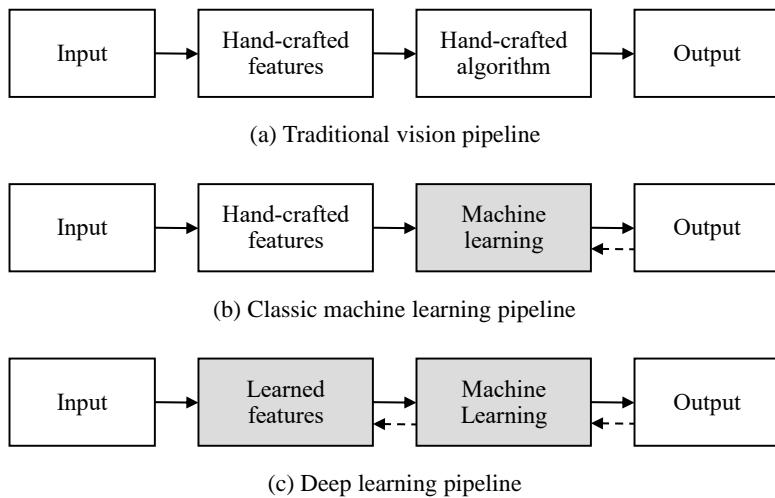


Figure 5.2 Traditional, machine learning, and deep learning pipelines, inspired by Goodfellow, Bengio, and Courville (2016, Figure 1.5). In a classic vision pipeline such as structure from motion, both the features and the algorithm were traditionally designed by hand (although learning techniques could be used, e.g., to design more repeatable features). Classic machine learning approaches take extracted features and use machine learning to build a classifier. Deep learning pipelines learn the whole pipeline, starting from pixels all the way to outputs, using end-to-end training (indicated by the backward dashed arrows) to fine-tune the model parameters.

The second half of this chapter focuses on *deep neural networks*, which, over the last decade, have become the method of choice for most computer vision recognition and lower-level vision tasks. We begin with the elements that make up deep neural networks, including weights and activations, regularization terms, and training using backpropagation and stochastic gradient descents. Next, we introduce convolutional layers, review some of the classic architectures, and talk about how to pre-train networks and visualize their performance. Finally, we briefly touch on more advanced networks, such as three-dimensional and spatio-temporal models, as well as recurrent and generative adversarial networks.

Because machine learning and deep learning are such rich and deep topics, this chapter just briefly summarizes some of the main concepts and techniques. Comprehensive texts on classic machine learning include Bishop (2006), Hastie, Tibshirani, and Friedman (2009), Murphy (2012), and Deisenroth, Faisal, and Ong (2020) while textbooks focusing on deep learning include Goodfellow, Bengio, and Courville (2016), Glassner (2018), Glassner (2021),

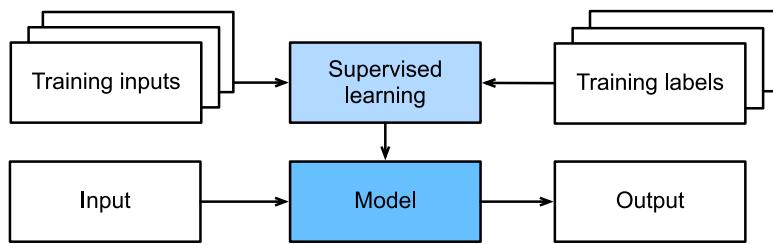


Figure 5.3 In supervised learning, paired training inputs and labels are used to estimate the model parameters that best predict the labels from their corresponding inputs. At run time, the model parameters are (usually) frozen, and the model is applied to new inputs to generate the desired outputs. © Zhang, Lipton et al. (2021, Figure 1.3)

and Zhang, Lipton *et al.* (2021).

5.1 Supervised learning

Machine learning algorithms are usually categorized as either *supervised*, where paired inputs and outputs are given to the learning algorithm (Figure 5.3), or *unsupervised*, where statistical samples are provided without any corresponding labeled outputs (Section 5.2).

As shown in Figure 5.3, supervised learning involves feeding pairs of inputs $\{\mathbf{x}_i\}$ and their corresponding *target* output values $\{t_i\}$ into a learning algorithm, which adjusts the model's parameters so as to maximize the agreement between the model's predictions and the target outputs. The outputs can either be discrete labels that come from a set of classes $\{\mathcal{C}_k\}$, or they can be a set of continuous, potentially vector-valued *values*, which we denote by \mathbf{y}_i to make the distinction between the two cases clearer. The first task is called *classification*, since we are trying to predict class membership, while the second is called *regression*, since historically, fitting a trend to data was called by that name (Section 4.1).¹

After a *training phase* during which all of the *training data* (labeled input-output pairs) have been processed (often by iterating over them many times), the trained model can now be used to predict new output values for previously unseen inputs. This phase is often called the *test phase*, although this sometimes fools people into focusing excessively on performance on a given test set, rather than building a system that works robustly for any plausible inputs that might arise.

¹Note that in software engineering, a *regression* sometimes means a change in the code that results in degraded performance. That is not the kind of regression we will be studying here.

In this section, we focus more on classification, since we've already covered some of the simpler (linear and kernel) methods for regression in the previous chapter. One of the most common applications of classification in computer vision is semantic *image classification*, where we wish to label a complete image (or predetermined portion) with its most likely semantic category, e.g., horse, cat, or car (Section 6.2). This is the main application for which deep networks (Sections 5.3–5.4) were originally developed. More recently, however, such networks have also been applied to continuous pixel labeling tasks such as semantic segmentation, image denoising, and depth and motion estimation. More sophisticated tasks, such as object detection and instance segmentation, will be covered in Chapter 6.

Before we begin our review of traditional supervised learning techniques, we should define a little more formally what the system is trying to learn, i.e., what we meant by “maximize the agreement between the model's predictions and the target outputs.” Ultimately, like any other computer algorithm that will occasionally make mistakes under uncertain, noisy, and/or incomplete data, we would like to maximize its expected utility, or conversely, minimize its expected *loss* or *risk*. This is the subject of *decision theory*, which is explained in more detail in textbooks on machine learning (Bishop 2006, Section 1.5; Hastie, Tibshirani, and Friedman 2009, Section 2.4; Murphy 2012, Section 6.5; Deisenroth, Faisal, and Ong 2020, Section 8.2).

We usually do not have access to the true probability distribution over the inputs, let alone the joint distribution over inputs and corresponding outputs. For this reason, we often use the training data distribution as a proxy for the real-world distribution. This approximation is known as *empirical risk minimization* (see above citations on decision theory), where the expected risk can be estimated with

$$E_{\text{Risk}}(\mathbf{w}) = \frac{1}{N} \sum L(\mathbf{y}_i, \mathbf{f}(\mathbf{x}_i; \mathbf{w})). \quad (5.1)$$

The loss function L measures the “cost” of predicting an output $\mathbf{f}(\mathbf{x}_i; \mathbf{w})$ for input \mathbf{x}_i and model parameters \mathbf{w} when the corresponding target is \mathbf{y}_i .²

This formula should by now be quite familiar, since it is the same one we introduced in the previous chapter (4.2; 4.15) for *regression*. In those cases, the cost (penalty) is a simple quadratic or robust function of the difference between the target output \mathbf{y}_i and the output predicted by the model $f(\mathbf{x}_i; \mathbf{w})$. In some situations, we may want the loss to model specific asymmetries in misprediction. For example, in autonomous navigation, it is usually more costly to over-estimate the distance to the nearest obstacle, potentially resulting in a collision, than to more conservatively under-estimate. We will see more examples of loss functions

²In the machine learning literature, it is more common to write the loss using the letter L . But since we have used the letter E for energy (or summed error) in the previous chapter, we will stick to that notation throughout the book.

later on in this chapter, including Section 5.1.3 on Bayesian classification (5.19–5.24) and Section 5.3.4 on neural network loss (5.54–5.56).

In classification tasks, it is common to minimize the *misclassification rate*, i.e., penalizing all class prediction errors equally using a class-agnostic delta function (Bishop 2006, Sections 1.5.1–1.5.2). However, asymmetries often exist. For example, the cost of producing a *false negative* diagnosis in medicine, which may result in an untreated illness, is often greater than that of a *false positive*, which may suggest further tests. We will discuss true and false positives and negatives, along with error rates, in more detail in Section 7.1.3.

Data preprocessing

Before we start our review of widely used machine learning techniques, we should mention that it is usually a good idea to *center*, *standardize*, and if possible, *whiten* the input data (Glassner 2018, Section 10.5; Bishop 2006, Section 12.1.3). *Centering* the feature vectors means subtracting their mean value, while *standardizing* means also re-scaling each component so that its variance (average squared distance from the mean) is 1.

Whitening is a more computationally expensive process, which involves computing the covariance matrix of the inputs, taking its SVD, and then rotating the coordinate system so that the final dimensions are uncorrelated and have unit variance (under a Gaussian model). While this may be quite practical and helpful for low-dimension inputs, it can become prohibitively expensive for large sets of images. (But see the discussion in Section 5.2.3 on principal component analysis, where it can be feasible and useful.)

With this background in place, we now turn our attention to some widely used supervised learning techniques, namely nearest neighbors, Bayesian classification, logistic regression, support vector machines, and decision trees and forests.

5.1.1 Nearest neighbors

Nearest neighbors is a very simple *non-parametric* technique, i.e., one that does not involve a low-parameter analytic form for the underlying distribution. Instead, the training examples are all retained, and at evaluation time the “nearest” k neighbors are found and then averaged to produce the output.³

Figure 5.4 shows a simple graphical example for various values of k , i.e., from using the $k = 1$ nearest neighbor all the way to finding the $k = 25$ nearest neighbors and selecting

³The reason I put “nearest” in quotations is that standardizing and/or whitening the data will affect distances between vectors, and is usually helpful.

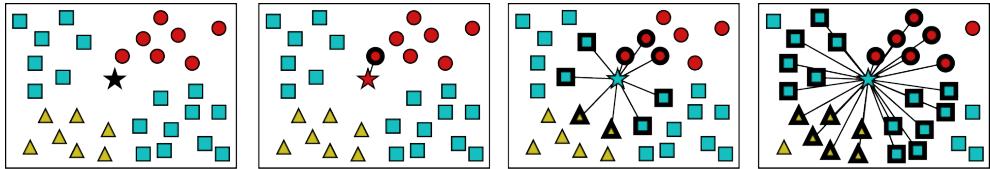


Figure 5.4 Nearest neighbor classification. To determine the class of the star (\star) test sample, we find the k nearest neighbors and select the most popular class. This figure shows the results for $k = 1, 9$, and 25 samples. © Glassner (2018)

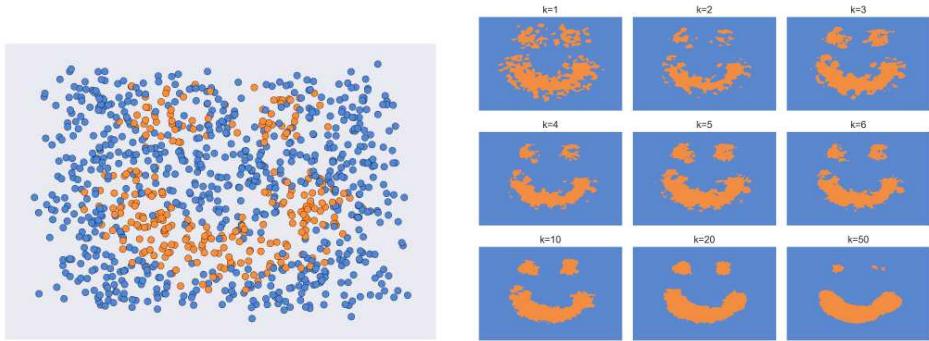


Figure 5.5 For noisy (intermingled) data, selecting too small a value of k results in irregular decision surfaces. Selecting too large a value can cause small regions to shrink or disappear. © Glassner (2018)

the class with the highest count as the output label. As you can see, changing the number of neighbors affects the final class label, which changes from red to blue.

Figure 5.5 shows the effect of varying the number of neighbors in another way. The left half of the figure shows the initial samples, which fall into either blue or orange categories. As you can see, the training samples are highly *intermingled*, i.e., there is no clear (plausible) boundary that will correctly label all of the samples. The right side of this figure shows the *decision boundaries* for a k -NN classifier as we vary the values of k from 1 to 50. When k is too small, the classifier acts in a very random way, i.e., it is *overfitting* to the training data (Section 4.1.2). As k gets larger, the classifier *underfits* (over-smooths) the data, resulting in the shrinkage of the two smaller regions. The optimal number of nearest neighbors to use k is a *hyperparameter* for this algorithm. Techniques for determining a good value include cross-validation, which we discussed in Section 4.1.2.

While nearest neighbors is a rather brute-force machine learning technique (although

Cover and Hart (1967) showed that it is statistically optimal in the large sample limit), but it can still be useful in many computer vision applications, such as large-scale matching and indexing (Section 7.1.4). As the number of samples gets large, however, efficient techniques must be used to find the (exact or approximate) nearest neighbors. Good algorithms for finding nearest neighbors have been developed in both the general computer science and more specialized computer vision communities.

Muja and Lowe (2014) developed a Fast Library for Approximate Nearest Neighbors (FLANN), which collects a number of previously developed algorithms and is incorporated as part of OpenCV. The library implements several powerful approximate nearest neighbor algorithms, including randomized k-d trees (Silpa-Anan and Hartley 2008), priority search k-means trees, approximate nearest neighbors (Friedman, Bentley, and Finkel 1977), and locality sensitive hashing (LSH) (Andoni and Indyk 2006). Their library can empirically determine which algorithm and parameters to use based on the characteristics of the data being indexed.

More recently, Johnson, Douze, and Jégou (2021) developed the GPU-enabled Faiss library⁴ for scaling similarity search (Section 6.2.3) to billions of vectors. The library is based on product quantization (Jégou, Douze, and Schmid 2010), which had been shown by the authors to perform better than LSH (Gordo, Perronnin *et al.* 2013) on the kinds of large-scale datasets the Faiss library was developed for.

5.1.2 Bayesian classification

For some simple machine learning problems, e.g., if we have an analytic model of feature construction and noising, or if we can gather enough samples, we can determine the probability distributions of the feature vectors for each class $p(\mathbf{x}|\mathcal{C}_k)$ as well as the prior class likelihoods $p(\mathcal{C}_k)$.⁵ According to Bayes' rule (4.33), the likelihood of class \mathcal{C}_k given a feature vector \mathbf{x} (Figure 5.6) is given by

$$p_k = p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{\sum_j p(\mathbf{x}|\mathcal{C}_j)p(\mathcal{C}_j)} \quad (5.2)$$

$$= \frac{\exp l_k}{\sum_j \exp l_j}, \quad (5.3)$$

⁴<https://github.com/facebookresearch/faiss>

⁵The following notation and equations are adapted from Bishop (2006, Section 4.2), which describes *probabilistic generative classification*.

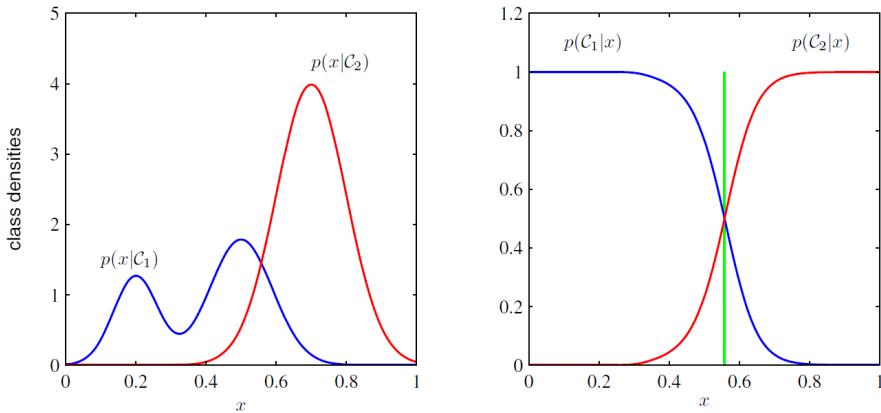


Figure 5.6 An example with two class conditional densities $p(x|\mathcal{C}_k)$ along with the corresponding posterior class probabilities $p(\mathcal{C}_k|x)$, which can be obtained using Bayes' rule, i.e., by dividing by the sum of the two curves (Bishop 2006) © 2006 Springer. The vertical green line is the optimal decision boundary for minimizing the misclassification rate.

where the second form (using the \exp functions) is known as the *normalized exponential* or *softmax* function.⁶ The quantity

$$l_k = \log p(\mathbf{x}|\mathcal{C}_k) + \log p(\mathcal{C}_k) \quad (5.4)$$

is the *log-likelihood* of sample \mathbf{x} being from class \mathcal{C}_k .⁷ It is sometimes convenient to denote the softmax function (5.3) as a vector-to-vector valued function,

$$\mathbf{p} = \text{softmax}(\mathbf{l}). \quad (5.5)$$

The softmax function can be viewed as a soft version of a maximum indicator function, which returns 1 for the largest value of l_k whenever it dominates the other values. It is widely used in machine learning and statistics, including its frequent use as the final non-linearity in deep neural classification networks (Figure 5.27).

The process of using formula (5.2) to determine the likelihood of a class \mathcal{C}_k given a feature vector \mathbf{x} is known as *Bayesian classification*, since it combines a conditional feature likelihood $p(\mathbf{x}|\mathcal{C}_k)$ with a prior distribution over classes $p(\mathcal{C}_k)$ using Bayes' rule to determine

⁶For better numerical stability, it is common to subtract the largest value of l_j from all of the input values so that the exponentials are in the range $(0, 1]$ and there is less chance of roundoff error.

⁷Some authors (e.g., Zhang, Lipton *et al.* 2021) use the term *logit* for the log-likelihood, although it is more commonly used to denote the *log odds*, discussed below, or the softmax function itself.

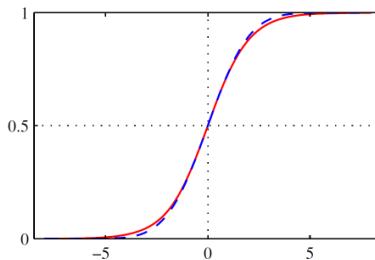


Figure 5.7 The logistic sigmoid function $\sigma(l)$, shown in red, along with a scaled error function, shown in dashed blue (Bishop 2006) © 2006 Springer.

the posterior class probabilities. In the case where the components of the feature vector are generated independently, i.e.,

$$p(\mathbf{x}|\mathcal{C}_k) = \prod_i p(x_i|\mathcal{C}_k), \quad (5.6)$$

the resulting technique is called a *naïve Bayes classifier*.

For the binary (two class) classification task, we can re-write (5.3) as

$$p(\mathcal{C}_0|\mathbf{x}) = \frac{1}{1 + \exp(-l)} = \sigma(l), \quad (5.7)$$

where $l = l_0 - l_1$ is the difference between the two class log likelihood and is known as the *log odds* or *logit*.

The $\sigma(l)$ function is called the *logistic sigmoid function* (or simply the *logistic function* or *sigmoid*) means an S-shaped curve (Figure 5.7). The sigmoid was a popular *activation function* in earlier neural networks, although it has now been replaced by functions, as discussed in Section 5.3.2.

Linear and quadratic discriminant analysis

While probabilistic generative classification based on the normalized exponential and sigmoid can be applied to any set of log likelihoods, the formulas become much simpler when the distributions are multi-dimensional Gaussians.

For Gaussians with identical covariance matrices Σ , we have

$$p(\mathbf{x}|\mathcal{C}_k) = \frac{1}{(2\pi)^{D/2}} \frac{1}{\|\Sigma\|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \right\} \quad (5.8)$$

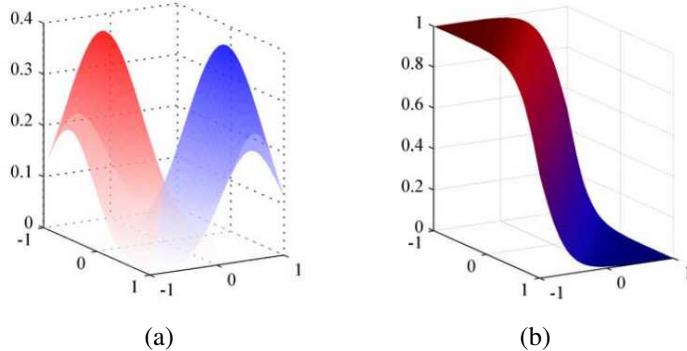


Figure 5.8 Logistic regression for two identically distributed Gaussian classes (Bishop 2006) © 2006 Springer: (a) two Gaussian distributions shown in red and blue; (b) the posterior probability $p(\mathcal{C}_0|\mathbf{x})$, shown as both the height of the function and the proportion of red ink.

In the case of two classes (binary classification), we obtain (Bishop 2006, Section 4.2.1)

$$p(\mathcal{C}_0|\mathbf{x}) = \sigma(\mathbf{w}^T \mathbf{x} + b), \quad (5.9)$$

with

$$\mathbf{w} = \Sigma^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1), \quad \text{and} \quad (5.10)$$

$$b = \frac{1}{2}\boldsymbol{\mu}_0^T \Sigma^{-1} \boldsymbol{\mu}_0 + \frac{1}{2}\boldsymbol{\mu}_1^T \Sigma^{-1} \boldsymbol{\mu}_1 + \log \frac{p(\mathcal{C}_0)}{p(\mathcal{C}_1)}. \quad (5.11)$$

Equation (5.9), which we will revisit shortly in the context of non-generative (discriminative) classification (5.18), is called *logistic regression*, since we pass the output of a linear regression formula

$$l(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b \quad (5.12)$$

through the logistic function to obtain a class probability. Figure 5.8 illustrates this in two dimensions, where the posterior likelihood of the red class $p(\mathcal{C}_0|\mathbf{x})$ is shown on the right side.

In linear regression (5.12), \mathbf{w} plays the role of the *weight* vector along which we project the feature vector \mathbf{x} , and b plays the role of the *bias*, which determines where to set the classification boundary. Note that the weight direction (5.10) aligns with the vector joining the distribution means (after rotating the coordinates by the inverse covariance Σ^{-1}), while the bias term is proportional to the mean squared moments and the log class prior ratio $\log(p(\mathcal{C}_0)/p(\mathcal{C}_1))$.

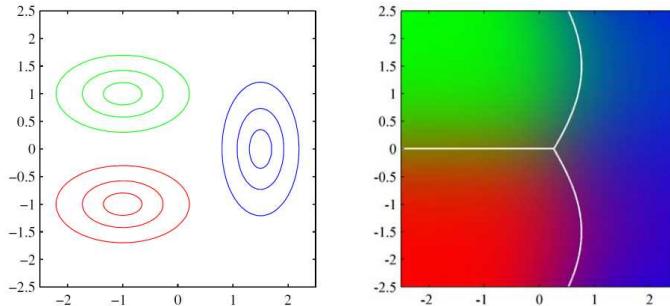


Figure 5.9 Quadratic discriminant analysis (Bishop 2006) © 2006 Springer. When the class covariances Σ_k are different, the decision surfaces between Gaussian distributions become quadratic surfaces.

For $K > 2$ classes, the softmax function (5.3) can be applied to the linear regression log likelihoods,

$$l_k(\mathbf{x}) = \mathbf{w}_k^T \mathbf{x} + b_k, \quad (5.13)$$

with

$$\mathbf{w}_k = \Sigma^{-1} \boldsymbol{\mu}_k, \quad \text{and} \quad (5.14)$$

$$b_k = -\frac{1}{2} \boldsymbol{\mu}_k^T \Sigma^{-1} \boldsymbol{\mu}_k + \log p(\mathcal{C}_k). \quad (5.15)$$

Because the decision boundaries along which the classification switches from one class to another are linear,

$$\mathbf{w}_k \mathbf{x} + b_k > \mathbf{w}_l \mathbf{x} + b_l, \quad (5.16)$$

the technique of classifying examples using such criteria is known as *linear discriminant analysis* (Bishop 2006, Section 4.1; Murphy 2012, Section 4.2.2).⁸

Thus far, we have looked at the case where all of the class covariance matrices Σ_k are identical. When they vary between classes, the decision surfaces are no longer linear and they become quadratic (Figure 5.9). The derivation of these quadratic decision surfaces is known as *quadratic discriminant analysis* (Murphy 2012, Section 4.2.1).

In the case where Gaussian class distributions are not available, we can still find the best discriminant direction using Fisher discriminant analysis (Bishop 2006, Section 4.1.4; Murphy 2012, Section 8.6.3), as shown in Figure 5.10. Such analysis can be useful in separately

⁸The acronym LDA is commonly used with linear discriminant analysis, but is sometimes also used for *latent Dirichlet allocation* in graphical models.

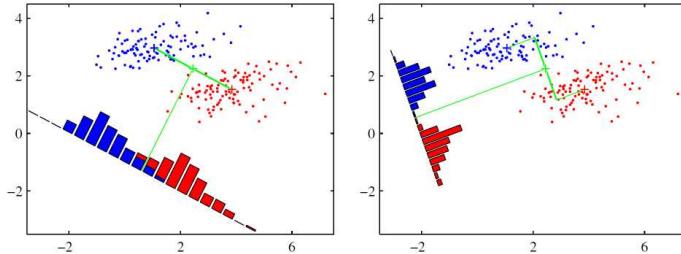


Figure 5.10 Fisher linear discriminant (Bishop 2006) © 2006 Springer. To find the projection direction to best separate two classes, we compute the sum of the two class covariances and then use its inverse to rotate the vector between the two class means.

modeling variability within different classes, e.g., the appearance variation of different people (Section 5.2.3).

5.1.3 Logistic regression

In the previous section, we derived classification rules based on posterior probabilities applied to multivariate Gaussian distributions. Quite often, however, Gaussians are not appropriate models of our class distributions and we must resort to alternative techniques.

One of the simplest among these is *logistic regression*, which applies the same ideas as in the previous section, i.e., a linear projection onto a weight vector,

$$l_i = \mathbf{w} \cdot \mathbf{x}_i + b \quad (5.17)$$

followed by a logistic function

$$p_i = p(\mathcal{C}_0 | \mathbf{x}_i) = \sigma(l_i) = \sigma(\mathbf{w}^T \mathbf{x}_i + b) \quad (5.18)$$

to obtain (binary) class probabilities. Logistic regression is a simple example of a *discriminative model*, since it does not construct or assume a prior distribution over unknowns, and hence is not *generative*, i.e., we cannot generate random samples from the class (Bishop 2006, Section 1.5.4).

As we no longer have analytic estimates for the class means and covariances (or they are poor models of the class distributions), we need some other method to determine the weights \mathbf{w} and bias b . We do this by maximizing the posterior log likelihoods of the correct labels.

For the binary classification task, let $t_i \in \{0, 1\}$ be the class label for each training sample \mathbf{x}_i and $p_i = p(\mathcal{C}_0 | \mathbf{x})$ be the estimated likelihood predicted by (5.18) for a given

weight and bias (\mathbf{w}, b) . We can maximize the likelihood of the correct labels being predicted by minimizing the negative log likelihood, i.e., the *cross-entropy loss* or error function,

$$E_{\text{CE}}(\mathbf{w}, b) = - \sum_i \{t_i \log p_i + (1 - t_i) \log(1 - p_i)\} \quad (5.19)$$

(Bishop 2006, Section 4.3.2).⁹ Note how whenever the label $t_i = 0$, we want $p_i = p(\mathcal{C}_0 | \mathbf{x}_i)$ to be high, and vice versa.

This formula can easily be extended to a multi-class loss by again defining the posterior probabilities as normalized exponentials over per-class linear regressions, as in (5.3) and (5.13),

$$p_{ik} = p(\mathcal{C}_k | \mathbf{x}_i) = \frac{\exp l_{ik}}{\sum_j \exp l_{ij}} = \frac{1}{Z_i} \exp l_{ik}, \quad (5.20)$$

with

$$l_{ik} = \mathbf{w}_k^T \mathbf{x}_i + b_k. \quad (5.21)$$

The term $Z_i = \sum_j \exp l_{ij}$ can be a useful shorthand in derivations and is sometimes called the *partition function*. After some manipulation (Bishop 2006, Section 4.3.4), the corresponding *multi-class cross-entropy loss* (a.k.a. *multinomial logistic regression objective*) becomes

$$E_{\text{MCCE}}(\{\mathbf{w}_k, b_k\}) = - \sum_i \sum_k \tilde{t}_{ik} \log p_{ik}, \quad (5.22)$$

where the 1-of-K (or *one-hot*) encoding has $\tilde{t}_{ik} = 1$ if sample i belongs to class k (and 0 otherwise).¹⁰ It is more common to simply use the integer class value t_i as the target, in which case we can re-write this even more succinctly as

$$E(\{\mathbf{w}_k, b_k\}) = - \sum_i \log p_{it_i}, \quad (5.23)$$

i.e., we simply sum up the log likelihoods of the correct class for each training sample. Substituting the softmax formula (5.20) into this loss, we can re-write it as

$$E(\{\mathbf{w}_k, b_k\}) = \sum_i (\log Z_i - l_{it_i}). \quad (5.24)$$

⁹Note, however, that since this derivation is based on the assumption of Gaussian noise, it may not perform well if there are outliers, e.g., errors in the labels. In such a case, a more robust measure such as mean absolute error (MAE) may be preferable (Ghosh, Kumar, and Sastry 2017) or it may be necessary to re-weight the training samples (Ren, Zeng *et al.* 2018).

¹⁰This kind of representation can be useful if we wish the target classes to be a mixture, e.g., in the *mixup* data augmentation technique of Zhang, Cisse *et al.* (2018).

To determine the best set of weights and biases, $\{\mathbf{w}_k, b_k\}$, we can use gradient descent, i.e., update their values using a Newton-Raphson second-order optimization scheme (Bishop 2006, Section 4.3.3),

$$\mathbf{w} \leftarrow \mathbf{w} - \mathbf{H}^{-1} \nabla E(\mathbf{w}), \quad (5.25)$$

where ∇E is the gradient of the loss function E with respect to the weight variables \mathbf{w} , and \mathbf{H} is the Hessian matrix of second derivatives of E . Because the cross-entropy functions are not linear in the unknown weights, we need to iteratively solve this equation a few times to arrive at a good solution. Since the elements in \mathbf{H} are updated after each iteration, this technique is also known as *iteratively reweighted least squares*, which we will study in more detail in Section 8.1.4. While many non-linear optimization problems have multiple local minima, the cross-entropy functions described in this section do not, so we are guaranteed to arrive at a unique solution.

Logistic regression does have some limitations, which is why it is often used for only the simplest classification tasks. If the classes in feature space are not linearly separable, using simple projections onto weight vectors may not produce adequate decision surfaces. In this case, *kernel methods* (Sections 4.1.1 and 5.1.4; Bishop 2006, Chapter 6; Murphy 2012, Chapter 14), which measure the distances between new (test) feature vectors and select training examples, can often provide good solutions.

Another problem with logistic regression is that if the classes actually are separable (either in the original feature space, or the lifted kernel space), there can be more than a single unique separating plane, as illustrated in Figure 5.11a. Furthermore, unless regularized, the weights \mathbf{w} will continue to grow larger, as larger values of \mathbf{w}_k lead to larger p_{ik} values (once a separating plane has been found) and hence a smaller overall loss.

For this reason, techniques that place the decision surfaces in a way that maximizes their separation to labeled examples have been developed, as we discuss next.

5.1.4 Support vector machines

As we have just mentioned, in some applications of logistic regression we cannot determine a single optimal decision surface (choice of weight and bias vectors $\{\mathbf{w}_k, b_k\}$ in (5.21)) because there are *gaps* in the feature space where any number of planes could be introduced. Consider Figure 5.11a, where the two classes are denoted in cyan and magenta colors. In addition to the two dashed lines and the solid line, there are infinitely many other lines that will also cleanly separate the two classes, including a swath of horizontal lines. Since the classification error for any of these lines is zero, how can we choose the best decision surface, keeping in mind that we only have a limited number of training examples, and that actual run-time examples

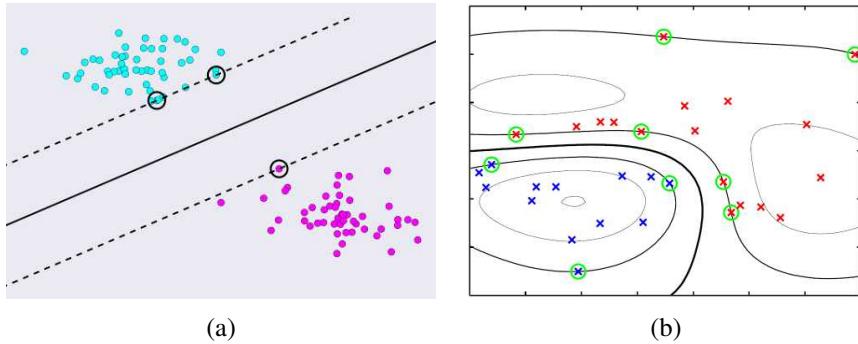


Figure 5.11 (a) A support vector machine (SVM) finds the linear decision surface (hyperplane) that maximizes the margin to the nearest training examples, which are called the support vectors © Glassner (2018). (b) A two-dimensional two class example of a Gaussian kernel support vector machine (Bishop 2006) © 2006 Springer. The red and blue \times s indicate the training samples, and the samples circled in green are the support vectors. The black lines indicate iso-contours of the kernel regression function, with the contours containing the blue and red support vectors indicating the ± 1 contours and the dark contour in between being the decision surface.

may fall somewhere in between?

The answer to this problem is to use *maximum margin classifiers* (Bishop 2006, Section 7.1), as shown in Figure 5.11a, where the dashed lines indicate two parallel decision surfaces that have the *maximum margin*, i.e., the largest perpendicular distance between them. The solid line, which represents the hyperplane half-way between the dashed hyperplanes, is the maximum margin classifier.

Why is this a good idea? There are several potential derivations (Bishop 2006, Section 7.1), but a fairly intuitive explanation is that there may be real-world examples coming from the cyan and magenta classes that we have not yet seen. Under certain assumptions, the maximum margin classifier provides our best bet for correctly classifying as many of these unseen examples as possible.

To determine the maximum margin classifier, we need to find a weight-bias pair (\mathbf{w}, b) for which all regression values $l_i = \mathbf{w} \cdot \mathbf{x}_i + b$ (5.17) have an absolute value of at least 1 as well as the correct sign. To denote this more compactly, let

$$\hat{t}_i = 2t_i - 1, \quad \hat{t}_i \in \{-1, 1\} \quad (5.26)$$

be the *signed class label*. We can now re-write the inequality condition as

$$\hat{t}_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1. \quad (5.27)$$

To maximize the margin, we simply find the smallest norm weight vector \mathbf{w} that satisfies (5.27), i.e., we solve the optimization problem

$$\arg \min_{\mathbf{w}, b} \|\mathbf{w}\|^2 \quad (5.28)$$

subject to (5.27). This is a classic *quadratic programming problem*, which can be solved using the method of Lagrange multipliers, as described in Bishop (2006, Section 7.1).

The inequality constraints are exactly satisfied, i.e., they turn into equalities, along the two dashed lines in Figure 5.11a, where we have $l_i = \mathbf{w} \cdot \mathbf{x}_i + b = \pm 1$. The circled points that touch the dashed lines are called the *support vectors*.¹¹ For a simple linear classifier, which can be denoted with a single weight and bias pair (\mathbf{w}, b) , there is no real advantage to computing the support vectors, except that they help us estimate the decision surface. However, as we will shortly see, when we apply kernel regression, having a small number of support vectors is a huge advantage.

What happens if the two classes are not *linearly separable*, and in fact require a complex curved surface to correctly classify samples, as in Figure 5.11b? In this case, we can replace linear regression with *kernel regression* (4.3), which we introduced in Section 4.1.1. Rather than multiplying the weight vector \mathbf{w} with the feature vector \mathbf{x} , we instead multiply it with the value of K kernel functions centered at the data point locations \mathbf{x}_k ,

$$l_i = f(\mathbf{x}_i; \mathbf{w}, b) = \sum_k w_k \phi(\|\mathbf{x}_i - \mathbf{x}_k\|) + b. \quad (5.29)$$

This is where the power of support vector machines truly comes in.

Instead of requiring the summation over all training samples \mathbf{x}_k , once we solve for the maximum margin classifier only a small subset of support vectors needs to be retained, as shown by the circled crosses in Figure 5.11b. As you can see in this figure, the decision boundary denoted by the dark black line nicely separates the red and blue class samples. Note that as with other applications of kernel regression, the width of the radial basis functions is still a free hyperparameter that must be reasonably tuned to avoid underfitting and overfitting.

¹¹While the cyan and magenta dots may just look like points, they are, of course, schematic representations of higher-dimensional *vectors* lying in feature space.

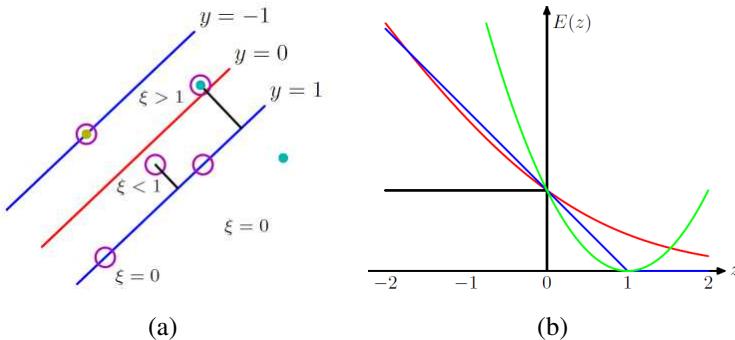


Figure 5.12 Support vector machine for overlapping class distributions (Bishop 2006) © 2006 Springer. (a) The green circled point is on the wrong side of the $y = 1$ decision contour and has a penalty of $\xi = 1 - y > 0$. (b) The “hinge” loss used in support vector machines is shown in blue, along with a rescaled version of the logistic regression loss function, shown in red, the misclassification error in black, and the squared error in green.

Hinge loss. So far, we have focused on classification problems that are separable, i.e., for which a decision boundary exists that correctly classifies all the training examples. Support vector machines can also be applied to overlapping (mixed) class distributions (Figure 5.12a), which we previously approached using logistic regression. In this case, we replace the inequality conditions (5.27), i.e., $\hat{t}_i l_i \geq 1$, with a *hinge loss* penalty

$$E_{\text{HL}}(l_i, \hat{t}_i) = [1 - \hat{t}_i l_i]_+, \quad (5.30)$$

where $[x]_+$ denotes the positive part, i.e. $[x]_+ = \max(0, x)$. The hinge loss penalty, shown in blue in Figure 5.12b, is 0 whenever the (previous) inequality is satisfied and ramps up linearly depending on how much the inequality is violated. To find the optimal weight values (\mathbf{w}, b) , we minimize the regularized sum of hinge loss values,

$$E_{\text{SV}}(\mathbf{w}, b) = \sum_i E_{\text{HL}}(l_i(\mathbf{x}_i; \mathbf{w}, b), \hat{t}_i) + \lambda \|\mathbf{w}\|^2. \quad (5.31)$$

Figure 5.12b compares the hinge loss to the logistic regression (cross-entropy) loss in (5.19). The hinge loss imposes no penalty on training samples that are on the correct side of the $|l_i| > 1$ boundary, whereas the cross-entropy loss prefers larger absolute values. While, in this section, we have focused on the two-class version of support vector machines, Bishop (2006, Chapter 7) describes the extension to multiple classes as well as efficient optimization algorithms such as *sequential minimal optimization* (SMO) (Platt 1989). There’s also a nice

online tutorial on the scikit-learn website.¹² A survey of SVMs and other kernel methods applied to computer vision can be found in Lampert (2008).

5.1.5 Decision trees and forests

In contrast to most of the supervised learning techniques we have studied so far in this chapter, which process complete feature vectors all at once (with either linear projections or distances to training examples), *decision trees* perform a sequence of simpler operations, often just looking at individual feature elements before deciding which element to look at next (Hastie, Tibshirani, and Friedman 2009, Chapter 17; Glassner 2018, Section 14.5; Criminisi, Shotton, and Konukoglu 2012; Criminisi and Shotton 2013). (Note that the *boosting approaches* we study in Section 6.3.1 also use similar simple *decision stumps*.) While decision trees have been used in statistical machine learning for several decades (Breiman, Friedman *et al.* 1984), the application of their more powerful extension, namely *decision forests*, only started gaining traction in computer vision a little over a decade ago (Lepetit and Fua 2006; Shotton, Johnson, and Cipolla 2008; Shotton, Girshick *et al.* 2013). Decision trees, like support vector machines, are discriminative classifiers (or regressors), since they never explicitly form a probabilistic (generative) model of the data they are classifying.

Figure 5.13 illustrates the basic concepts behind decision trees and random forests. In this example, training samples come from four different classes, each shown in a different color (a). A decision tree (b) is constructed top-to-bottom by selecting decisions at each node that split the training samples that have made it to that node into more specific (lower entropy) distributions. The thickness of each link shows the number of samples that get classified along that path, and the color of the link is the blend of the class colors that flow through that link. The color histograms show the class distributions at a few of the interior nodes.

A random forest (c) is created by building a set of decision trees, each of which makes slightly different decisions. At test (classification) time, a new sample is classified by each of the trees in the random forest, and the class distributions at the final leaf nodes are averaged to provide an answer that is more accurate than could be obtained with a single tree (with a given depth).

Random forests have several design parameters, which can be used to tailor their accuracy, generalization, and run-time and space complexity. These parameters include:

- the depth of each tree D ,
- the number of trees T , and

¹²<https://scikit-learn.org/stable/modules/svm.html#svm-classification>

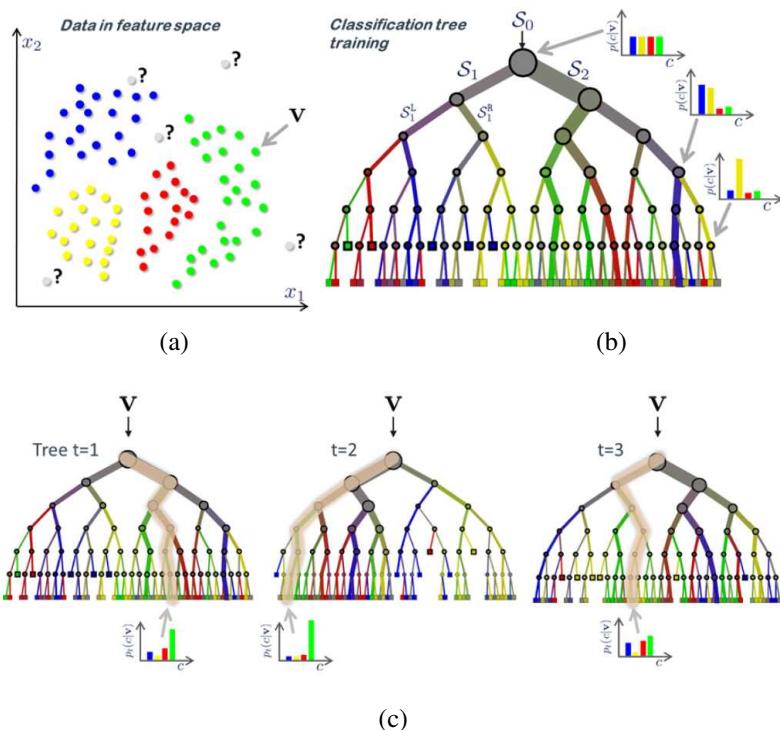


Figure 5.13 Decision trees and forests (Criminisi and Shotton 2013) © 2013 Springer. The top left figure (a) shows a set of training samples tags with four different class colors. The top right (b) shows a single decision tree with a distribution of classes at each node (the root node has the same distribution as the entire training set). During testing (c), each new example (feature vector) is tested at the root node, and depending on this test result (e.g., the comparison of some element to a threshold), a decision is made to walk down the tree to one of its children. This continues until a leaf node with a particular class distribution is reached. During training (b), decisions are selected such that they reduce the entropy (increase class specificity) at the node's children. The bottom diagram (c) shows an ensemble of three trees. After a particular test example has been classified by each tree, the class distributions of the leaf nodes of all the constituent trees are averaged.

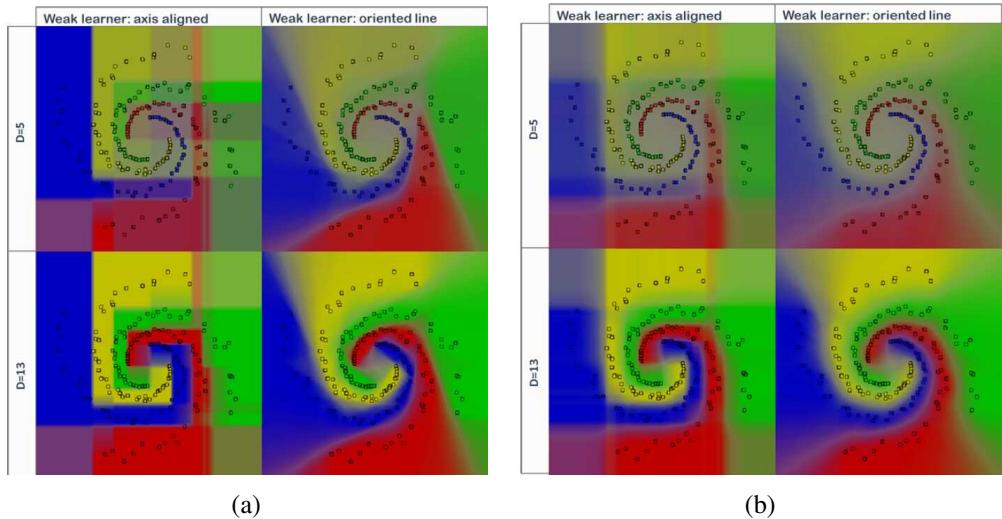


Figure 5.14 Random forest decision surfaces (Criminisi and Shotton 2013) © 2013 Springer. Figures (a) and (b) show smaller and larger amounts of “noise” between the $T = 400$ tree forests obtained by using $\rho = 500$ and $\rho = 5$ random hypotheses at each split node. Within each figure, the two rows show trees of different depths ($D = 5$ and 13), while the columns show the effects of using axis-aligned or linear decision surfaces (“weak learners”).

- the number of samples examined at node construction time ρ .

By only looking at a random subset ρ of all the training examples, each tree ends up having different decision functions at each node, so that the ensemble of trees can be averaged to produce softer decision boundaries.

Figure 5.14 shows the effects of some of these parameters on a simple four-class two-dimensional spiral dataset. In this figure, the number of trees has been fixed to $T = 400$. Criminisi and Shotton (2013, Chapter 4) have additional figures showing the effect of varying more parameters. The left (a) and right (b) halves of this figure show the effects of having less randomness ($\rho = 500$) and more randomness ($\rho = 5$) at the decision nodes. Less random trees produce sharper decision surfaces but may not generalize as well. Within each 2×2 grid of images, the top row shows a shallower $D = 5$ tree, while the bottom row shows a deeper $D = 13$ tree, which leads to finer details in the decision boundary. (As with all machine learning, better performance on training data may not lead to better generalization because of overfitting.) Finally, the right column shows what happens if axis-aligned (single element) decisions are replaced with linear combinations of feature elements.

When applied to computer vision, decision trees first made an impact in keypoint recognition (Lepetit and Fua 2006) and image segmentation (Shotton, Johnson, and Cipolla 2008). They were one of the key ingredients (along with massive amounts of synthetic training data) in the breakthrough success of human pose estimation from Kinect depth images (Shotton, Girshick *et al.* 2013). They also led to state-of-the-art medical image segmentation systems (Criminisi, Robertson *et al.* 2013), although these have now been supplanted by deep neural networks (Kamnitsas, Ferrante *et al.* 2016). Most of these applications, along with additional ones, are reviewed in the book edited by Criminisi and Shotton (2013).

5.2 Unsupervised learning

Thus far in this chapter, we have focused on *supervised learning* techniques where we are given training data consisting of paired input and target examples. In some applications, however, we are only given a set of data, which we wish to characterize, e.g., to see if there are any patterns, regularities, or typical distributions. This is typically the realm of classical statistics. In the machine learning community, this scenario is usually called *unsupervised learning*, since the sample data comes without labels. Examples of applications in computer vision include image segmentation (Section 7.5) and face and body recognition and reconstruction (Sections 13.6.2).

In this section, we look at some of the more widely used techniques in computer vision, namely clustering and mixture modeling (e.g., for segmentation) and principal component analysis (for appearance and shape modeling). Many other techniques are available, and are covered in textbooks on machine learning, such as Bishop (2006, Chapter 9), Hastie, Tibshirani, and Friedman (2009, Chapter 14), and Murphy (2012, Section 1.3).

5.2.1 Clustering

One of the simplest things you can do with your sample data is to group it into sets based on similarities (e.g., vector distances). In statistics, this problem is known as *cluster analysis* and is a widely studied area with hundreds of different algorithms (Jain and Dubes 1988; Kaufman and Rousseeuw 1990; Jain, Duin, and Mao 2000; Jain, Topchy *et al.* 2004). Murphy (2012, Chapter 25) has a nice exposition on clustering algorithms, including affinity propagation, spectral clustering, graph Laplacian, hierarchical, agglomerative, and divisive clustering. The survey by Xu and Wunsch (2005) is even more comprehensive, covering almost 300 different papers and such topics as similarity measures, vector quantization, mixture modeling, kernel methods, combinatorial and neural network algorithms, and visualization. Figure 5.15 shows

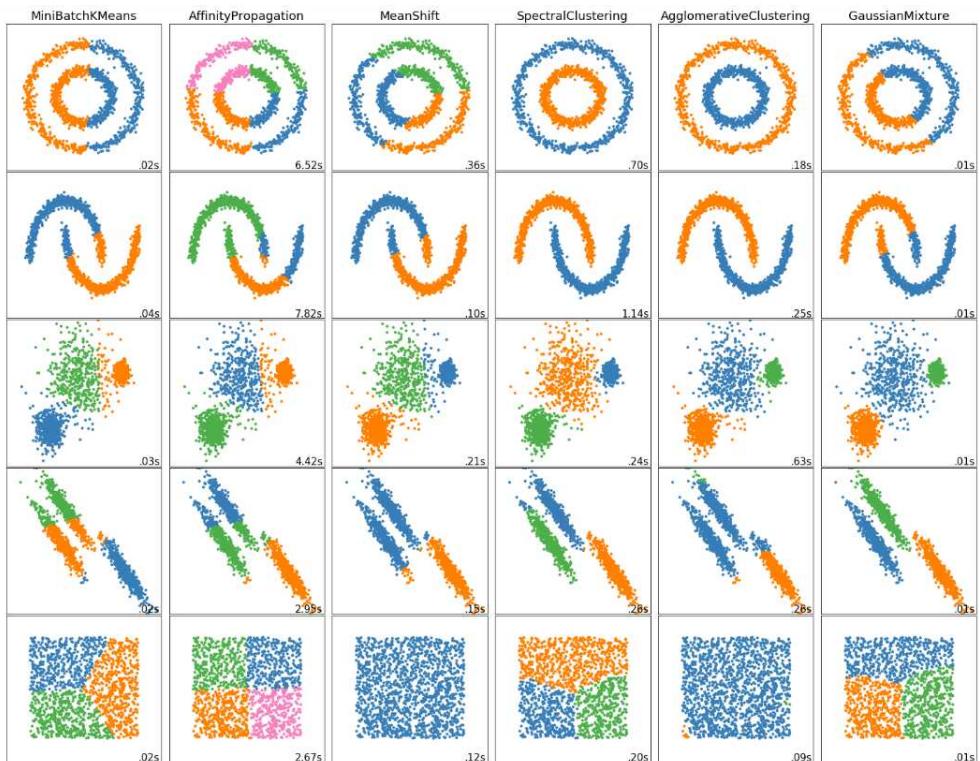


Figure 5.15 Comparison of different clustering algorithms on some toy datasets, generated using a simplified version of https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html#sphx-glr-auto-examples-cluster-plot-cluster-comparison-py.

some of the algorithms implemented in the <https://scikit-learn.org> cluster analysis package applied to some simple two-dimensional examples.

Splitting an image into successively finer regions (divisive clustering) is one of the oldest techniques in computer vision. Ohlander, Price, and Reddy (1978) present such a technique, which first computes a histogram for the whole image and then finds a threshold that best separates the large peaks in the histogram. This process is repeated until regions are either fairly uniform or below a certain size. More recent splitting algorithms often optimize some metric of intra-region similarity and inter-region dissimilarity. These are covered in Sections 7.5.3 and 4.3.2.

Region merging techniques also date back to the beginnings of computer vision. Brice and Fennema (1970) use a dual grid for representing boundaries between pixels and merge

regions based on their relative boundary lengths and the strength of the visible edges at these boundaries.

In data clustering, algorithms can link clusters together based on the distance between their closest points (single-link clustering), their farthest points (complete-link clustering), or something in between (Jain, Topchy *et al.* 2004). Kamvar, Klein, and Manning (2002) provide a probabilistic interpretation of these algorithms and show how additional models can be incorporated within this framework. Applications of such agglomerative clustering (region merging) algorithms to image segmentation are discussed in Section 7.5.

Mean-shift (Section 7.5.2) and mode finding techniques, such as k-means and mixtures of Gaussians, model the feature vectors associated with each pixel (e.g., color and position) as samples from an unknown probability density function and then try to find clusters (modes) in this distribution.

Consider the color image shown in Figure 7.53a. How would you segment this image based on color alone? Figure 7.53b shows the distribution of pixels in $L^*u^*v^*$ space, which is equivalent to what a vision algorithm that ignores spatial location would see. To make the visualization simpler, let us only consider the L^*u^* coordinates, as shown in Figure 7.53c. How many obvious (elongated) clusters do you see? How would you go about finding these clusters?

The k-means and mixtures of Gaussians techniques use a *parametric* model of the density function to answer this question, i.e., they assume the density is the superposition of a small number of simpler distributions (e.g., Gaussians) whose locations (centers) and shape (covariance) can be estimated. Mean shift, on the other hand, smoothes the distribution and finds its peaks as well as the regions of feature space that correspond to each peak. Since a complete density is being modeled, this approach is called *non-parametric* (Bishop 2006).

5.2.2 K-means and Gaussians mixture models

K-means implicitly model the probability density as a superposition of spherically symmetric distributions and does not require any probabilistic reasoning or modeling (Bishop 2006). Instead, the algorithm is given the number of clusters k it is supposed to find and is initialized by randomly sampling k centers from the input feature vectors. It then iteratively updates the cluster center location based on the samples that are closest to each center (Figure 5.16). Techniques have also been developed for splitting or merging cluster centers based on their statistics, and for accelerating the process of finding the nearest mean center (Bishop 2006).

In mixtures of Gaussians, each cluster center is augmented by a covariance matrix whose values are re-estimated from the corresponding samples (Figure 5.17). Instead of using nearest neighbors to associate input samples with cluster centers, a *Mahalanobis distance* (Ap-

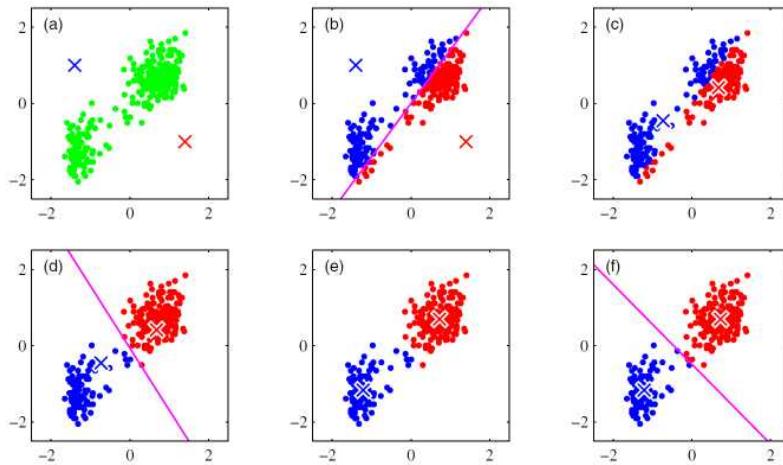


Figure 5.16 The k -means algorithm starts with a set of samples and the number of desired clusters (in this case, $k = 2$) (Bishop 2006) © 2006 Springer. It iteratively assigns samples to the nearest mean, and then re-computes the mean center until convergence.

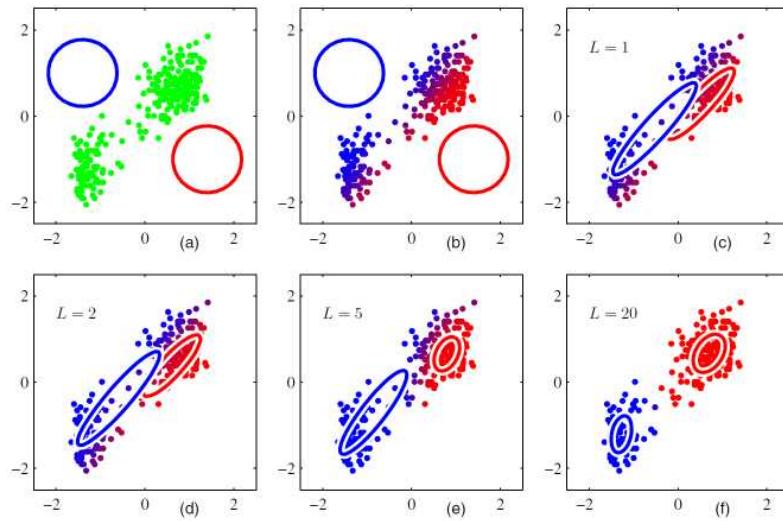


Figure 5.17 Gaussian mixture modeling (GMM) using expectation maximization (EM) (Bishop 2006) © 2006 Springer. Samples are softly assigned to cluster centers based on their Mahalanobis distance (inverse covariance weighted distance), and the new means and covariances are recomputed based on these weighted assignments.

pendix B.1) is used:

$$d(\mathbf{x}_i, \boldsymbol{\mu}_k; \boldsymbol{\Sigma}_k) = \|\mathbf{x}_i - \boldsymbol{\mu}_k\|_{\boldsymbol{\Sigma}_k^{-1}} = (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \quad (5.32)$$

where \mathbf{x}_i are the input samples, $\boldsymbol{\mu}_k$ are the cluster centers, and $\boldsymbol{\Sigma}_k$ are their covariance estimates. Samples can be associated with the nearest cluster center (a *hard assignment* of membership) or can be *softly assigned* to several nearby clusters.

This latter, more commonly used, approach corresponds to iteratively re-estimating the parameters for a Gaussians mixture model,

$$p(\mathbf{x} | \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}) = \sum_k \pi_k \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad (5.33)$$

where π_k are the *mixing coefficients*, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the Gaussian means and covariances, and

$$\mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{|\boldsymbol{\Sigma}_k|} e^{-d(\mathbf{x}, \boldsymbol{\mu}_k; \boldsymbol{\Sigma}_k)} \quad (5.34)$$

is the *normal* (Gaussian) distribution (Bishop 2006).

To iteratively compute (a local) maximum likely estimate for the unknown mixture parameters $\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$, the *expectation maximization* (EM) algorithm (Shlezinger 1968; Dempster, Laird, and Rubin 1977) proceeds in two alternating stages:

1. The *expectation* stage (E step) estimates the *responsibilities*

$$z_{ik} = \frac{1}{Z_i} \pi_k \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad \text{with} \quad \sum_k z_{ik} = 1, \quad (5.35)$$

which are the estimates of how likely a sample \mathbf{x}_i was generated from the k th Gaussian cluster.

2. The *maximization* stage (M step) updates the parameter values

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_i z_{ik} \mathbf{x}_i, \quad (5.36)$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_i z_{ik} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T, \quad (5.37)$$

$$\pi_k = \frac{N_k}{N}, \quad (5.38)$$

where

$$N_k = \sum_i z_{ik}. \quad (5.39)$$

is an estimate of the number of sample points assigned to each cluster.

Bishop (2006) has a wonderful exposition of both mixture of Gaussians estimation and the more general topic of expectation maximization.

In the context of image segmentation, Ma, Derksen *et al.* (2007) present a nice review of segmentation using mixtures of Gaussians and develop their own extension based on Minimum Description Length (MDL) coding, which they show produces good results on the Berkeley segmentation dataset.

5.2.3 Principal component analysis

As we just saw in mixture analysis, modeling the samples within a cluster with a multivariate Gaussian can be a powerful way to capture their distribution. Unfortunately, as the dimensionality of our sample space increases, estimating the full covariance quickly becomes infeasible.

Consider, for example, the space of all frontal faces (Figure 5.18). For an image consisting of P pixels, the covariance matrix has a size of $P \times P$. Fortunately, the full covariance normally does not have to be modeled, since a lower-rank approximation can be estimated using *principal component analysis*, as described in Appendix A.1.2.

PCA was originally used in computer vision for modeling faces, i.e., *eigenfaces*, initially for gray-scale images (Kirby and Sirovich 1990; Turk and Pentland 1991), and then for 3D models (Blanz and Vetter 1999; Egger, Smith *et al.* 2020) (Section 13.6.2) and active appearance models (Section 6.2.4), where they were also used to model facial shape deformations (Rowland and Perrett 1995; Cootes, Edwards, and Taylor 2001; Matthews, Xiao, and Baker 2007).

Eigenfaces. Eigenfaces rely on the observation first made by Kirby and Sirovich (1990) that an arbitrary face image \mathbf{x} can be compressed and reconstructed by starting with a mean image \mathbf{m} (Figure 6.1b) and adding a small number of scaled signed images \mathbf{u}_i ,

$$\tilde{\mathbf{x}} = \mathbf{m} + \sum_{i=0}^{M-1} a_i \mathbf{u}_i, \quad (5.40)$$

where the signed basis images (Figure 5.18b) can be derived from an ensemble of training images using *principal component analysis* (also known as *eigenvalue analysis* or the *Karhunen–Loève transform*). Turk and Pentland (1991) recognized that the coefficients a_i in the eigenface expansion could themselves be used to construct a fast image matching algorithm.

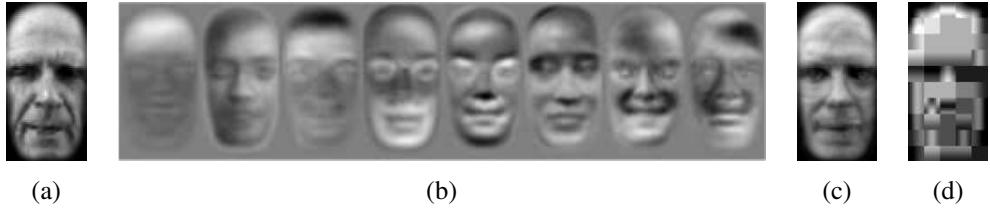


Figure 5.18 Face modeling and compression using eigenfaces (Moghaddam and Pentland 1997) © 1997 IEEE: (a) input image; (b) the first eight eigenfaces; (c) image reconstructed by projecting onto this basis and compressing the image to 85 bytes; (d) image reconstructed using JPEG (530 bytes).

In more detail, we start with a collection of *training images* $\{\mathbf{x}_j\}$, from which we compute the mean image \mathbf{m} and a *scatter* or *covariance* matrix

$$\mathbf{C} = \frac{1}{N} \sum_{j=0}^{N-1} (\mathbf{x}_j - \mathbf{m})(\mathbf{x}_j - \mathbf{m})^T. \quad (5.41)$$

We can apply the eigenvalue decomposition (A.6) to represent this matrix as

$$\mathbf{C} = \mathbf{U}\Lambda\mathbf{U}^T = \sum_{i=0}^{N-1} \lambda_i \mathbf{u}_i \mathbf{u}_i^T, \quad (5.42)$$

where the λ_i are the eigenvalues of \mathbf{C} and the \mathbf{u}_i are the *eigenvectors*. For general images, Kirby and Sirovich (1990) call these vectors *eigenpictures*; for faces, Turk and Pentland (1991) call them *eigenfaces* (Figure 5.18b).¹³

Two important properties of the eigenvalue decomposition are that the optimal (best approximation) coefficients a_i for any new image \mathbf{x} can be computed as

$$a_i = (\mathbf{x} - \mathbf{m}) \cdot \mathbf{u}_i, \quad (5.43)$$

and that, assuming the eigenvalues $\{\lambda_i\}$ are sorted in decreasing order, truncating the approximation given in (5.40) at any point M gives the best possible approximation (least error) between $\tilde{\mathbf{x}}$ and \mathbf{x} . Figure 5.18c shows the resulting approximation corresponding to Figure 5.18a and shows how much better it is at compressing a face image than JPEG.

Truncating the eigenface decomposition of a face image (5.40) after M components is equivalent to projecting the image onto a linear subspace F , which we can call the *face space*

¹³In actual practice, the full $P \times P$ scatter matrix (5.41) is never computed. Instead, a smaller $N \times N$ matrix consisting of the inner products between all the signed deviations $(\mathbf{x}_i - \mathbf{m})$ is accumulated instead. See Appendix A.1.2 (A.13–A.14) for details.

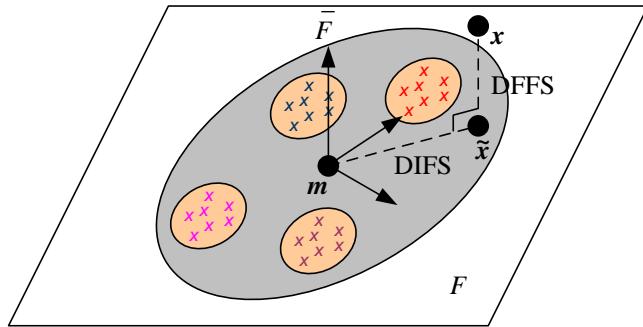


Figure 5.19 Projection onto the linear subspace spanned by the eigenface images (Moghaddam and Pentland 1997) © 1997 IEEE. The distance from face space (DFFS) is the orthogonal distance to the plane, while the distance in face space (DIFS) is the distance along the plane from the mean image. Both distances can be turned into Mahalanobis distances and given probabilistic interpretations.

(Figure 5.19). Because the eigenvectors (eigenfaces) are orthogonal and of unit norm, the distance of a projected face $\tilde{\mathbf{x}}$ to the mean face \mathbf{m} can be written as

$$\text{DIFS} = \|\tilde{\mathbf{x}} - \mathbf{m}\| = \left[\sum_{i=0}^{M-1} a_i^2 \right]^{1/2}, \quad (5.44)$$

where DIFS stands for *distance in face space* (Moghaddam and Pentland 1997). The remaining distance between the original image \mathbf{x} and its projection onto face space $\tilde{\mathbf{x}}$, i.e., the *distance from face space* (DFFS), can be computed directly in pixel space and represents the “faceness” of a particular image. It is also possible to measure the distance between two different faces in face space by taking the norm of their eigenface coefficients difference.

Computing such distances in Euclidean vector space, however, does not exploit the additional information that the eigenvalue decomposition of the covariance matrix (5.42) provides. To properly weight the distance based on the measured covariance, we can use the *Mahalanobis distance* (5.32) (Appendix B.1). A similar analysis can be performed for computing a sensible difference from face space (DFFS) (Moghaddam and Pentland 1997) and the two terms can be combined to produce an estimate of the likelihood of being a true face, which can be useful in doing face detection (Section 6.3.1). More detailed explanations of probabilistic and Bayesian PCA can be found in textbooks on statistical learning (Bishop 2006; Hastie, Tibshirani, and Friedman 2009; Murphy 2012), which also discuss techniques for selecting the optimum number of components M to use in modeling a distribution.

The original work on eigenfaces for recognition (Turk and Pentland 1991) was extended in Moghaddam and Pentland (1997), Heisele, Ho *et al.* (2003), and Heisele, Serre, and Poggio (2007) to include *modular eigenespaces* for separately modeling the appearance of different facial components such as the eyes, nose, and mouth, as well as *view-based eigenspaces* to separately model different views of a face. It was also extended by Belhumeur, Hespanha, and Kriegman (1997) to handle appearance variation due to illumination, modeling *intrapersonal* and *extrapersonal* variability separately, and using Fisher linear discriminant analysis (Figure 5.10) to perform recognition. A Bayesian extension of this work was subsequently developed by Moghaddam, Jebara, and Pentland (2000). These extensions are described in more detail in the cited papers, as well as the first edition of this book (Szeliski 2010, Section 14.2).

It is also possible to generalize the bilinear factorization implicit in PCA and SVD approaches to multilinear (tensor) formulations that can model several interacting factors simultaneously (Vasilescu and Terzopoulos 2007). These ideas are related to additional topics in machine learning such as *subspace learning* (Cai, He *et al.* 2007), *local distance functions* (Frome, Singer *et al.* 2007; Ramanan and Baker 2009), and *metric learning* (Kulis 2013).

5.2.4 Manifold learning

In many cases, the data we are analyzing does not reside in a globally linear subspace, but does live on a lower-dimensional manifold. In this case, non-linear dimensionality reduction can be used (Lee and Verleysen 2007). Since these systems extract lower-dimensional manifolds in a higher-dimensional space, they are also known as *manifold learning* techniques (Zheng and Xue 2009). Figure 5.20 shows some examples of two-dimensional manifolds extracted from the three-dimensional S-shaped ribbon using the scikit-learn manifold learning package.¹⁴

These results are just a small sample from the large number of algorithms that have been developed, which include multidimensional scaling (Kruskal 1964a,b), Isomap (Tenenbaum, De Silva, and Langford 2000), Local Linear Embedding (Roweis and Saul 2000), Hessian Eigenmaps (Donoho and Grimes 2003), Laplacian Eigenmaps (Belkin and Niyogi 2003), local tangent space alignment (Zhang and Zha 2004), Dimensionality Reduction by Learning an Invariant Mapping (Hadsell, Chopra, and LeCun 2006), Modified LLE (Zhang and Wang 2007), t-distributed Stochastic Neighbor Embedding (t-SNE) (van der Maaten and Hinton 2008; van der Maaten 2014), and UMAP (McInnes, Healy, and Melville 2018). Many of these algorithms are reviewed in Lee and Verleysen (2007), Zheng and Xue (2009), and on

¹⁴<https://scikit-learn.org/stable/modules/manifold.html>

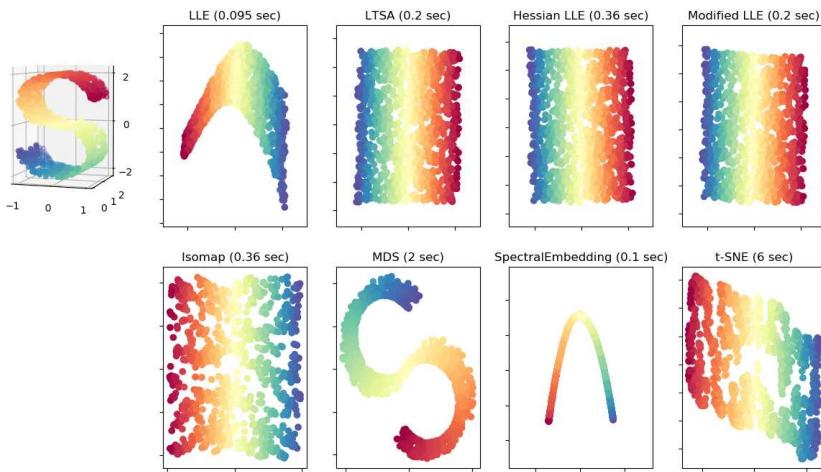


Figure 5.20 Examples of manifold learning, i.e., non-linear dimensionality reduction, applied to 1,000 points with 10 neighbors each, from <https://scikit-learn.org/stable/modules/manifold.html>. The eight sample outputs were produced by eight different embedding algorithms, as described in the scikit-learn manifold learning documentation page.

Wikipedia.¹⁵ Bengio, Paiement *et al.* (2004) describe a method for extending such algorithms to compute the embedding of new (“out-of-sample”) data points. McQueen, Meila *et al.* (2016) describe their megaman software package, which can efficiently solve embedding problems with millions of data points.

In addition to dimensionality reduction, which can be useful for regularizing data and accelerating similarity search, manifold learning algorithms can be used for visualizing input data distributions or neural network layer activations. Figure 5.21 show an example of applying two such algorithms (UMAP and t-SNE) to three different computer vision datasets.

5.2.5 Semi-supervised learning

In many machine learning settings, we have a modest amount of accurately labeled data and a far larger set of unlabeled or less accurate data. For example, an image classification dataset such as ImageNet may only contain one million labeled images, but the total number of images that can be found on the web is orders of magnitudes larger. Can we use this larger dataset, which still captures characteristics of our expect future inputs, to construct a better classifier or predictor?

¹⁵https://en.wikipedia.org/wiki/Nonlinear_dimensionality_reduction

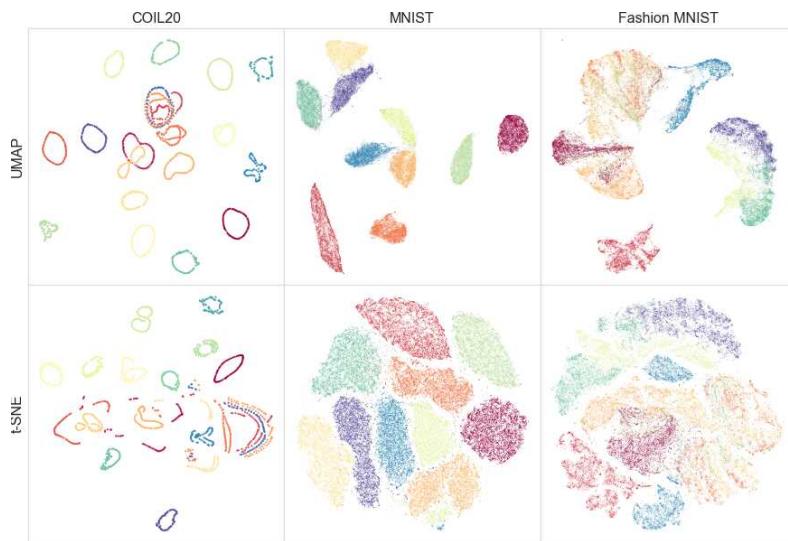


Figure 5.21 Comparison of UMAP and t-SNE manifold learning algorithms © McInnes, Healy, and Melville (2018) on three different computer vision learning recognition tasks: COIL (Nene, Nayar, and Murase 1996), MNIST (LeCun, Cortes, and Burges 1998), and Fashion MNIST (Xiao, Rasul, and Vollgraf 2017).

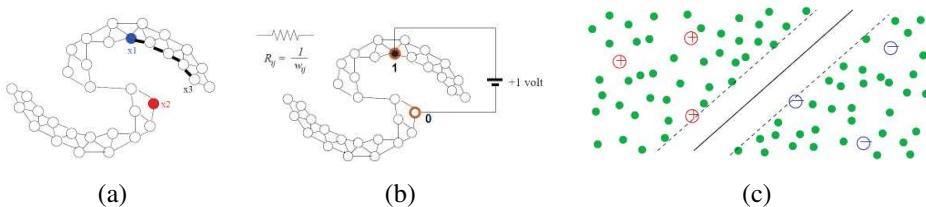


Figure 5.22 Examples of semi-supervised learning (Zhu and Goldberg 2009) © 2009 Morgan & Claypool: (a) two labeled samples and a graph connecting all of the samples; (b) solving binary labeling with harmonic functions, interpreted as a resistive electrical network; (c) using semi-supervised support vector machine (S3VM).

Consider the simple diagrams in Figure 5.22. Even if only a small number of examples are labeled with the correct class (in this case, indicated by red and blue circles or dots), we can still imagine extending these labels (inductively) to nearby samples and therefore not only labeling all of the data, but also constructing appropriate decision surfaces for future inputs.

This area of study is called *semi-supervised learning* (Zhu and Goldberg 2009; Subramanya and Talukdar 2014). In general, it comes in two varieties. In *transductive learning*, the goal is to classify all of the unlabeled inputs that are given as one batch at the same time as the labeled examples, i.e., all of the dots and circles shown in Figure 5.22. In *inductive learning*, we train a machine learning system that will classify all future inputs, i.e., all the regions in the input space. The second form is much more widely used, since in practice, most machine learning systems are used for online applications such as autonomous driving or new content classification.

Semi-supervised learning is a subset of the larger class of *weakly supervised learning* problems, where the training data may not only be missing labels, but also have labels of questionable accuracy (Zhou 2018). Some early examples from computer vision (Torresani 2014) include building whole image classifiers from image labels found on the internet (Fergus, Perona, and Zisserman 2004; Fergus, Weiss, and Torralba 2009) and object detection and/or segmentation (localization) with missing or very rough delineations in the training data (Nguyen, Torresani *et al.* 2009; Deselaers, Alexe, and Ferrari 2012). In the deep learning era, weakly supervised learning continues to be widely used (Pathak, Krahenbuhl, and Darrell 2015; Bilen and Vedaldi 2016; Arandjelovic, Gronat *et al.* 2016; Khoreva, Benenson *et al.* 2017; Novotny, Larlus, and Vedaldi 2017; Zhai, Oliver *et al.* 2019). A recent example of weakly supervised learning being applied to billions of noisily labeled images is pre-training deep neural networks on Instagram images with hashtags (Mahajan, Girshick *et al.* 2018). We will look at weakly and self-supervised learning techniques for *pre-training* neural networks in Section 5.4.7.

5.3 Deep neural networks

As we saw in the introduction to this chapter (Figure 5.2), deep learning pipelines take an end-to-end approach to machine learning, optimizing every stage of the processing by searching for parameters that minimize the training loss. In order for such search to be feasible, it helps if the loss is a differentiable function of all these parameters. Deep neural networks provide a uniform, differentiable computation architecture, while also automatically discovering useful internal representations.

Interest in building computing systems that mimic neural (biological) computation has

waxed and waned since the late 1950s, when Rosenblatt (1958) developed the *perceptron* and Widrow and Hoff (1960) derived the weight adaptation *delta rule*. Research into these topics was revitalized in the late 1970s by researchers who called themselves *connectionists*, organizing a series of meetings around this topic, which resulted in the foundation of the Neural Information Processing Systems (NeurIPS) conference in 1987. The recent book by Sejnowski (2018) has a nice historical review of this field’s development, as do the introductions in Goodfellow, Bengio, and Courville (2016) and Zhang, Lipton *et al.* (2021), the review paper by Rawat and Wang (2017), and the Turing Award lecture by Bengio, LeCun, and Hinton (2021). And while most of the deep learning community has moved away from biologically plausible models, some research still studies the connection between biological visual systems and neural network models (Yamins and DiCarlo 2016; Zhuang, Yan *et al.* 2020).

A good collection of papers from this era can be found in McClelland, Rumelhart, and PDP Research Group (1987), including the seminal paper on *backpropagation* (Rumelhart, Hinton, and Williams 1986a), which laid the foundation for the training of modern feedforward neural networks. During that time, and in the succeeding decades, a number of alternative neural network architectures were developed, including ones that used stochastic units such as Boltzmann Machines (Ackley, Hinton, and Sejnowski 1985) and Restricted Boltzmann Machines (Hinton and Salakhutdinov 2006; Salakhutdinov and Hinton 2009). The survey by Bengio (2009) has a review of some of these earlier approaches to deep learning. Many of these architectures are examples of the *generative graphical models* we saw in Section 4.3.

Today’s most popular deep neural networks are deterministic *discriminative feedforward networks* with real-valued activations, trained using gradient descent, i.e., the backpropagation training rule (Rumelhart, Hinton, and Williams 1986b). When combined with ideas from convolutional networks (Fukushima 1980; LeCun, Bottou *et al.* 1998), deep multi-layer neural networks produced the breakthroughs in speech recognition (Hinton, Deng *et al.* 2012) and visual recognition (Krizhevsky, Sutskever, and Hinton 2012; Simonyan and Zisserman 2014b) seen in the early 2010s. Zhang, Lipton *et al.* (2021, Chapter 7) have a nice description of the components that went into these breakthroughs and the rapid evolution in deep networks that has occurred since then, as does the earlier review paper by (Rawat and Wang 2017).

Compared to other machine learning techniques, which normally rely on several pre-processing stages to extract features on which classifiers can be built, deep learning approaches are usually trained *end-to-end*, going directly from raw pixels to final desired outputs (be they classifications or other images). In the next few sections, we describe the basic

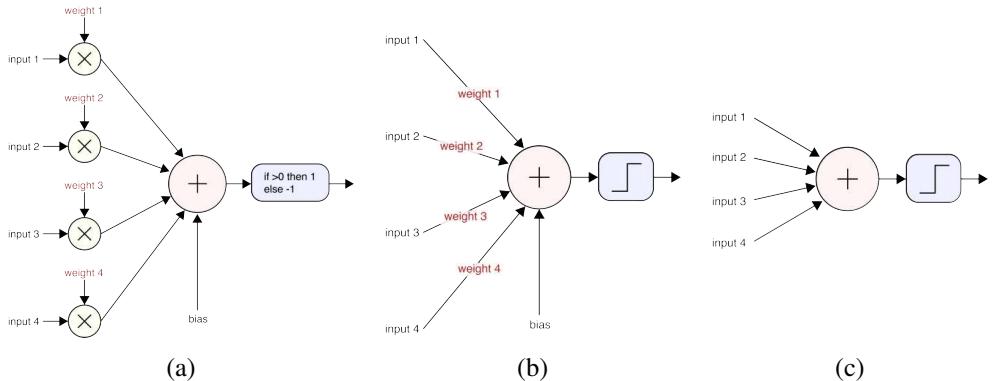


Figure 5.23 A perceptron unit (a) explicitly showing the weights being multiplied by the inputs, (b) with the weights written on the input connections, and (c) the most common form, with the weights and bias omitted. A non-linear activation function follows the weighted summation. © Glassner (2018)

components that go into constructing and training such neural networks. More detailed explanations on each topic can be found in textbooks on deep learning (Nielsen 2015; Goodfellow, Bengio, and Courville 2016; Glassner 2018, 2021; Zhang, Lipton *et al.* 2021) as well as the excellent course notes by Li, Johnson, and Yeung (2019) and Johnson (2020).

5.3.1 Weights and layers

Deep neural networks (DNNs) are *feedforward* computation graphs composed of thousands of simple interconnected “neurons” (*units*), which, much like logistic regression (5.18), perform weighted sums of their inputs

$$s_i = \mathbf{w}_i^T \mathbf{x}_i + b_i \quad (5.45)$$

followed by a non-linear *activation function* re-mapping,

$$y_i = h(s_i), \quad (5.46)$$

as illustrated in Figure 5.23. The \mathbf{x}_i are the inputs to the i th unit, \mathbf{w}_i and b_i are its learnable *weights* and *bias*, s_i is the output of the weighted linear sum, and y_i is the final output after s_i is fed through the activation function h .¹⁶ The outputs of each stage, which are often called

¹⁶Note that we have switched to using s_i for the weighted summations, since we will want to use l to index neural network layers.

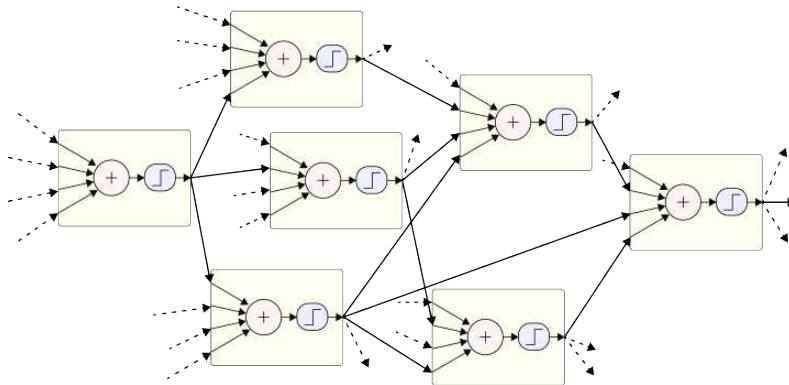


Figure 5.24 A multi-layer network, showing how the outputs of one unit are fed into additional units. © Glassner (2018)

the *activations*, are then fed into units in later stages, as shown in Figure 5.24.¹⁷

The earliest such units were called *perceptrons* (Rosenblatt 1958) and were diagrammed as shown in Figure 5.23a. Note that in this first diagram, the weights, which are optimized during the learning phase (Section 5.3.5), are shown explicitly along with the element-wise multiplications. Figure 5.23b shows a form in which the weights are written on top of the *connections* (arrows between units, although the arrowheads are often omitted). It is even more common to diagram nets as in Figure 5.23c, in which the weights (and bias) are completely omitted and assumed to be present.

Instead of being connected into an irregular computation graph as in Figure 5.24, neural networks are usually organized into consecutive *layers*, as shown in Figure 5.25. We can now think of all the units within a layer as being a vector, with the corresponding linear combinations written as

$$\mathbf{s}_l = \mathbf{W}_l \mathbf{x}_l, \quad (5.47)$$

where \mathbf{x}_l are the inputs to layer l , \mathbf{W}_l is a weight matrix, and \mathbf{s}_l is the weighted sum, to which an element-wise non-linearity is applied using a set of activation functions,

$$\mathbf{x}_{l+1} = \mathbf{y}_l = \mathbf{h}(\mathbf{s}_l). \quad (5.48)$$

A layer in which a full (dense) weight matrix is used for the linear combination is called a *fully connected* (FC) layer, since all of the inputs to one layer are connected to all of its

¹⁷Note that while almost all feedforward neural networks use linear weighted summations of their inputs, the Neocognitron (Fukushima 1980) also included a divisive normalization stage inspired by the behavior of biological neurons. Some of the latest DNNs also support multiplicative interactions between activations using *conditional batch norm* (Section 5.3.3).

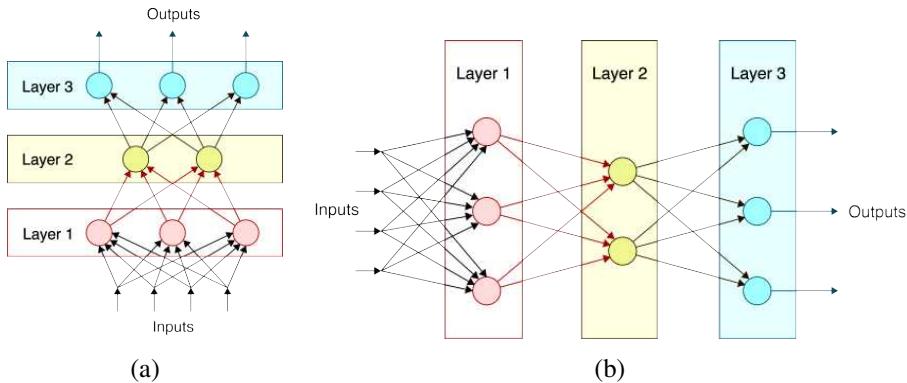


Figure 5.25 Two different ways to draw neural networks: (a) inputs at bottom, outputs at top, (b) inputs at left, outputs at right. © Glassner (2018)

outputs. As we will see in Section 5.4, when processing pixels (or other signals), early stages of processing use convolutions instead of dense connections for both spatial invariance and better efficiency.¹⁸ A network that consists only of fully connected (and no convolutional) layers is now often called a *multi-layer perceptron* (MLP).

5.3.2 Activation functions

Most early neural networks (Rumelhart, Hinton, and Williams 1986b; LeCun, Bottou *et al.* 1998) used sigmoidal functions similar to the ones used in logistic regression. Newer networks, starting with Nair and Hinton (2010) and Krizhevsky, Sutskever, and Hinton (2012), use Rectified Linear Units (ReLU) or variants. The ReLU activation function is defined as

$$h(y) = \max(0, y) \quad (5.49)$$

and is shown in the upper-left corner of Figure 5.26, along with some other popular functions, whose definitions can be found in a variety of publications (e.g., Goodfellow, Bengio, and Courville 2016, Section 6.3; Clevert, Unterthiner, and Hochreiter 2015; He, Zhang *et al.* 2015) and the Machine Learning Cheatsheet.¹⁹

While the ReLU is currently the most popular activation function, a widely cited observation in the CS231N course notes (Li, Johnson, and Yeung 2019) attributed to Andrej Karpathy

¹⁸Heads up for more confusing abbreviations: While a fully connected (dense) layer is often abbreviated as FC, a fully convolutional network, which is the opposite, i.e., sparsely connected with shared weights, is often abbreviated as FCN.

¹⁹https://ml-cheatsheet.readthedocs.io/en/latest/activation_functions.html

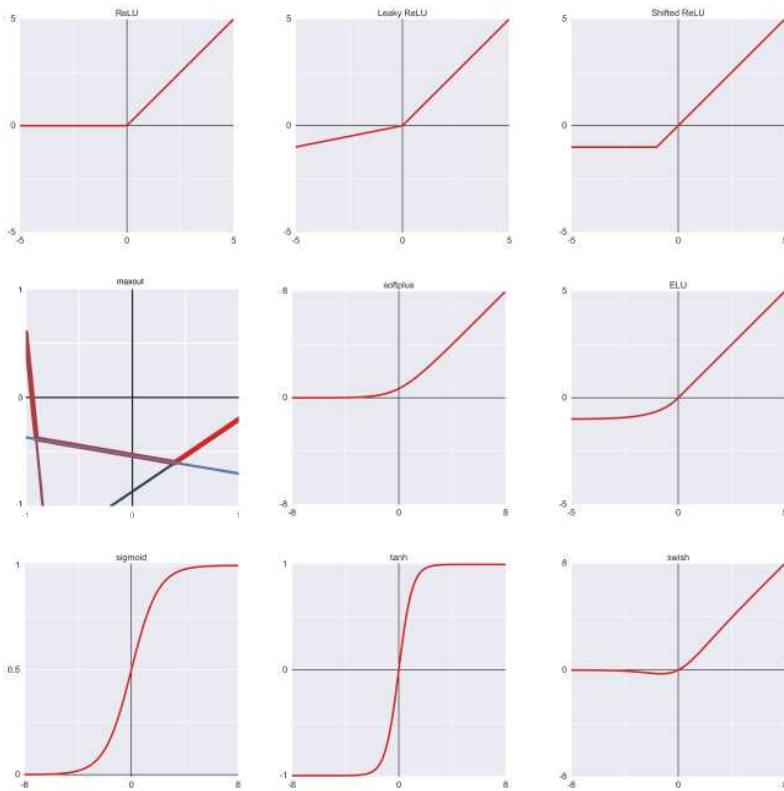


Figure 5.26 Some popular non-linear activation functions from © Glassner (2018): From top-left to bottom-right: ReLU, leaky ReLU, shifted ReLU, maxout, softplus, ELU, sigmoid, tanh, swish.

warns that²⁰

Unfortunately, ReLU units can be fragile during training and can “die”. For example, a large gradient flowing through a ReLU neuron could cause the weights to update in such a way that the neuron will never activate on any data-point again. If this happens, then the gradient flowing through the unit will forever be zero from that point on. That is, the ReLU units can irreversibly die during training since they can get knocked off the data manifold. ... With a proper setting of the learning rate this is less frequently an issue.

The CS231n course notes advocate trying some alternative non-clipping activation functions

²⁰<http://cs231n.github.io/neural-networks-1/#actfun>

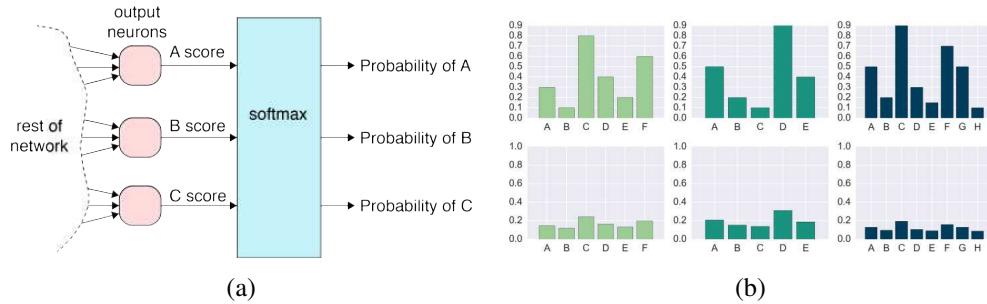


Figure 5.27 (a) A softmax layer used to convert from neural network activations (“score”) to class likelihoods (b) The top row shows the activations, while the bottom shows the result of running the scores through softmax to obtain properly normalized likelihoods. © Glassner (2018).

if this problem arises.

For the final layer in networks used for classification, the softmax function (5.3) is normally used to convert from real-valued activations to class likelihoods, as shown in Figure 5.27. We can thus think of the penultimate set of neurons as determining directions in activation space that most closely match the log likelihoods of their corresponding class, while minimizing the log likelihoods of alternative classes. Since the inputs flow forward to the final output classes and probabilities, feedforward networks are discriminative, i.e., they have no statistical model of the classes they are outputting, nor any straightforward way to generate samples from such classes (but see Section 5.5.4 for techniques to do this).

5.3.3 Regularization and normalization

As with other forms of machine learning, regularization and other techniques can be used to prevent neural networks from overfitting so they can better generalize to unseen data. In this section, we discuss traditional methods such as regularization and data augmentation that can be applied to most machine learning systems, as well as techniques such as dropout and batch normalization, which are specific to neural networks.

Regularization and weight decay

As we saw in Section 4.1.1, quadratic or p -norm penalties on the weights (4.9) can be used to improve the conditioning of the system and to reduce overfitting. Setting $p = 2$ results in the usual L_2 regularization and makes large weights smaller, whereas using $p = 1$ is called



Figure 5.28 An original “6” digit from the MNIST database and two elastically distorted versions (Simard, Steinkraus, and Platt 2003) © 2003 IEEE.

lasso (least absolute shrinkage and selection operator) and can drive some weights all the way to zero. As the weights are being optimized inside a neural network, these terms make the weights smaller, so this kind of regularization is also known as *weight decay* (Bishop 2006, Section 3.1.4; Goodfellow, Bengio, and Courville 2016, Section 7.1; Zhang, Lipton *et al.* 2021, Section 4.5).²¹ Note that for more complex optimization algorithms such as Adam, L_2 regularization and weight decay are *not* equivalent, but the desirable properties of weight decay can be restored using a modified algorithm (Loshchilov and Hutter 2019).

Dataset augmentation

Another powerful technique to reduce over-fitting is to add more training samples by perturbing the inputs and/or outputs of the samples that have already been collected. This technique is known as *dataset augmentation* (Zhang, Lipton *et al.* 2021, Section 13.1) and can be particularly effective on image classification tasks, since it is expensive to obtain labeled examples, and also since image classes should not change under small local perturbations.

An early example of such work applied to a neural network classification task is the *elastic distortion* technique proposed by Simard, Steinkraus, and Platt (2003). In their approach, random low-frequency displacement (warp) fields are synthetically generated for each training example and applied to the inputs during training (Figure 5.28). Note how such distortions are *not* the same as simply adding pixel noise to the inputs. Instead, distortions move pixels around, and therefore introduce much larger changes in the input vector space, while still preserving the semantic meaning of the examples (in this case, MNIST digits (LeCun, Cortes, and Burges 1998)).

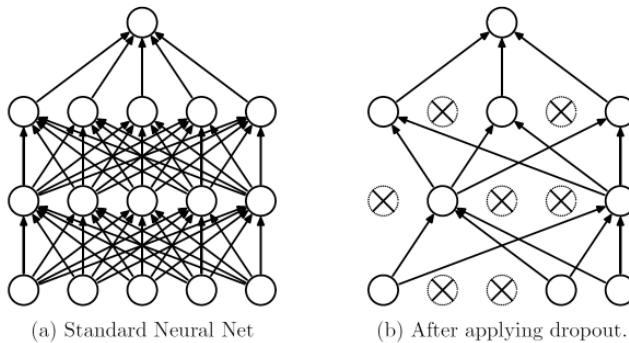


Figure 5.29 When using dropout, during training some fraction of units p is removed from the network (or, equivalently, clamped to zero) © Srivastava, Hinton et al. (2014). Doing this randomly for each mini-batch injects noise into the training process (at all levels of the network) and prevents the network from overly relying on particular units.

Dropout

Dropout is a regularization technique introduced by Srivastava, Hinton et al. (2014), where at each mini-batch during training (Section 5.3.6), some percentage p (say 50%) of the units in each layer are clamped to zero, as shown in Figure 5.29. Randomly setting units to zero injects noise into the training process and also prevents the network from overly specializing units to particular samples or tasks, both of which can help reduce overfitting and improve generalization.

Because dropping (zeroing out) p of the units reduces the expected value of any sum the unit contributes to by a fraction $(1 - p)$, the weighted sums s_i in each layer (5.45) are multiplied (during training) by $(1 - p)^{-1}$. At test time, the network is run with no dropout and no compensation on the sums. A more detailed description of dropout can be found in Zhang, Lipton et al. (2021, Section 4.6) and Johnson (2020, Lecture 10).

Batch normalization

Optimizing the weights in a deep neural network, which we discuss in more detail in Section 5.3.6, is a tricky process and may be slow to converge.

One of the classic problems with iterative optimization techniques is poor *conditioning*, where the components of the gradient vary greatly in magnitude. While it is sometimes

²¹From a Bayesian perspective, we can also think of this penalty as a Gaussian prior on the weight distribution (Appendix B.4).

possible to reduce these effects with preconditioning techniques that scale individual elements in a gradient before taking a step (Section 5.3.6 and Appendix A.5.2), it is usually preferable to control the condition number of the system during the problem formulation.

In deep networks, one way in which poor conditioning can manifest itself is if the sizes of the weights or activations in successive layers become imbalanced. Say we take a given network and scale all of the weights in one layer by $100\times$ and scale down the weights in the next layer by the same amount. Because the ReLU activation function is linear in both of its domains, the outputs of the second layer will still be the same, although the activations at the output of the first layer will be 100 times larger. During the gradient descent step, the derivatives with respect to the weights will be vastly different after this rescaling, and will in fact be opposite in magnitude to the weights themselves, requiring tiny gradient descent steps to prevent overshooting (see Exercise 5.4).²²

The idea behind batch normalization (Ioffe and Szegedy 2015) is to re-scale (and re-center) the activations at a given unit so that they have unit variance and zero mean (which, for a ReLU activation function, means that the unit will be active half the time). We perform this normalization by considering all of the training samples n in a given minibatch \mathcal{B} (5.71) and computing the mean and variance statistics for unit i as

$$\mu_i = \frac{1}{|\mathcal{B}|} \sum_{n \in \mathcal{B}} s_i^{(n)} \quad (5.50)$$

$$\sigma_i^2 = \frac{1}{|\mathcal{B}|} \sum_{n \in \mathcal{B}} (s_i^{(n)} - \mu_i)^2 \quad (5.51)$$

$$\hat{s}_i^{(n)} = \frac{s_i^{(n)} - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}}, \quad (5.52)$$

where s_i^n is the weighted sum of unit i for training sample n , $\hat{s}_i^{(n)}$ is the corresponding batch normalized sum, and ϵ (often 10^{-5}) is a small constant to prevent division by zero.

After batch normalization, the $\hat{s}_i^{(n)}$ activations now have zero mean and unit variance. However, this normalization may run at cross-purpose to the minimization of the loss function during training. For this reason, Ioffe and Szegedy (2015) add an extra gain γ_i and bias β_i parameter to each unit i and define the output of a batch normalization stage to be

$$y_i = \gamma_i \hat{s}_i + \beta_i. \quad (5.53)$$

²²This motivating paragraph is my own explanation of why batch normalization might be a good idea, and is related to the idea that batch normalization reduces *internal covariate shift*, used by (Ioffe and Szegedy 2015) to justify their technique. This hypothesis is now being questioned and alternative theories are being developed (Bjorck, Gomes *et al.* 2018; Santurkar, Tsipras *et al.* 2018; Kohler, Daneshmand *et al.* 2019).

These parameters act just like regular weights, i.e., they are modified using gradient descent during training to reduce the overall training loss.²³

One subtlety with batch normalization is that the μ_i and σ_i^2 quantities depend analytically on all of the activation for a given unit in a minibatch. For gradient descent to be properly defined, the derivatives of the loss function with respect to these variables, and the derivatives of the quantities \hat{s}_i and y_i with respect to these variables, must be computed as part of the gradient computation step, using similar chain rule computations as the original backpropagation algorithm (5.65–5.68). These derivations can be found in Ioffe and Szegedy (2015) as well as several blogs.²⁴

When batch normalization is applied to convolutional layers (Section 5.4), one could in principle compute a normalization separately for each pixel, but this would add a tremendous number of extra learnable bias and gain parameters (β_i, γ_i). Instead, batch normalization is usually implemented by computing the statistics as sums over all the pixels with the same convolution kernel, and then adding a single bias and gain parameter for each convolution kernel (Ioffe and Szegedy 2015; Johnson 2020, Lecture 10; Zhang, Lipton *et al.* 2021, Section 7.5).

Having described how batch normalization operates during training, we still need to decide what to do at test or inference time, i.e., when applying the trained network to new data. We cannot simply skip this stage, as the network was trained while removing common mean and variance estimates. For this reason, the mean and variance estimates are usually recomputed over the *whole* training set, or some running average of the per-batch statistics are used. Because of the linear form of (5.45) and (5.52–5.53), it is possible to fold the μ_i and σ_i estimates and learned (β_i, γ_i) parameters into the original weight and bias terms in (5.45).

Since the publication of the seminal paper by Ioffe and Szegedy (2015), a number of variants have been developed, some of which are illustrated in Figure 5.30. Instead of accumulating statistics over the samples in a minibatch \mathcal{B} , we can compute them over different subsets of activations in a layer. These subsets include:

- all the activations in a layer, which is called *layer normalization* (Ba, Kiros, and Hinton 2016);
- all the activations in a given convolutional output channel (see Section 5.4), which is called *instance normalization* (Ulyanov, Vedaldi, and Lempitsky 2017);

²³There is a trick used by those in the know, which relies on the observation that any bias term b_i in the original summation s_i (5.45) shows up in the mean μ_i and gets subtracted out. For this reason, the bias term is often omitted when using batch (or other kinds of) normalization.

²⁴<https://kratzert.github.io/2016/02/12/understanding-the-gradient-flow-through-the-batch-normalization-layer.html>, https://kevinzakka.github.io/2016/09/14/batch_normalization.html, <https://deepnotes.io/batchnorm>

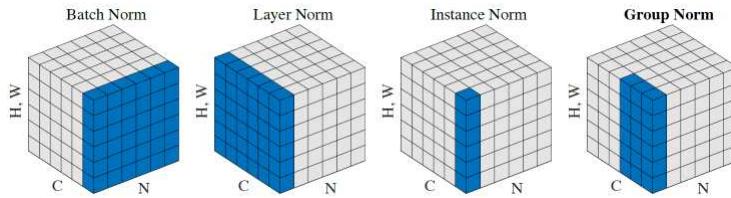


Figure 5.30 *Batch norm, layer norm, instance norm, and group norm*, from Wu and He (2018) © 2018 Springer. The (H, W) dimension denotes pixels, C denotes channels, and N denotes training samples in a minibatch. The pixels in blue are normalized by the same mean and variance.

- different sub-groups of output channels, which is called *group normalization* (Wu and He 2018).

The paper by Wu and He (2018) describes each of these in more detail and also compares them experimentally. More recent work by Qiao, Wang *et al.* (2019a) and Qiao, Wang *et al.* (2019b) discusses some of the disadvantages of these newer variants and proposes two new techniques called *weight standardization* and *batch channel normalization* to mitigate these problems.

Instead of modifying the activations in a layer using their statistics, it is also possible to modify the weights in a layer to explicitly make the weight norm and weight vector direction separate parameters, which is called *weight normalization* (Salimans and Kingma 2016). A related technique called *spectral normalization* (Miyato, Kataoka *et al.* 2018) constrains the largest singular value of the weight matrix in each layer to be 1.

The bias and gain parameters (β_i, γ_i) may also depend on the activations in some other layer in the network, e.g., derived from a guide image.²⁵ Such techniques are referred to as *conditional batch normalization* and have been used to select between different artistic styles (Dumoulin, Shlens, and Kudlur 2017) and to enable local semantic guidance in image synthesis (Park, Liu *et al.* 2019). Related techniques and applications are discussed in more detail in Section 14.6 on neural rendering.

The reasons why batch and other kinds of normalization help deep networks converge faster and generalize better are still being debated. Some recent papers on this topic include Bjorck, Gomes *et al.* (2018), Hoffer, Banner *et al.* (2018), Santurkar, Tsipras *et al.* (2018), and Kohler, Daneshmand *et al.* (2019).

²⁵Note that this gives neural networks the ability to multiply two layers in a network, which we used previously to perform locally (Section 3.5.5).

5.3.4 Loss functions

In order to optimize the weights in a neural network, we need to first define a *loss function* that we minimize over the training examples. We have already seen the main loss functions used in machine learning in previous parts of this chapter.

For classification, most neural networks use a final softmax layer (5.3), as shown in Figure 5.27. Since the outputs are meant to be class probabilities that sum up to 1, it is natural to use the cross-entropy loss given in (5.19) or (5.23–5.24) as the function to minimize during training. Since in our description of the feedforward networks we have used indices i and j to denote neural units, we will, in this section, use n to index a particular training example.

The multi-class cross-entropy loss can thus be re-written as

$$E(\mathbf{w}) = \sum_n E_n(\mathbf{w}) = -\sum_n \log p_{nt_n}, \quad (5.54)$$

where \mathbf{w} is the vector of all weights, biases, and other model parameters, and p_{nk} is the network's current estimate of the probability of class k for sample n , and t_n is the integer denoting the correct class. Substituting the definition of p_{nk} from (5.20) with the appropriate replacement of l_{ik} with s_{nk} (the notation we use for neural nets), we get

$$E_n(\mathbf{w}) = \log Z_n - s_{nt_n} \quad (5.55)$$

with $Z_n = \sum_j \exp s_{nj}$. Gómez (2018) has a nice discussion of some of the losses widely used in deep learning.

For networks that perform regression, i.e., generate one or more continuous variables such as depth maps or denoised images, it is common to use an L_2 loss,

$$E(\mathbf{w}) = \sum_n E_n(\mathbf{w}) = -\sum_n \|\mathbf{y}_n - \mathbf{t}_n\|^2, \quad (5.56)$$

where \mathbf{y}_n is the network output for sample n and \mathbf{t}_n is the corresponding training (target) value, since this is a natural measure of error between continuous variables. However, if we believe there may be outliers in the training data, or if gross errors are not so harmful as to merit a quadratic penalty, more robust norms such as L_1 can be used (Barron 2019; Ranftl, Lasinger *et al.* 2020). (It is also possible to use robust norms for classification, e.g., adding an outlier probability to the class labels.)

As it is common to interpret the final outputs of a network as a probability distribution, we need to ask whether it is wise to use such probabilities as a measure of confidence in a particular answer. If a network is properly trained and predicting answers with good accuracy, it is tempting to make this assumption. The training losses we have presented so far, however,

only encourage the network to maximize the probability-weighted correct answers, and do not, in fact, encourage the network outputs to be properly *confidence calibrated*. Guo, Pleiss *et al.* (2017) discuss this issue, and present some simple measures, such as multiplying the log-likelihoods by a *temperature* (Platt 2000a), to improve the match between classifier probabilities and true reliability. The GrokNet image recognition system (Bell, Liu *et al.* 2020), which we discuss in Section 6.2.3, uses calibration to obtain better attribute probability estimates.

For networks that hallucinate new images, e.g., when introducing missing high-frequency details (Section 10.3) or doing image transfer tasks (Section 14.6), we may want to use a *perceptual loss* (Johnson, Alahi, and Fei-Fei 2016; Dosovitskiy and Brox 2016; Zhang, Isola *et al.* 2018), which uses intermediate layer neural network responses as the basis of comparison between target and output images. It is also possible to train a separate *discriminator* network to evaluate the quality (and plausibility) of synthesized images, as discussed in Section 5.5.4 More details on the application of loss functions to image synthesis can be found in Section 14.6 on neural rendering.

While loss functions are traditionally applied to supervised learning tasks, where the correct label or target value t_n is given for each input, it is also possible to use loss functions in an unsupervised setting. An early example of this was the *contrastive loss* function proposed by Hadsell, Chopra, and LeCun (2006) to cluster samples that are similar together while spreading dissimilar samples further apart. More formally, we are given a set of inputs $\{\mathbf{x}_i\}$ and pairwise indicator variables $\{t_{ij}\}$ that indicate whether two inputs are similar.²⁶ The goal is now to compute an embedding \mathbf{v}_i for each input \mathbf{x}_i such that similar input pairs have similar embeddings (low distances), while dissimilar inputs have large embedding distances. Finding mappings or embeddings that create useful distances between samples is known as (*distance*) *metric learning* (Köstinger, Hirzer *et al.* 2012; Kulis 2013) and is a commonly used tool in machine learning. The losses used to encourage the creation of such meaningful distances are collectively known as *ranking losses* (Gómez 2019) and can be used to relate features from different domains such as text and images (Karpathy, Joulin, and Fei-Fei 2014).

The contrastive loss from (Hadsell, Chopra, and LeCun 2006) is defined as

$$E_{\text{CL}} = \sum_{(i,j) \in \mathcal{P}} \{t_{ij} \log L_S(d_{ij}) + (1 - t_{ij}) \log L_D(d_{ij})\}, \quad (5.57)$$

where \mathcal{P} is the set of all labeled input pairs, L_S and L_D are the similar and dissimilar loss functions, and $d_{ij} = \|\mathbf{v}_i - \mathbf{v}_j\|$ are the pairwise distance between paired embeddings.²⁷

²⁶Indicator variables are often denoted as y_{ij} , but we will stick to the t_{ij} notation to be consistent with Section 5.1.3.

²⁷In metric learning, the embeddings are very often normalized to unit length.

This has a form similar to the cross-entropy loss given in (5.19), except that we measure squared distances between encodings \mathbf{v}_i and \mathbf{v}_j . In their paper, Hadsell, Chopra, and LeCun (2006) suggest using a quadratic function for L_S and a quadratic hinge loss (c.f. (5.30)) $L_D = [m - d_{ij}]_+^2$ for dissimilarity, where m is a margin beyond which there is no penalty.

To train with a contrastive loss, you can run both pairs of inputs through the neural network, compute the loss, and then backpropagate the gradients through both instantiations (activations) of the network. This can also be thought of as constructing a *Siamese network* consisting of two copies with shared weights (Bromley, Guyon *et al.* 1994; Chopra, Hadsell, and LeCun 2005). It is also possible to construct a *triplet loss* that takes as input a pair of matching samples and a third non-matching sample and ensures that the distance between non-matching samples is greater than the distance between matches plus some margin (Weinberger and Saul 2009; Weston, Bengio, and Usunier 2011; Schroff, Kalenichenko, and Philbin 2015; Rawat and Wang 2017).

Both pairwise contrastive and triplet losses can be used to learn embeddings for visual similarity search (Bell and Bala 2015; Wu, Manmatha *et al.* 2017; Bell, Liu *et al.* 2020), as discussed in more detail in Section 6.2.3. They have also been recently used for unsupervised pre-training of neural networks (Wu, Xiong *et al.* 2018; He, Fan *et al.* 2020; Chen, Kornblith *et al.* 2020), which we discuss in Section 5.4.7. In this case, it is more common to use a different contrastive loss function, inspired by softmax (5.3) and multi-class cross-entropy (5.20–5.22), which was first proposed by (Sohn 2016). Before computing the loss, the embeddings are all normalized to unit norm, $\|\hat{\mathbf{v}}_i\|^2 = 1$. Then, the following loss is summed over all matching embeddings,

$$l_{ij} = -\log \frac{\exp(\hat{\mathbf{v}}_i \cdot \hat{\mathbf{v}}_j / \tau)}{\sum_k \exp(\hat{\mathbf{v}}_i \cdot \hat{\mathbf{v}}_k / \tau)}, \quad (5.58)$$

with the denominator summed over non-matches as well. The τ variable denotes the “temperature” and controls how tight the clusters will be; it is sometimes replaced with an s multiplier parameterizing the hyper-sphere radius (Deng, Guo *et al.* 2019). The exact details of how the matches are computed vary by exact implementation.

This loss goes by several names, including InfoNCE (Oord, Li, and Vinyals 2018), and NT-Xent (normalized temperature cross-entropy loss) in Chen, Kornblith *et al.* (2020). Generalized versions of this loss called SphereFace, CosFace, and ArcFace are discussed and compared in the ArcFace paper (Deng, Guo *et al.* 2019) and used by Bell, Liu *et al.* (2020) as part of their visual similarity search system. The *smoothed average precision* loss recently proposed by Brown, Xie *et al.* (2020) can sometimes be used as an alternative to the metric losses discussed in this section. Some recent papers that compare and discuss deep metric learning approaches include (Jacob, Picard *et al.* 2019; Musgrave, Belongie, and Lim 2020).

Weight initialization

Before we can start optimizing the weights in our network, we must first initialize them. Early neural networks used small random weights to break the symmetry, i.e., to make sure that all of the gradients were not zero. It was observed, however, that in deeper layers, the activations would get progressively smaller.

To maintain a comparable variance in the activations of successive layers, we must take into account the *fan-in* of each layer, i.e., the number of incoming connections where activations get multiplied by weights. Glorot and Bengio (2010) did an initial analysis of this issue, and came up with a recommendation to set the random initial weight variance as the inverse of the fan-in. Their analysis, however, assumed a linear activation function (at least around the origin), such as a tanh function.

Since most modern deep neural networks use the ReLU activation function (5.49), He, Zhang *et al.* (2015) updated this analysis to take into account this asymmetric non-linearity. If we initialize the weights to have zero mean and variance V_l for layer l and set the original biases to zero, the linear summation in (5.45) will have a variance of

$$\text{Var}[s_l] = n_l V_l E[x_l^2], \quad (5.59)$$

where n_l is the number of incoming activations/weights and $E[x_l^2]$ is the expectation of the squared incoming activations. When the summations s_l , which have zero mean, are fed through the ReLU, the negative ones will get clamped to zero, so the expectation of the squared output $E[y_l^2]$ is half the variance of s_l , $\text{Var}[s_l]$.

In order to avoid decaying or increasing average activations in deeper layers, we want the magnitude of the activations in successive layers to stay about the same. Since we have

$$E[y_l^2] = \frac{1}{2} \text{Var}[s_l] = \frac{1}{2} n_l V_l E[x_l^2], \quad (5.60)$$

we conclude that the variance in the initial weights V_l should be set to

$$V_l = \frac{2}{n_l}, \quad (5.61)$$

i.e., the inverse of half the fan-in of a given unit or layer. This weight initialization rule is commonly called *He initialization*.

Neural network initialization continues to be an active research area, with publications that include Krähenbühl, Doersch *et al.* (2016), Mishkin and Matas (2016), Frankle and Carbin (2019), and Zhang, Dauphin, and Ma (2019)

5.3.5 Backpropagation

Once we have set up our neural network by deciding on the number of layers, their widths and depths, added some regularization terms, defined the loss function, and initialized the weights, we are ready to train the network with our sample data. To do this, we use gradient descent or one of its variants to iteratively modify the weights until the network has converged to a good set of values, i.e., an acceptable level of performance on the training and validation data.

To do this, we compute the derivatives (gradients) of the loss function E_n for training sample n with respect to the weights w using the chain rule, starting with the outputs and working our way back through the network towards the inputs, as shown in Figure 5.31. This procedure is known as *backpropagation* (Rumelhart, Hinton, and Williams 1986b) and stands for *backward propagation of errors*. You can find alternative descriptions of this technique in textbooks and course notes on deep learning, including Bishop (2006, Section 5.3.1), Goodfellow, Bengio, and Courville (2016, Section 6.5), Glassner (2018, Chapter 18), Johnson (2020, Lecture 6), and Zhang, Lipton *et al.* (2021).

Recall that in the forward (evaluation) pass of a neural network, activations (layer outputs) are computed layer-by-layer, starting with the first layer and finishing at the last. We will see in the next section that many newer DNNs have an acyclic graph structure, as shown in Figures 5.42–5.43, rather than just a single linear pipeline. In this case, any breadth-first traversal of the graph can be used. The reason for this evaluation order is computational efficiency. Activations need only be computed once for each input sample and can be re-used in succeeding stages of computation.

During backpropagation, we perform a similar breadth-first traversal of the reverse graph. However, instead of computing activations, we compute derivatives of the loss with respect to the weights and inputs, which we call *errors*. Let us look at this in more detail, starting with the loss function.

The derivative of the cross-entropy loss E_n (5.54) with respect to the output probability p_{nk} is simply $-\delta_{nt_n}/p_{nk}$. What is more interesting is the derivative of the loss with respect to the *scores* s_{nk} going into the softmax layer (5.55) shown in Figure 5.27,

$$\frac{\partial E_n}{\partial s_{nk}} = -\delta_{nt_n} + \frac{1}{Z_n} \exp s_{nk} = p_{nk} - \delta_{nt_n} = p_{nk} - \tilde{t}_{nk}. \quad (5.62)$$

(The last form is useful if we are using one-hot encoding or the targets have non-binary probabilities.) This has a satisfyingly intuitive explanation as the difference between the predicted class probability p_{nk} and the true class identity t_{nk} .

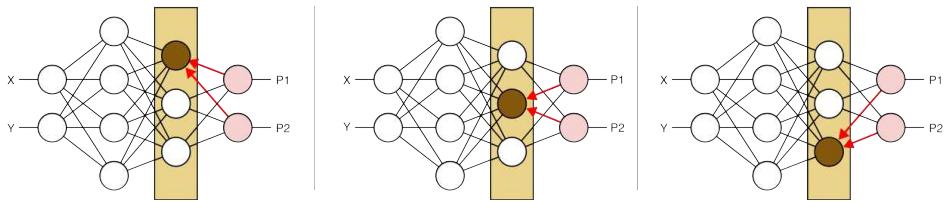


Figure 5.31 Backpropagating the derivatives (errors) through an intermediate layer of the deep network © Glassner (2018). The derivatives of the loss function applied to a single training example with respect to each of the pink unit inputs are summed together and the process is repeated chaining backward through the network.

For the L_2 loss in (5.56), we get a similar result,

$$\frac{\partial E_n}{\partial y_{nk}} = y_{nk} - t_{nk}, \quad (5.63)$$

which in this case denotes the real-valued difference between the predicted and target values.

In the rest of this section, we drop the sample index n from the activations x_{in} and y_{in} , since the derivatives for each sample n can typically be computed independently from other samples.²⁸

To compute the partial derivatives of the loss term with respect to earlier weights and activations, we work our way back through the network, as shown in Figure 5.31. Recall from (5.45–5.46) that we first compute a weighted sum s_i by taking a dot product between the input activations \mathbf{x}_i and the unit's weight vector \mathbf{w}_i ,

$$s_i = \mathbf{w}_i^T \mathbf{x}_i + b_i = \sum_j w_{ij} x_{ij} + b_i. \quad (5.64)$$

We then pass this weighted sum through an activation function h to obtain $y_i = h(s_i)$.

To compute the derivative of the loss E_n with respect to the weights, bias, and input

²⁸This is not the case if batch or other kinds of normalization (Section 5.3.3) are being used. For batch normalization, we have to accumulate the statistics across all the samples in the batch and then take their derivatives with respect to each weight (Ioffe and Szegedy 2015). For instance and group norm, we compute the statistics across all the pixels in a given channel or group, and then have to compute these additional derivatives as well.

activations, we use the chain rule,

$$e_i = \frac{\partial E_n}{\partial s_i} = h'(s_i) \frac{\partial E_n}{\partial y_i}, \quad (5.65)$$

$$\frac{\partial E_n}{\partial w_{ij}} = x_{ij} \frac{\partial E_n}{\partial s_i} = x_{ij} e_i, \quad (5.66)$$

$$\frac{\partial E_n}{\partial b_i} = \frac{\partial E_n}{\partial s_i} = e_i, \quad \text{and} \quad (5.67)$$

$$\frac{\partial E_n}{\partial x_{ij}} = w_{ij} \frac{\partial E_n}{\partial s_i} = w_{ij} e_i. \quad (5.68)$$

We call the term $e_i = \partial E_n / \partial s_i$, i.e., the partial derivative of the loss E_n with respect to the summed activation s_i , the *error*, as it gets propagated backward through the network.

Now, where do these errors come from, i.e., how do we obtain $\partial E_n / \partial y_i$? Recall from Figure 5.24 that the outputs from one unit or layer become the inputs for the next layer. In fact, for a simple network like the one in Figure 5.24, if we let x_{ij} be the activation that unit i receives from unit j (as opposed to just the j th input to unit i), we can simply set $x_{ij} = y_j$.

Since y_i , the output of unit i , now serves as input for the other units $k > i$ (assuming the units are ordered breadth first), we have

$$\frac{\partial E_n}{\partial y_i} = \sum_{k>i} \frac{\partial E_n}{\partial x_{ki}} = \sum_{k>i} w_{ki} e_k \quad (5.69)$$

and

$$e_i = h'(s_i) \frac{\partial E_n}{\partial y_i} = h'(s_i) \sum_{k>i} w_{ki} e_k. \quad (5.70)$$

In other words, to compute a unit's (backpropagation) error, we compute a weighted sum of the errors coming from the units it feeds into and then multiply this by the derivative of the current activation function $h'(s_i)$. This backward flow of errors is shown in Figure 5.31, where the errors for the three units in the shaded box are computed using weighted sums of the errors coming from later in the network.

This backpropagation rule has a very intuitive explanation. The error (derivative of the loss) for a given unit depends on the errors of the units that it feeds multiplied by the weights that couple them together. This is a simple application of the chain rule. The slope of the activation function $h'(s_i)$ modulates this interaction. If the unit's output is clamped to zero or small, e.g., with a negative-input ReLU or the “flat” part of a sigmoidal response, the unit's error is itself zero or small. The gradient of the weight, i.e., how much the weight should be perturbed to reduce the loss, is a signed product of the incoming activation and the unit's error, $x_{ij} e_i$. This is closely related to the Hebbian update rule (Hebb 1949), which observes that synaptic efficiency in biological neurons increases with correlated firing in the presynaptic

and postsynaptic cells. An easier way to remember this rule is “neurons wire together if they fire together” (Lowel and Singer 1992).

There are, of course, other computational elements in modern neural networks, including convolutions and pooling, which we cover in the next section. The derivatives and error propagation through such other units follows the same procedure as we sketched here, i.e., recursively apply the chain rule, taking analytic derivatives of the functions being applied, until you have the derivatives of the loss function with respect to all the parameters being optimized, i.e., the gradient of the loss.

As you may have noticed, the computation of the gradients with respect to the weights requires the unit activations computed in the forward pass. A typical implementation of neural network training stores the activations for a given sample and uses these during the backprop (backward error propagation) stage to compute the weight derivatives. Modern neural networks, however, may have millions of units and hence activations (Figure 5.44). The number of activations that need to be stored can be reduced by only storing them at certain layers and then re-computing the rest as needed, which goes under the name *gradient checkpointing* (Griewank and Walther 2000; Chen, Xu *et al.* 2016; Bulatov 2018).²⁹ A more extensive review of low-memory training can be found in the technical report by Sohoni, Aberger *et al.* (2019).

5.3.6 Training and optimization

At this point, we have all of the elements needed to train a neural network. We have defined the network’s topology in terms of the sizes and depths of each layer, specified our activation functions, added regularization terms, specified our loss function, and initialized the weights. We have even described how to compute the gradients, i.e., the derivatives of the regularized loss with respect to all of our weights. What we need at this point is some algorithm to turn these gradients into weight updates that will optimize the loss function and produce a network that generalizes well to new, unseen data.

In most computer vision algorithms such as optical flow (Section 9.1.3), 3D reconstruction using bundle adjustment (Section 11.4.2), and even in smaller-scale machine learning problems such as logistic regression (Section 5.1.3), the method of choice is linearized least squares (Appendix A.3). The optimization is performed using a second-order method such as Gauss-Newton, in which we evaluate all of the terms in our loss function and then take an optimally-sized downhill step using a direction derived from the gradients and the Hessian of the energy function.

²⁹This name seems a little weird, since it’s actually the activations that are saved instead of the gradients.

Unfortunately, deep learning problems are far too large (in terms of number of parameters and training samples; see Figure 5.44) to make this approach practical. Instead, practitioners have developed a series of optimization algorithms based on extensions to *stochastic gradient descent* (SGD) (Zhang, Lipton *et al.* 2021, Chapter 11). In SGD, instead of evaluating the loss function by summing over all the training samples, as in (5.54) or (5.56), we instead just evaluate a single training sample n and compute the derivatives of the associated loss $E_n(\mathbf{w})$. We then take a tiny downhill step along the direction of this gradient.

In practice, the directions obtained from just a single sample are incredibly noisy estimates of a good descent direction, so the losses and gradients are usually summed over a small subset of the training data,

$$E_{\mathcal{B}}(\mathbf{w}) = \sum_{n \in \mathcal{B}} E_n(\mathbf{w}), \quad (5.71)$$

where each subset \mathcal{B} is called a *minibatch*. Before we start to train, we randomly assign the training samples into a fixed set of minibatches, each of which has a fixed size that commonly ranges from 32 at the low end to 8k at the higher end (Goyal, Dollár *et al.* 2017). The resulting algorithm is called *minibatch stochastic gradient descent*, although in practice, most people just call it SGD (omitting the reference to minibatches).³⁰

After evaluating the gradients $\mathbf{g} = \nabla_{\mathbf{w}} E_{\mathcal{B}}$ by summing over the samples in the minibatch, it is time to update the weights. The simplest way to do this is to take a small step in the gradient direction,

$$\mathbf{w} \leftarrow \mathbf{w} - \alpha \mathbf{g} \quad \text{or} \quad (5.72)$$

$$\mathbf{w}_{t+i} = \mathbf{w}_t - \alpha_t \mathbf{g}_t \quad (5.73)$$

where the first variant looks more like an assignment statement (see, e.g., Zhang, Lipton *et al.* 2021, Chapter 11; Loshchilov and Hutter 2019), while the second makes the temporal dependence explicit, using t to denote each successive step in the gradient descent.³¹

The step size parameter α is often called the *learning rate* and must be carefully adjusted to ensure good progress while avoiding overshooting and exploding gradients. In practice, it is common to start with a larger (but still small) learning rate α_t and to decrease it over time so that the optimization settles into a good minimum (Johnson 2020, Lecture 11; Zhang, Lipton *et al.* 2021, Chapter 11).

³⁰In the deep learning community, classic algorithms that sum over all the measurements are called *batch gradient descent*, although this term is not widely used elsewhere, as it is assumed that using all measurement at once is the preferred approach. In large-scale problems such as bundle adjustment, it's possible that using minibatches may result in better performance, but this has so far not been explored.

³¹I use the index k in discussing iterative algorithms in Appendix A.5.

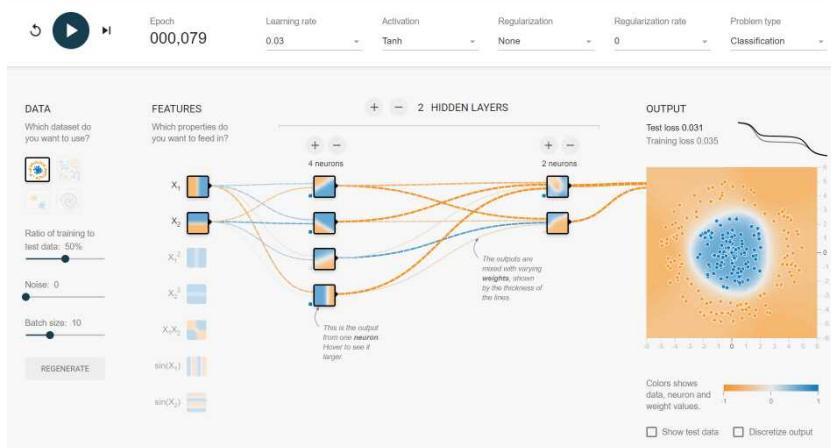


Figure 5.32 Screenshot from <http://playground.tensorflow.org>, where you can build and train your own small network in your web browser. Because the input space is two-dimensional, you can visualize the responses to all 2D inputs at each unit in the network.

Regular gradient descent is prone to stalling when the current solution reaches a “flat spot” in the search space, and stochastic gradient descent only pays attention to the errors in the current minibatch. For these reasons, the SGD algorithms may use the concept of *momentum*, where an exponentially decaying (“leaky”) running average of the gradients is accumulated and used as the update direction,

$$\mathbf{v}_{t+i} = \rho \mathbf{v}_t + \mathbf{g}_t \quad (5.74)$$

$$\mathbf{w}_{t+i} = \mathbf{w}_t - \alpha_t \mathbf{v}_t. \quad (5.75)$$

A relatively large value of $\rho \in [0.9, 0.99]$ is used to give the algorithm good memory, effectively averaging gradients over more batches.³²

Over the last decade, a number of more sophisticated optimization techniques have been applied to deep network training, as described in more detail in Johnson (2020, Lecture 11) and Zhang, Lipton *et al.* (2021, Chapter 11)). These algorithms include:

- *Nesterov momentum*, where the gradient is (effectively) computed at the state predicted from the velocity update;

³²Note that a recursive formula such as (5.74), which is the same as a temporal infinite impulse response filter (3.2.3) converges in the limit to a value of $\mathbf{g}/(1 - \rho)$, so α needs to be correspondingly adjusted.

- *AdaGrad* (Adaptive Gradient), where each component in the gradient is divided by the square root of the per-component summed squared gradients (Duchi, Hazan, and Singer 2011);
- *RMSProp*, where the running sum of squared gradients is replaced with a leaky (decaying) sum (Hinton 2012);
- *Adadelta*, which augments RMSProp with a leaky sum of the actual per-component changes in the parameters and uses these in the gradient re-scaling equation (Zeiler 2012);
- *Adam*, which combines elements of all the previous ideas into a single framework and also de-biases the initial leaky estimates (Kingma and Ba 2015); and
- *AdamW*, which is Adam with decoupled weight decay (Loshchilov and Hutter 2019).

Adam and AdamW are currently the most popular optimizers for deep networks, although even with all their sophistication, learning rates need to be set carefully (and probably decayed over time) to achieve good results. Setting the right *hyperparameters*, such as the learning rate initial value and decay rate, momentum terms such as ρ , and amount of regularization, so that the network achieves good performance within a reasonable training time is itself an open research area. The lecture notes by Johnson (2020, Lecture 11) provide some guidance, although in many cases, people perform a search over hyperparameters to find which ones produce the best performing network.

A simple two-input example

A great way to get some intuition on how deep networks update the weights and carve out a solution space during training is to play with the interactive visualization at <http://playground.tensorflow.org>.³³ As shown in Figure 5.32, just click the “run” (\triangleright) button to get started, then reset the network to a new start (button to the left of run) and try single-stepping the network, using different numbers of units per hidden layer and different activation functions. Especially when using ReLUs, you can see how the network carves out different parts of the input space and then combines these sub-pieces together. Section 5.4.5 discusses visualization tools to get insights into the behavior of larger, deeper networks.

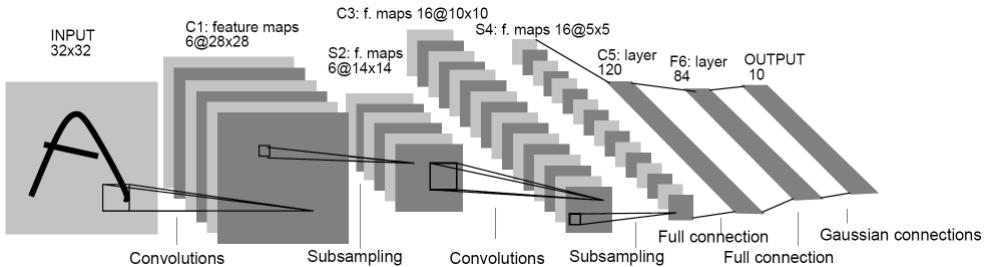


Figure 5.33 Architecture of LeNet-5, a convolutional neural network for digit recognition (LeCun, Bottou *et al.* 1998) © 1998 IEEE. This network uses multiple channels in each layer and alternates multi-channel convolutions with downsampling operations, followed by some fully connected layers that produce one activation for each of the 10 digits being classified.

5.4 Convolutional neural networks

The previous sections on deep learning have covered all of the essential elements of constructing and training deep networks. However, they have omitted what is likely the most crucial component of deep networks for image processing and computer vision, which is the use of trainable multi-layer convolutions. The idea of convolutional neural networks was popularized by LeCun, Bottou *et al.* (1998), where they introduced the LeNet-5 network for digit recognition shown in Figure 5.33.³⁴

Instead of connecting all of the units in a layer to all the units in a preceding layer, convolutional networks organize each layer into *feature maps* (LeCun, Bottou *et al.* 1998), which you can think of as parallel planes or *channels*, as shown in Figure 5.33. In a convolutional layer, the weighted sums are only performed within a small local window, and weights are identical for all pixels, just as in regular shift-invariant image convolution and correlation (3.12–3.15).

Unlike image convolution, however, where the same filter is applied to each (color) channel, neural network convolutions typically linearly combine the activations from each of the C_1 input channels in a previous layer and use different convolution kernels for each of the C_2 output channels, as shown in Figures 5.34–5.35.³⁵ This makes sense, as the main task in

³³Additional informative interactive demonstrations can be found at <https://cs.stanford.edu/people/karpathy/convnetjs>.

³⁴A similar convolutional architecture, but without the gradient descent training procedure, was earlier proposed by Fukushima (1980).

³⁵The number of channels in a given network layer is sometimes called its *depth*, but the number of layers in a deep network is also called its depth. So, be careful when reading network descriptions.

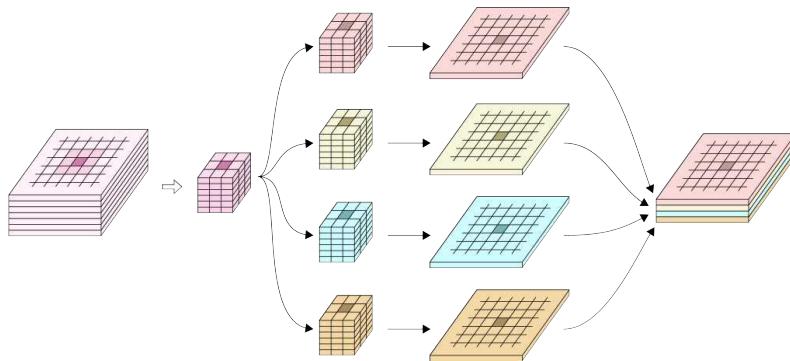


Figure 5.34 2D convolution with multiple input and output channels © Glassner (2018). Each 2D convolution kernel takes as input all of the C_1 channels in the preceding layer, windowed to a small area, and produces the values (after the activation function non-linearity) in one of the C_2 channels in the next layer. For each of the output channels, we have $S^2 \times C_1$ kernel weights, so the total number of learnable parameters in each convolutional layer is $S^2 \times C_1 \times C_2$. In this figure, we have $C_1 = 6$ input channels and $C_2 = 4$ output channels, with an $S = 3$ convolution window, for a total of $9 \times 6 \times 4$ learnable weights, shown in the middle column of the figure. Since the convolution is applied at each of the $W \times H$ pixels in a given layer, the amount of computation (multiply-adds) in each forward and backward pass over one sample in a given layer is $WHS^2C_1C_2$.

convolutional neural network layers is to construct local features (Figure 3.40c) and to then combine them in different ways to produce more discriminative and semantically meaningful features.³⁶ Visualizations of the kinds of features that deep networks extract are shown in Figure 5.47 in Section 5.4.5.

With these intuitions in place, we can write the weighted linear sums (5.45) performed in a convolutional layer as

$$s(i, j, c_2) = \sum_{c_1 \in \{C_1\}} \sum_{(k, l) \in \mathcal{N}} w(k, l, c_1, c_2)x(i + k, j + l, c_1) + b(c_2), \quad (5.76)$$

where the $x(i, j, c_1)$ are the activations in the previous layer, just as in (5.45), \mathcal{N} are the S^2 signed offsets in the 2D spatial kernel, and the notation $c_1 \in \{C_1\}$ denotes $c_1 \in [0, C_1)$. Note that because the offsets (k, l) are added to (instead of subtracted from) the (i, j) pixel coordinates, this operation is actually a *correlation* (3.13), but this distinction is usually glossed

³⁶Note that pixels in different input and output channels (within the convolution window size) are fully connected, unless grouped convolutions, discussed below, are used.

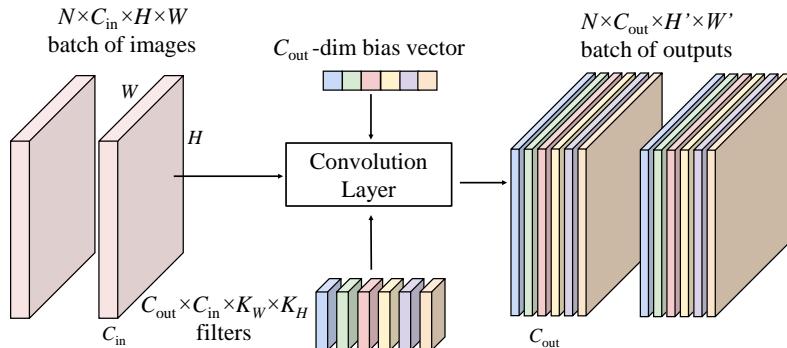


Figure 5.35 2D convolution with multiple batches, input, and output channels, © Johnson (2020). When doing mini-batch gradient descent, a whole batch of training images or features is passed into a convolutional layer, which takes as input all of the C_{in} channels in the preceding layer, windowed to a small area, and produces the values (after the activation function non-linearity) in one of the C_{out} channels in the next layer. As before, for each of the output channels, we have $K_w \times K_h \times C_{\text{in}}$ kernel weights, so the total number of learnable parameters in each convolutional layer is $K_w \times K_h \times C_{\text{in}} \times C_{\text{in}}$. In this figure, we have $C_{\text{in}} = 3$ input channels and $C_{\text{out}} = 6$ output channels.

over.³⁷

In neural network diagrams such as those shown in Figures 5.33 and 5.39–5.43, it is common to indicate the convolution kernel size S and the number of channels in a layer C , and only sometimes to show the image dimensions, as in Figures 5.33 and 5.39. Note that some neural networks such as the Inception module in GoogLeNet (Szegedy, Liu *et al.* 2015) shown in Figure 5.42 use 1×1 convolutions, which do not actually perform convolutions but rather combine various channels on a per-pixel basis, often with the goal of reducing the dimensionality of the feature space.

Because the weights in a convolution kernel are the same for all of the pixels within a given layer and channel, these weights are actually *shared* across what would result if we drew all of the connections between different pixels in different layers. This means that there are many fewer weights to learn than in fully connected layers. It also means that during backpropagation, kernel weight updates are summed over all of the pixels in a given layer/channel.

To fully determine the behavior of a convolutional layer, we still need to specify a few

³⁷Since the weights in a neural network are learned, this reversal does not really matter.

additional parameters.³⁸ These include:

- *Padding*. Early networks such as LeNet-5 did not pad the image, which therefore shrank after each convolution. Modern networks can optionally specify a padding width and mode, using one of the choices used with traditional image processing, such as zero padding or pixel replication, as shown in Figure 3.13.
- *Stride*. The default stride for convolution is 1 pixel, but it is also possible to only evaluate the convolution at every n th column and row. For example, the first convolution layer in AlexNet (Figure 5.39) uses a stride of 4. Traditional image pyramids (Figure 3.31) use a stride of 2 when constructing the coarser levels.
- *Dilation*. Extra “space” (skipped rows and column) can be inserted between pixel samples during convolution, also known as dilated or *à trous* (with holes, in French, or often just “atrous”) convolution (Yu and Koltun 2016; Chen, Papandreou *et al.* 2018). While in principle this can lead to aliasing, it can also be effective at pooling over a larger region while using fewer operations and learnable parameters.
- *Grouping*. While, by default, all input channels are used to produce each output channel, we can also group the input and output layers into G separate groups, each of which is convolved separately (Xie, Girshick *et al.* 2017). $G = 1$ corresponds to regular convolution, while $G = C_1$ means that each corresponding input channel is convolved independently from the others, which is known as *depthwise* or *channel-separated* convolution (Howard, Zhu *et al.* 2017; Tran, Wang *et al.* 2019).

A nice animation of the effects of these different parameters created by Vincent Dumoulin can be found at https://github.com/vdumoulin/conv_arithmetic as well as Dumoulin and Visin (2016).

In certain applications such as image inpainting (Section 10.5.1), the input image may come with an associated binary *mask*, indicating which pixels are valid and which need to be filled in. This is similar to the concept of alpha-matted images we studied in Section 3.1.3. In this case, one can use *partial convolutions* (Liu, Reda *et al.* 2018), where the input pixels are multiplied by the mask pixels and then normalized by the count of non-zero mask pixels. The mask channel output is set to 1 if any input mask pixels are non-zero. This resembles the pull-push algorithm of Gortler, Grzeszczuk *et al.* (1996) that we presented in Figure 4.2, except that the convolution weights are learned.

³⁸Most of the neural network building blocks we present in this chapter have corresponding functions in widely used deep learning frameworks, where you can get more detailed information about their operation. For example, the 2D convolution operator is called *Conv2d* in PyTorch and is documented at <https://pytorch.org/docs/stable/nn.html#convolution-layers>.

A more sophisticated version of partial convolutions is *gated convolutions* (Yu, Lin *et al.* 2019; Chang, Liu *et al.* 2019), where the per-pixel masks are derived from the previous layer using a learned convolution followed by a sigmoid non-linearity. This enables the network not only to learn a better measure of per-pixel confidence (weighting), but also to incorporate additional features such as user-drawn sketches or derived semantic information.

5.4.1 Pooling and unpooling

As we just saw in the discussion of convolution, strides of greater than 1 can be used to reduce the resolution of a given layer, as in the first convolutional layer of AlexNet (Figure 5.39). When the weights inside the convolution kernel are identical and sum up to 1, this is called *average pooling* and is typically applied in a channel-wise manner.

A widely used variant is to compute the maximum response within a square window, which is called *max pooling*. Common strides and window sizes for max pooling are a stride of 2 and 2×2 non-overlapping windows or 3×3 overlapping windows. Max pooling layers can be thought of as a “logical or”, since they only require one of the units in the pooling region to be turned on. They are also supposed to provide some shift invariance over the inputs. However, most deep networks are not all that shift-invariant, which degrades their performance. The paper by Zhang (2019) has a nice discussion of this issue and some simple suggestions to mitigate this problem.

One issue that commonly comes up is how to backpropagate through a max pooling layer. The max pool operator acts like a “switch” that shunts (connects) one of the input units to the output unit. Therefore, during backpropagation, we only need to pass the error and derivatives down to this maximally active unit, as long as we have remembered which unit has this response.

This same *max unpooling* mechanism can be used to create a “deconvolution network” when searching for the stimulus (Figure 5.47) that most strongly activates a particular unit (Zeiler and Fergus 2014).

If we want a more continuous behavior, we could construct a pooling unit that computes an L_p norm over its inputs, since the $L_{p \rightarrow \infty}$ effectively computes a maximum over its components (Springenberg, Dosovitskiy *et al.* 2015). However, such a unit requires more computation, so it is not widely used in practice, except sometimes at the final layer, where it is known as *generalized mean (GeM) pooling* (Dollár, Tu *et al.* 2009; Tolias, Sicre, and Jégou 2016; Gordo, Almazán *et al.* 2017; Radenović, Tolias, and Chum 2019) or *dynamic mean (DAME) pooling* (Yang, Kien Nguyen *et al.* 2019). In their paper, Springenberg, Dosovitskiy *et al.* (2015) also show that using strided convolution instead of max pooling can produce competitive results.

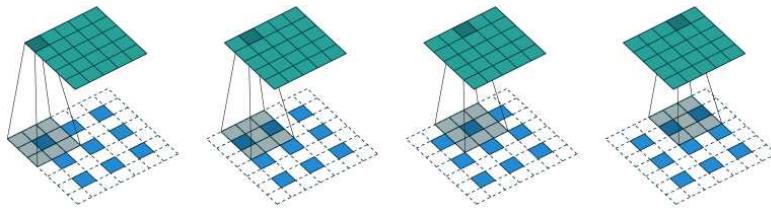


Figure 5.36 Transposed convolution (© Dumoulin and Visin (2016)) can be used to upsample (increase the size of) an image. Before applying the convolution operator, $(s - 1)$ extra rows and columns of zeros are inserted between the input samples, where s is the upsampling stride.

While unpooling can be used to (approximately) reverse the effect of max pooling operation, if we want to reverse a convolutional layer, we can look at learned variants of the interpolation operator we studied in Sections 3.5.1 and 3.5.3. The easiest way to visualize this operation is to add extra rows and columns of zeros between the pixels in the input layer, and to then run a regular convolution (Figure 5.36). This operation is sometimes called *backward convolution* with a *fractional stride* (Long, Shelhamer, and Darrell 2015), although it is more commonly known as *transposed convolution* (Dumoulin and Visin 2016), because when convolutions are written in matrix form, this operation is a multiplication with a transposed sparse weight matrix. Just as with regular convolution, padding, stride, dilation, and grouping parameters can be specified. However, in this case, the stride specifies the factor by which the image will be upsampled instead of downsampled.

U-Nets and Feature Pyramid Networks

When discussing the Laplacian pyramid in Section 3.5.3, we saw how image downsampling and upsampling can be combined to achieve a variety of multi-resolution image processing tasks (Figure 3.33). The same kinds of combinations can be used in deep convolutional networks, in particular, when we want the output to be a full-resolution image. Examples of such applications include pixel-wise semantic labeling (Section 6.4), image denoising and super-resolution (Section 10.3), monocular depth inference (Section 12.8), and neural style transfer (Section 14.6). The idea of reducing the resolution of a network and then expanding it again is sometimes called a *bottleneck* and is related to earlier self-supervised network training using *autoencoders* (Hinton and Zemel 1994; Goodfellow, Bengio, and Courville 2016, Chapter 14).

One of the earliest applications of this idea was the *fully convolutional network* developed

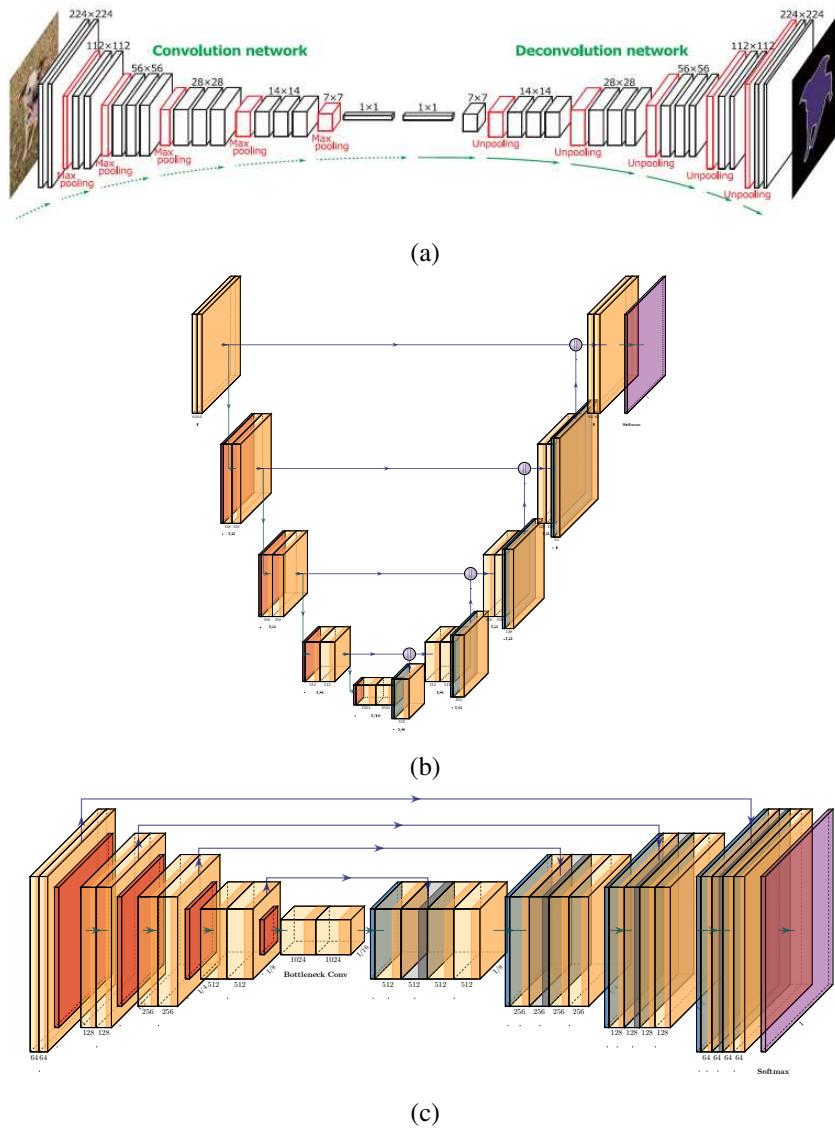


Figure 5.37 (a) The deconvolution network of Noh, Hong, and Han (2015) © 2015 IEEE and (b–c) the U-Net of Ronneberger, Fischer, and Brox (2015), drawn using the PlotNeural-Net LaTeX package. In addition to the fine-to-coarse-to-fine bottleneck used in (a), the U-Net also has skip connections between encoding and decoding layers at the same resolution.



Figure 5.38 Screenshot from Andrej Karpathy’s web browser demos at <https://cs.stanford.edu/people/karpathy/convnetjs>, where you can run a number of small neural networks, including CNNs for digit and tiny image classification.

by Long, Shelhamer, and Darrell (2015). This paper inspired myriad follow-on architectures, including the hourglass-shaped “deconvolution” network of Noh, Hong, and Han (2015), the U-Net of Ronneberger, Fischer, and Brox (2015), the atrous convolution network with CRF refinement layer of Chen, Papandreou *et al.* (2018), and the panoptic feature pyramid networks of Kirillov, Girshick *et al.* (2019). Figure 5.37 shows the general layout of two of these networks, which are discussed in more detail in Section 6.4 on semantic segmentation. We will see other uses of these kinds of *backbone networks* (He, Gkioxari *et al.* 2017) in later sections on image denoising and super-resolution (Section 10.3), monocular depth inference (Section 12.8), and neural style transfer (Section 14.6).

5.4.2 Application: Digit classification

One of the earliest commercial application of convolutional neural networks was the LeNet-5 system created by LeCun, Bottou *et al.* (1998) whose architecture is shown in Figure 5.33. This network contained most of the elements of modern CNNs, although it used sigmoid non-linearities, average pooling, and Gaussian RBF units instead of softmax at its output. If you want to experiment with this simple digit recognition CNN, you can visit the interactive JavaScript demo created by Andrej Karpathy at <https://cs.stanford.edu/people/karpathy/convnetjs> (Figure 5.38).

The network was initially deployed around 1995 by AT&T to automatically read checks deposited in NCR ATM machines to verify that the written and keyed check amounts were the same. The system was then incorporated into NCR’s high-speed check reading systems,

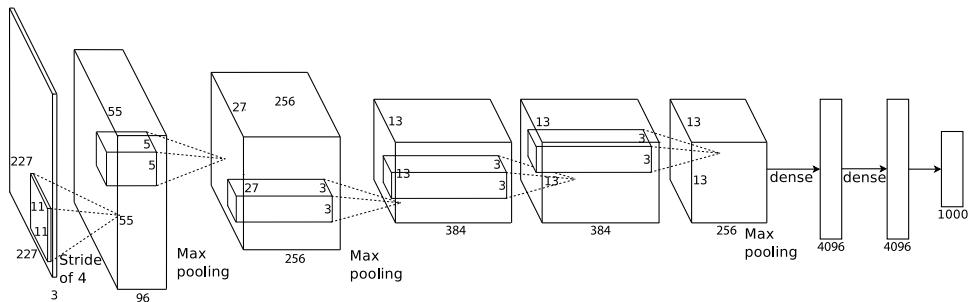


Figure 5.39 Architecture of the SuperVision deep neural network (more commonly known as “AlexNet”), courtesy of Matt Deitke (redrawn from (Krizhevsky, Sutskever, and Hinton 2012)). The network consists of multiple convolutional layers with ReLU activations, max pooling, some fully connected layers, and a softmax to produce the final class probabilities.

which at some point were processing somewhere between 10% and 20% of all the checks in the US.³⁹

Today, variants of the LeNet-5 architecture (Figure 5.33) are commonly used as the first convolutional neural network introduced in courses and tutorials on the subject.⁴⁰ Although the MNIST dataset (LeCun, Cortes, and Burges 1998) originally used to train LeNet-5 is still sometimes used, it is more common to use the more challenging CIFAR-10 (Krizhevsky 2009) or Fashion MNIST (Xiao, Rasul, and Vollgraf 2017) as datasets for training and testing.

5.4.3 Network architectures

While modern convolutional neural networks were first developed and deployed in the late 1990s, it was not until the breakthrough publication by Krizhevsky, Sutskever, and Hinton (2012) that they started outperforming more traditional techniques on natural image classification (Figure 5.40). As you can see in this figure, the AlexNet system (the more widely used name for their SuperVision network) led to a dramatic drop in error rates from 25.8% to 16.4%. This was rapidly followed in the next few years with additional dramatic performance improvements, due to further developments as well as the use of deeper networks, e.g., from the original 8-layer AlexNet to a 152-layer ResNet.

Figure 5.39 shows the architecture of the SuperVision network, which contains a series of convolutional layers with ReLU (rectified linear) non-linearities, max pooling, some fully

³⁹This information courtesy of Yann LeCun and Larry Jackel, who were two of the principals in the development of this system.

⁴⁰See, e.g., https://pytorch.org/tutorials/beginner/blitz/cifar10_tutorial.html.

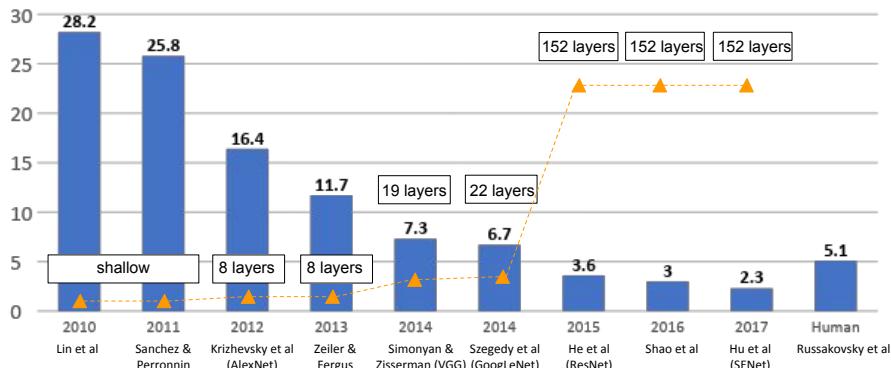


Figure 5.40 Top-5 error rate and network depths of winning entries from the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) © Li, Johnson, and Yeung (2019).

connected layers, and a final softmax layer, which is fed into a multi-class cross-entropy loss. Krizhevsky, Sutskever, and Hinton (2012) also used dropout (Figure 5.29), small translation and color manipulation for data augmentation, momentum, and weight decay (L_2 weight penalties).

The next few years after the publication of this paper saw dramatic improvement in the classification performance on the ImageNet Large Scale Visual Recognition Challenge (Russakovsky, Deng *et al.* 2015), as shown in Figure 5.40. A nice description of the innovations in these various networks, as well as their capacities and computational cost, can be found in the lecture slides by Justin Johnson (2020, Lecture 8).

The winning entry from 2013 by Zeiler and Fergus (2014) used a larger version of AlexNet with more channels in the convolution stages and lowered the error rate by about 30%. The 2014 Oxford Visual Geometry Group (VGG) winning entry by Simonyan and Zisserman (2014b) used repeated 3×3 convolution/ReLU blocks interspersed with 2×2 max pooling and channel doubling (Figure 5.41), followed by some fully connected layers, to produce 16–19 layer networks that further reduced the error by 40%. However, as shown in Figure 5.44, this increased performance came at a greatly increased amount of computation.

The 2015 GoogLeNet of Szegedy, Liu *et al.* (2015) focused instead on efficiency. GoogLeNet begins with an aggressive stem network that uses a series of strided and regular convolutions and max pool layers to quickly reduce the image resolutions from 224^2 to 28^2 . It then uses a number of *Inception modules* (Figure 5.42), each of which is a small branching neural network whose features get concatenated at the end. One of the important characteristics of this module is that it uses 1×1 “bottleneck” convolutions to reduce the number of

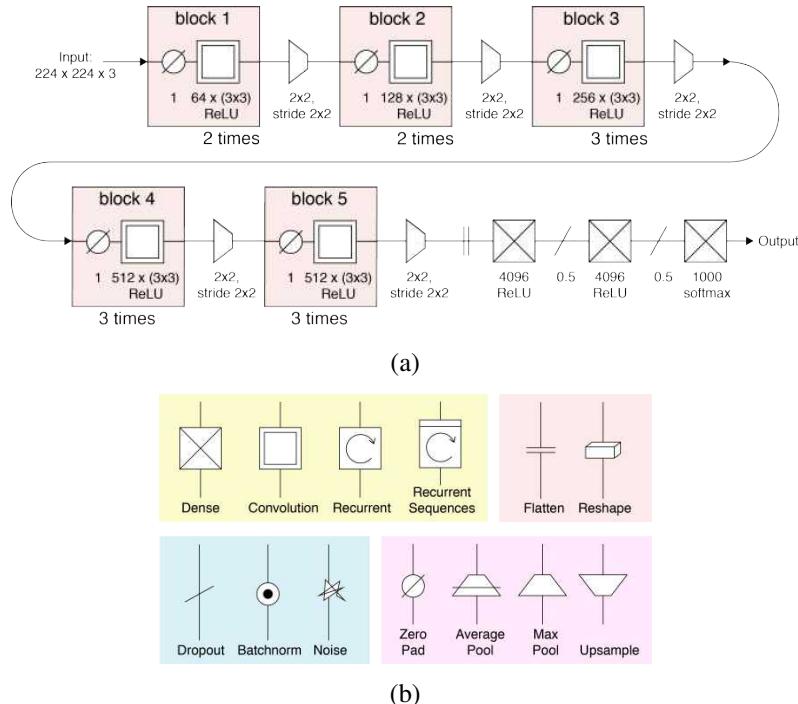


Figure 5.41 The VGG16 network of Simonyan and Zisserman (2014b) © Glassner (2018).
 (a) The network consists of repeated zero-pad, 3×3 convolution, ReLU blocks interspersed with 2×2 max pooling and a doubling in the number of channels. This is followed by some fully connected and dropout layers, with a final softmax into the 1,000 Imagenet categories.
 (b) Some of the schematic neural network symbols used by Glassner (2018).

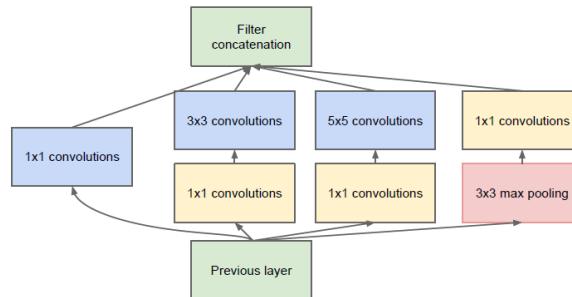


Figure 5.42 An Inception module from (Szegedy, Liu et al. 2015) © 2015 IEEE, which combines dimensionality reduction, multiple convolution sizes, and max pooling as different channels that get stacked together into a final feature map.

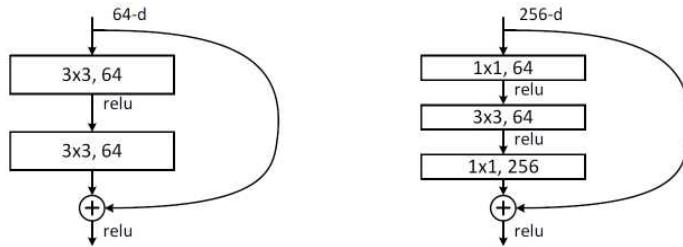


Figure 5.43 ResNet residual networks (He, Zhang et al. 2016a) © 2016 IEEE, showing skip connections going around a series of convolutional layers. The figure on the right uses a bottleneck to reduce the number of channels before the convolution. Having direct connections that shortcut the convolutional layer allows gradients to more easily flow backward through the network during training.

channels before performing larger 3×3 and 5×5 convolutions, thereby saving a significant amount of computation. This kind of projection followed by an additional convolution is similar in spirit to the approximation of filters as a sum of separable convolutions proposed by Perona (1995). GoogLeNet also removed the fully connected (MLP) layers at the end, relying instead on global average pooling followed by one linear layer before the softmax. Its performance was similar to that of VGG but at dramatically lower computation and model size costs (Figure 5.44).

The following year saw the introduction of Residual Networks (He, Zhang et al. 2016a), which dramatically expanded the number of layers that could be successfully trained (Figure 5.40). The main technical innovation was the introduction of *skip connections* (originally called “shortcut connections”), which allow information (and gradients) to flow around a set of convolutional layers, as shown in Figure 5.43. The networks are called *residual networks* because they allow the network to learn the residuals (differences) between a set of incoming and outgoing activations. A variant on the basic residual block is the “bottleneck block” shown on the right side of Figure 5.43, which reduces the number of channels before performing the 3×3 convolutional layer. A further extension, described in (He, Zhang et al. 2016b), moves the ReLU non-linearity to before the residual summation, thereby allowing true identity mappings to be modeled at no cost.

To build a ResNet, various residual blocks are interspersed with strided convolutions and channel doubling to achieve the desired number of layers. (Similar downsampling stems and average pooled softmax layers as in GoogLeNet are used at the beginning and end.) By combining various numbers of residual blocks, ResNets consisting of 18, 34, 50, 101, and 152 layers have been constructed and evaluated. The deeper networks have higher accuracy

but more computational cost (Figure 5.44). In 2015, ResNet not only took first place in the ILSVRC (ImageNet) classification, detection, and localization challenges, but also took first place in the detection and segmentation challenges on the newer COCO dataset and benchmark (Lin, Maire *et al.* 2014).

Since then, myriad extensions and variants have been constructed and evaluated. The ResNeXt system from Xie, Girshick *et al.* (2017) used grouped convolutions to slightly improve accuracy. Denseley connected CNNs (Huang, Liu *et al.* 2017) added skip connections between non-adjacent convolution and/or pool blocks. Finally, the Squeeze-and-Excitation network (SENet) by Hu, Shen, and Sun (2018) added global context (via global pooling) to each layer to obtain a noticeable increase in accuracy. More information about these and other CNN architectures can be found in both the original papers as well as class notes on this topic (Li, Johnson, and Yeung 2019; Johnson 2020).

Mobile networks

As deep neural networks were getting deeper and larger, a countervailing trend emerged in the construction of smaller, less computationally expensive networks that could be used in mobile and embedded applications. One of the earliest networks tailored for lighter-weight execution was MobileNets (Howard, Zhu *et al.* 2017), which used *depthwise convolutions*, a special case of grouped convolutions where the number of groups equals the number of channels. By varying two hyperparameters, namely a *width multiplier* and a *resolution multiplier*, the network architecture could be tuned along an accuracy vs. size vs. computational efficiency tradeoff. The follow-on MobileNetV2 system (Sandler, Howard *et al.* 2018) added an “inverted residual structure”, where the shortcut connections were between the bottleneck layers. ShuffleNet (Zhang, Zhou *et al.* 2018) added a “shuffle” stage between grouped convolutions to enable channels in different groups to co-mingle. ShuffleNet V2 (Ma, Zhang *et al.* 2018) added a channel split operator and tuned the network architectures using end-to-end performance measures. Two additional networks designed for computational efficiency are ESPNet (Mehta, Rastegari *et al.* 2018) and ESPNetv2 (Mehta, Rastegari *et al.* 2019), which use pyramids of (depth-wide) dilated separable convolutions.

The concepts of grouped, depthwise, and channel-separated convolutions continue to be a widely used tool for managing computational efficiency and model size (Choudhary, Mishra *et al.* 2020), not only in mobile networks, but also in video classification (Tran, Wang *et al.* 2019), which we discuss in more detail in Section 5.5.2.

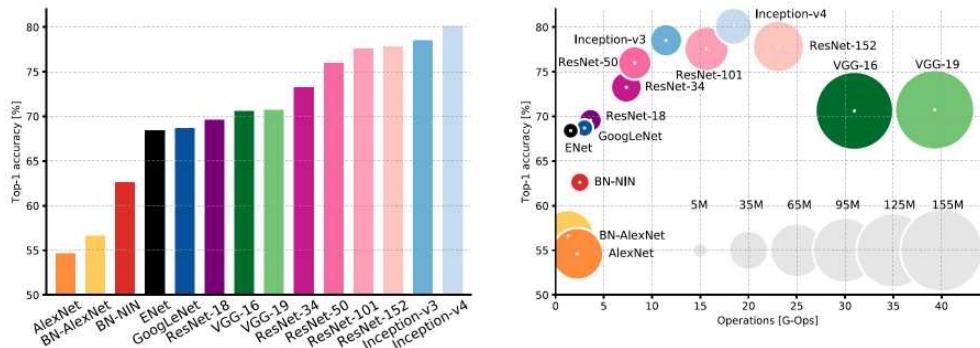


Figure 5.44 Network accuracy vs. size and operation counts (Canziani, Culurciello, and Paszke 2017) © 2017 IEEE: In the right figure, the network accuracy is plotted against operation count (1–40 G-Ops), while the size of the circle indicates the number of parameters (10–155 M). The initials BN indicate a batch normalized version of a network.

5.4.4 Model zoos

A great way to experiment with these various CNN architectures is to download pre-trained models from a *model zoo*⁴¹ such as the TorchVision library at <https://github.com/pytorch/vision>. If you look in the `torchvision/models` folder, you will find implementations of AlexNet, VGG, GoogleNet, Inception, ResNet, DenseNet, MobileNet, and ShuffleNet, along with other models for classification, object detection, and image segmentation. Even more recent models, some of which are discussed in the upcoming sections, can be found in the PyTorch Image Models library (timm), <https://github.com/rwightman/pytorch-image-models>. Similar collections of pre-trained models exist for other languages, e.g., <https://www.tensorflow.org/lite/models> for efficient (mobile) TensorFlow models.

While people often download and use pre-trained neural networks for their applications, it is more common to at least *fine-tune* such networks on data more characteristic of the application (as opposed to the public benchmark data on which most zoo models are trained).⁴² It is also quite common to replace the last few layers, i.e., the *head* of the network (so called because it lies at the top of a layer diagram when the layers are stacked bottom-to-top) while leaving the *backbone* intact. The terms *backbone* and *head(s)* are widely used and were popularized by the Mask-RCNN paper (He, Gkioxari *et al.* 2017). Some more recent papers

⁴¹The name “model zoo” itself is a fanciful invention of Evan Shelhamer, lead developer on Caffe (Jia, Shelhamer *et al.* 2014), who first used it on https://caffe.berkeleyvision.org/model_zoo.html to describe a collection of various pre-trained DNN models (personal communication).

⁴²See, e.g., https://classyvision.ai/tutorials/fine_tuning and (Zhang, Lipton *et al.* 2021, Section 13.2).

refer to the backbone and head as the *trunk* and its *branches* (Ding and Tao 2018; Kirillov, Girshick *et al.* 2019; Bell, Liu *et al.* 2020), with the term *neck* also being occasionally used (Chen, Wang *et al.* 2019).⁴³

When adding a new head, its parameters can be trained using the new data specific to the intended application. Depending on the amount and quality of new training data available, the head can be as simple as a linear model such as an SVM or logistic regression/softmax (Donahue, Jia *et al.* 2014; Sharif Razavian, Azizpour *et al.* 2014), or as complex as a fully connected or convolutional network (Xiao, Liu *et al.* 2018). Fine-tuning some of the layers in the backbone is also an option, but requires sufficient data and a slower learning rate so that the benefits of the pre-training are not lost. The process of pre-training a machine learning system on one dataset and then applying it to another domain is called *transfer learning* (Pan and Yang 2009). We will take a closer look at transfer learning in Section 5.4.7 on self-supervised learning.

Model size and efficiency

As you can tell from the previous discussion, neural network models come in a large variety of sizes (typically measured in number of parameters, i.e., weights and biases) and computational loads (often measured in FLOPs per forward inference pass). The evaluation by Canziani, Culurciello, and Paszke (2017), summarized in Figure 5.44, gives a snapshot of the performance (accuracy and size+operations) of the top-performing networks on the ImageNet challenge from 2012–2017. In addition to the networks we have already discussed, the study includes Inception-v3 (Szegedy, Vanhoucke *et al.* 2016) and Inception-v4 (Szegedy, Ioffe *et al.* 2017).

Because deep neural networks can be so memory- and compute-intensive, a number of researchers have investigated methods to reduce both, using lower precision (e.g., fixed-point) arithmetic and weight compression (Han, Mao, and Dally 2016; Iandola, Han *et al.* 2016). The XNOR-Net paper by Rastegari, Ordonez *et al.* (2016) investigates using binary weights (on-off connections) and optionally binary activations. It also has a nice review of previous binary networks and other compression techniques, as do more recent survey papers (Sze, Chen *et al.* 2017; Gu, Wang *et al.* 2018; Choudhary, Mishra *et al.* 2020).

Neural Architecture Search (NAS)

One of the most recent trends in neural network design is the use of *Neural Architecture Search* (NAS) algorithms to try different network topologies and parameterizations (Zoph

⁴³Classy Vision uses a trunk and heads terminology, https://classyvision.ai/tutorials/classy_model.

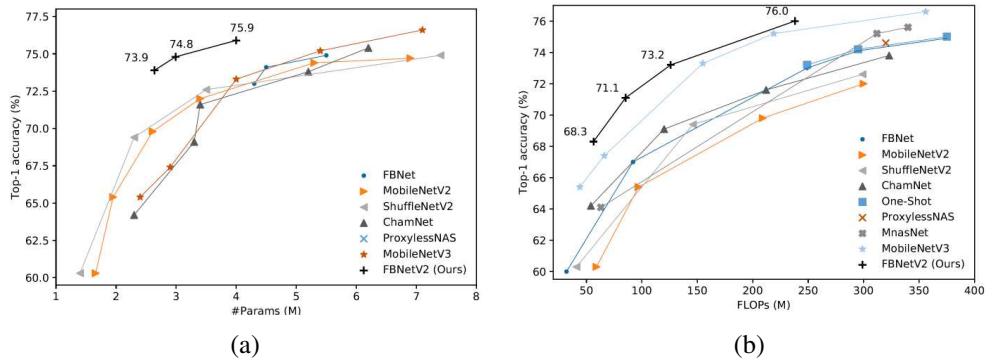


Figure 5.45 ImageNet accuracy vs. (a) size (# of parameters) and (b) operation counts for a number of recent efficient networks (Wan, Dai et al. 2020) © 2020 IEEE.

and Le 2017; Zoph, Vasudevan et al. 2018; Liu, Zoph et al. 2018; Pham, Guan et al. 2018; Liu, Simonyan, and Yang 2019; Hutter, Kotthoff, and Vanschoren 2019). This process is more efficient (in terms of a researcher’s time) than the trial-and-error approach that characterized earlier network design. Elsken, Metzen, and Hutter (2019) survey these and additional papers on this rapidly evolving topic. More recent publications include FBNet (Wu, Dai et al. 2019), RandomNets (Xie, Kirillov et al. 2019), EfficientNet (Tan and Le 2019), RegNet (Radosavovic, Kosaraju et al. 2020), FBNetV2 (Wan, Dai et al. 2020), and EfficientNetV2 (Tan and Le 2021). It is also possible to do unsupervised neural architecture search (Liu, Dollár et al. 2020). Figure 5.45 shows the top-1% accuracy on ImageNet vs. the network size (# of parameters) and forward inference operation counts for a number of recent network architectures (Wan, Dai et al. 2020). Compared to the earlier networks shown in Figure 5.44, the newer networks use 10× (or more) fewer parameters.

Deep learning software

Over the last decade, a large number of deep learning software frameworks and programming language extensions have been developed. The Wikipedia entry on deep learning software lists over twenty such frameworks, about a half of which are still being actively developed.⁴⁴ While Caffe (Jia, Shelhamer et al. 2014) was one of the first to be developed and used for computer vision applications, it has mostly been supplanted by PyTorch and TensorFlow, at least if we judge by the open-source implementations that now accompany most computer vision research papers.

⁴⁴https://en.wikipedia.org/wiki/Comparison_of_deep-learning_software

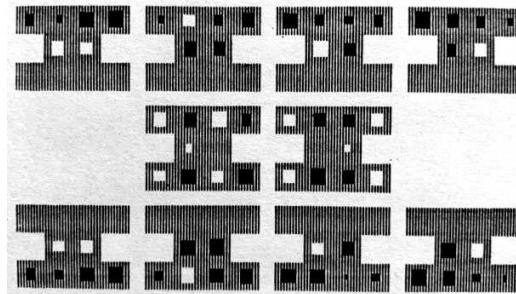


Figure 5.46 A Hinton diagram showing the weights connecting the units in a three layer neural network, courtesy of Geoffrey Hinton. The size of each small box indicates the magnitude of each weight and its color (black or white) indicates the sign.

Andrew Glassner's (2018) introductory deep learning book uses the Keras library because of its simplicity. The Dive into Deep Learning book (Zhang, Lipton *et al.* 2021) and associated course (Smola and Li 2019) use MXNet for all the examples in the text, but they have recently released PyTorch and TensorFlow code samples as well. Stanford's CS231n (Li, Johnson, and Yeung 2019) and Johnson (2020) include a lecture on the fundamentals of PyTorch and TensorFlow. Some classes also use simplified frameworks that require the students to implement more components, such as the Educational Framework (EDF) developed by McAllester (2020) and used in Geiger (2021).

In addition to software frameworks and libraries, deep learning code development usually benefits from good visualization libraries such as TensorBoard⁴⁵ and Visdom.⁴⁶ And in addition to the model zoos mentioned earlier in this section, there are even higher-level packages such as Classy Vision,⁴⁷ which allow you to train or fine-tune your own classifier with no or minimal programming. Andrej Karpathy also provides a useful guide for training neural networks at <http://karpathy.github.io/2019/04/25/recipe>, which may help avoid common issues.

5.4.5 Visualizing weights and activations

Visualizing intermediate and final results has always been an integral part of computer vision algorithm development (e.g., Figures 1.7–1.11) and is an excellent way to develop intuitions and debug or refine results. In this chapter, we have already seen examples of tools for simple

⁴⁵<https://www.tensorflow.org/tensorboard>

⁴⁶<https://github.com/fossasia/visdom>

⁴⁷<https://classyvision.ai>

two-input neural network visualizations, e.g., the TensorFlow Playground in Figure 5.32 and ConvNetJS in Figure 5.38. In this section, we discuss tools for visualizing network weights and, more importantly, the *response functions* of different units or layers in a network.

For a simple small network such as the one shown in Figure 5.32, we can indicate the strengths of connections using line widths and colors. What about networks with more units? A clever way to do this, called *Hinton diagrams* in honor of its inventor, is to indicate the strengths of the incoming and outgoing weights as black or white boxes of different sizes, as shown in Figure 5.46 (Ackley, Hinton, and Sejnowski 1985; Rumelhart, Hinton, and Williams 1986b).⁴⁸

If we wish to display the set of activations in a given layer, e.g., the response of the final 10-category layer in MNIST or CIFAR-10, across some or all of the inputs, we can use non-linear dimensionality reduction techniques such as t-SNE and UMap discussed in Section 5.2.4 and Figure 5.21.

How can we visualize what individual units (“neurons”) in a deep network respond to? For the first layer in a network (Figure 5.47, upper left corner), the response can be read directly from the incoming weights (grayish images) for a given channel. We can also find the patches in the validation set that produce the largest responses across the units in a given channel (colorful patches in the upper left corner of Figure 5.47). (Remember that in a convolutional neural network, different units in a particular channel respond similarly to shifted versions of the input, ignoring boundary and aliasing effects.)

For deeper layers in a network, we can again find maximally activating patches in the input images. Once these are found, Zeiler and Fergus (2014) pair a *deconvolution network* with the original network to backpropagate feature activations all the way back to the image patch, which results in the grayish images in layers 2–5 in Figure 5.47. A related technique called *guided backpropagation* developed by Springenberg, Dosovitskiy *et al.* (2015) produces slightly higher contrast results.

Another way to probe a CNN feature map is to determine how strongly parts of an input image activate units in a given channel. Zeiler and Fergus (2014) do this by masking sub-regions of the input image with a gray square, which not only produces activation maps, but can also show the most likely labels associated with each image region (Figure 5.48). Simonyan, Vedaldi, and Zisserman (2013) describe a related technique they call *saliency maps*, Nguyen, Yosinski, and Clune (2016) call their related technique *activation maximization*, and Selvaraju, Cogswell *et al.* (2017) call their visualization technique *gradient-weighted class activation mapping* (Grad-CAM).

⁴⁸In the early days of neural networks, bit-mapped displays and printers only supported 1-bit black and white images.

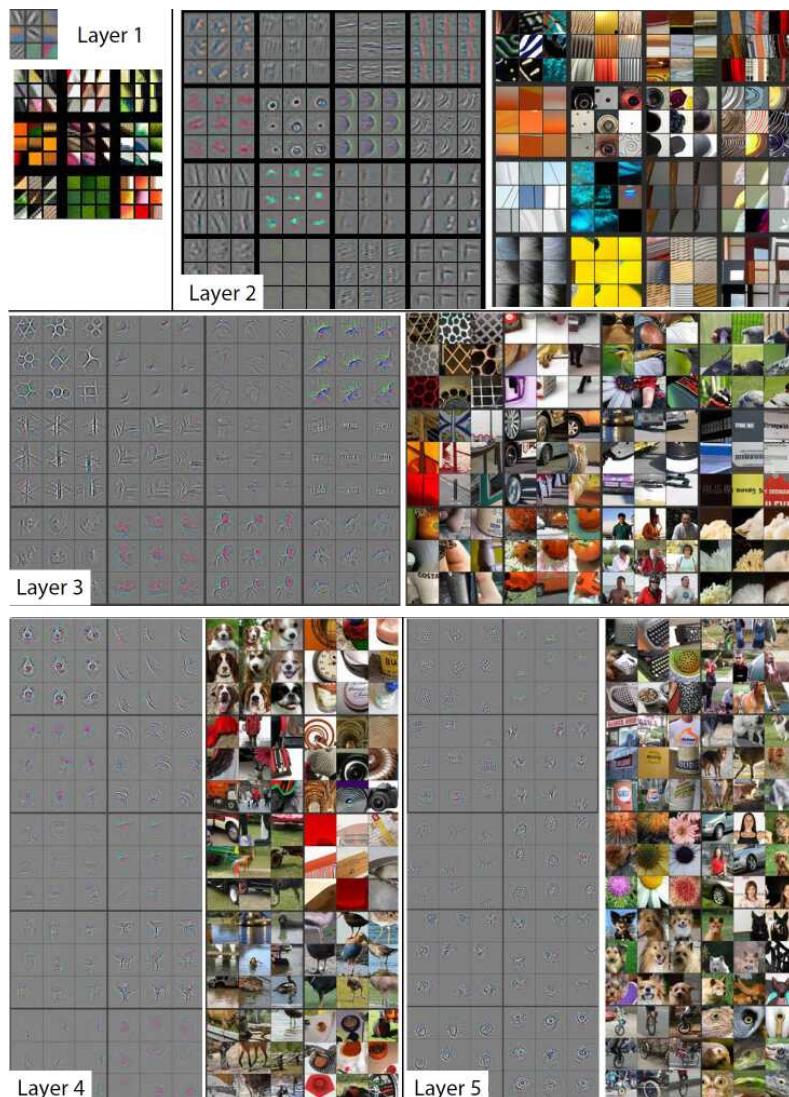


Figure 5.47 Visualizing network weights and features (Zeiler and Fergus 2014) © 2014 Springer. Each visualized convolutional layer is taken from a network adapted from the SuperVision net of Krizhevsky, Sutskever, and Hinton (2012). The 3×3 subimages denote the top nine responses in one feature map (channel in a given layer) projected back into pixel space (higher layers project to larger pixel patches), with the color images on the right showing the most responsive image patches from the validation set, and the grayish signed images on the left showing the corresponding maximum stimulus pre-images.

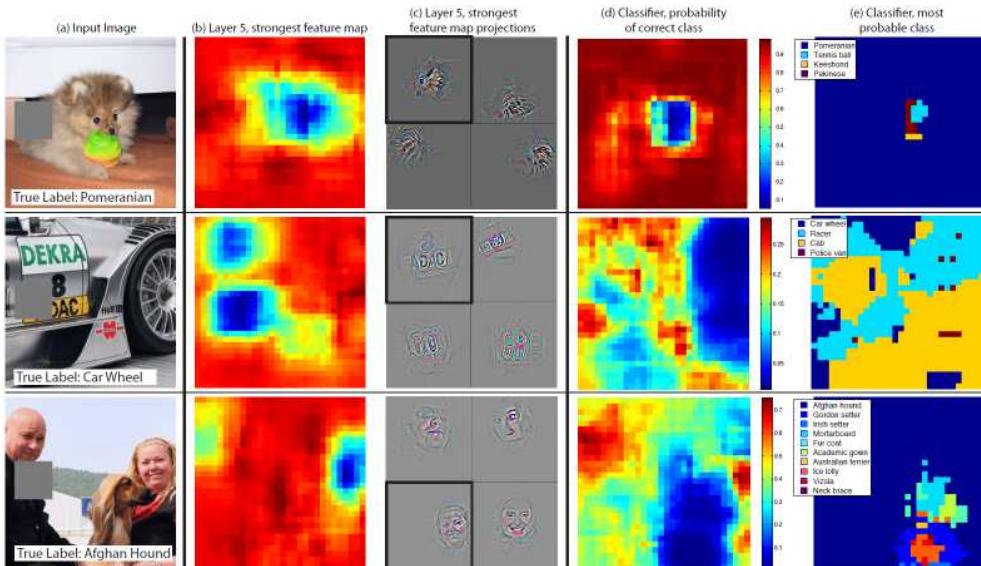


Figure 5.48 Heat map visualization from Zeiler and Fergus (2014) © 2014 Springer. By covering up portions of the input image with a small gray square, the response of a highly active channel in layer 5 can be visualized (second column), as can the feature map projections (third column), the likelihood of the correct class, and the most likely class per pixel.

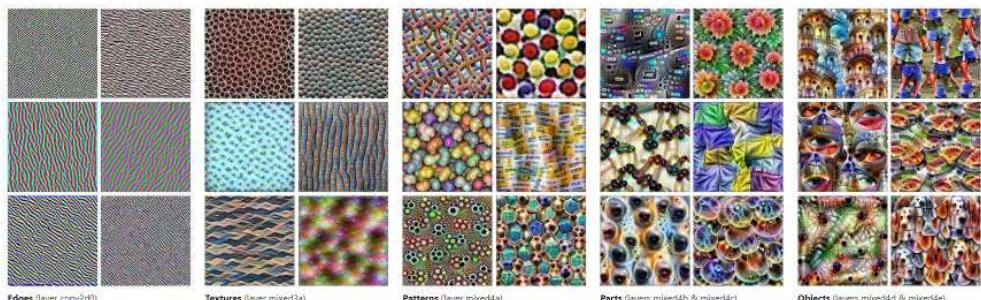


Figure 5.49 Feature visualization of how GoogLeNet (Szegedy, Liu et al. 2015) trained on ImageNet builds up its representations over different layers, from Olah, Mordvintsev, and Schubert (2017).

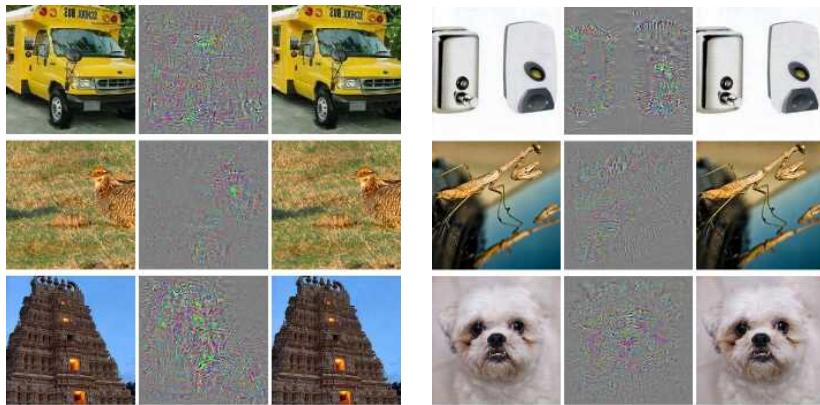


Figure 5.50 Examples of adversarial images © Szegedy, Zaremba et al. (2013). For each original image in the left column, a small random perturbation (shown magnified by 10× in the middle column) is added to obtain the image in the right column, which is always classified as an ostrich.

Many more techniques for visualizing neural network responses and behaviors have been described in various papers and blogs (Mordvintsev, Olah, and Tyka 2015; Zhou, Khosla *et al.* 2015; Nguyen, Yosinski, and Clune 2016; Bau, Zhou *et al.* 2017; Olah, Mordvintsev, and Schubert 2017; Olah, Satyanarayan *et al.* 2018; Cammarata, Carter *et al.* 2020), as well as the extensive lecture slides by Johnson (2020, Lecture 14). Figure 5.49 shows one example, visualizing different layers in a pre-trained GoogLeNet. OpenAI also recently released a great interactive tool called Microscope,⁴⁹ which allows people to visualize the significance of every neuron in many common neural networks.

5.4.6 Adversarial examples

While techniques such as guided backpropagation can help us better visualize neural network responses, they can also be used to “trick” deep networks into misclassifying inputs by subtly perturbing them, as shown in Figure 5.50. The key to creating such images is to take a set of final activations and to then backpropagate a gradient in the direction of the “fake” class, updating the input image until the fake class becomes the dominant activation. Szegedy, Zaremba *et al.* (2013) call such perturbed images *adversarial examples*.

Running this backpropagation requires access to the network and its weights, which means that this is a *white box attack*, as opposed to a *black box attack*, where nothing is

⁴⁹<https://microscope.openai.com/models>

known about the network. Surprisingly, however, the authors find “... that adversarial examples are relatively robust, and are shared by neural networks with varied number of layers, activations or trained on different subsets of the training data.”

The initial discovery of adversarial attacks spurred a flurry of additional investigations (Goodfellow, Shlens, and Szegedy 2015; Nguyen, Yosinski, and Clune 2015; Kurakin, Goodfellow, and Bengio 2016; Moosavi-Dezfooli, Fawzi, and Frossard 2016; Goodfellow, Papernot *et al.* 2017). Eykholt, Evtimov *et al.* (2018) show how adding simple stickers to real world objects (such as stop signs) can cause neural networks to misclassify photographs of such objects. Hendrycks, Zhao *et al.* (2021) have created a database of natural images that consistently fool popular deep classification networks trained on ImageNet. And while adversarial examples are mostly used to demonstrate the weaknesses of deep learning models, they can also be used to improve recognition (Xie, Tan *et al.* 2020).

Ilyas, Santurkar *et al.* (2019) try to demystify adversarial examples, finding that instead of making the anticipated large-scale perturbations that affect a human label, they are performing a type of shortcut learning (Lapuschkin, Wäldchen *et al.* 2019; Geirhos, Jacobsen *et al.* 2020). They find that optimizers are exploiting the *non-robust features* for an image label; that is, non-random correlations for an image class that exist in the dataset, but are not easily detected by humans. These non-robust features look merely like noise to a human observer, leaving images perturbed by them predominantly the same. Their claim is supported by training classifiers solely on non-robust features and finding that they correlate with image classification performance.

Are there ways to guard against adversarial attacks? The `cleverhans` software library (Papernot, Faghri *et al.* 2018) provides implementations of adversarial example construction techniques and adversarial training. There’s also an associated <http://www.cleverhans.io> blog on security and privacy in machine learning. Madry, Makelov *et al.* (2018) show how to train a network that is robust to bounded additive perturbations in known test images. There’s also recent work on detecting (Qin, Frosst *et al.* 2020b) and *deflecting* adversarial attacks (Qin, Frosst *et al.* 2020a) by forcing the perturbed images to visually resemble their (false) target class. This continues to be an evolving area, with profound implications for the robustness and safety of machine learning-based applications, as is the issue of *dataset bias* (Torralba and Efros 2011), which can be guarded against, to some extent, by testing *cross-dataset transfer* performance (Ranftl, Lasinger *et al.* 2020).

5.4.7 Self-supervised learning

As we mentioned previously, it is quite common to *pre-train* a *backbone* (or *trunk*) network for one task, e.g., whole image classification, and to then replace the final (few) layers with a

new *head* (or one or more *branches*), which are then trained for a different task, e.g., semantic image segmentation (He, Gkioxari *et al.* 2017). Optionally, the last few layers of the original backbone network can be *fine-tuned*.

The idea of training on one task and then using the learning on another is called *transfer learning*, while the process of modifying the final network to its intended application and statistics is called *domain adaptation*. While this idea was originally applied to backbones trained on labeled datasets such as ImageNet, i.e., in a *fully supervised* manner, the possibility of pre-training on the immensely larger set of unlabeled real-world images always remained a tantalizing possibility.

The central idea in *self-supervised* learning is to create a supervised *pretext task* where the labels can be automatically derived from unlabeled images, e.g., by asking the network to predict a subset of the information from the rest. Once pre-trained, the network can then be modified and fine-tuned on the final intended *downstream task*. Weng (2019) has a wonderful introductory blog post on this topic, and Zisserman (2018) has a great lecture, where the term *proxy task* is used. Additional good introductions can be found in the survey by Jing and Tian (2020) and the bibliography by Ren (2020).

Figure 5.51 shows some examples of pretext tasks that have been proposed for pre-training image classification networks. These include:

- Context prediction (Doersch, Gupta, and Efros 2015): take nearby image patches and predict their relative positions.
- Context encoders (Pathak, Krahenbuhl *et al.* 2016): inpaint one or more missing regions in an image.
- 9-tile jigsaw puzzle (Noroozi and Favaro 2016): rearrange the tiles into their correct positions.
- Colorizing black and white images (Zhang, Isola, and Efros 2016).
- Rotating images by multiples of 90° to make them upright (Gidaris, Singh, and Komodakis 2018). The paper compares itself against 11 previous self-supervised techniques.

In addition to using single-image pretext tasks, many researchers have used video clips, since successive frames contain semantically related content. One way to use this information is to order video frames correctly in time, i.e., to use a temporal version of context prediction and jigsaw puzzles (Misra, Zitnick, and Hebert 2016; Wei, Lim *et al.* 2018). Another is to extend colorization to video, with the colors in the first frame given (Vondrick,

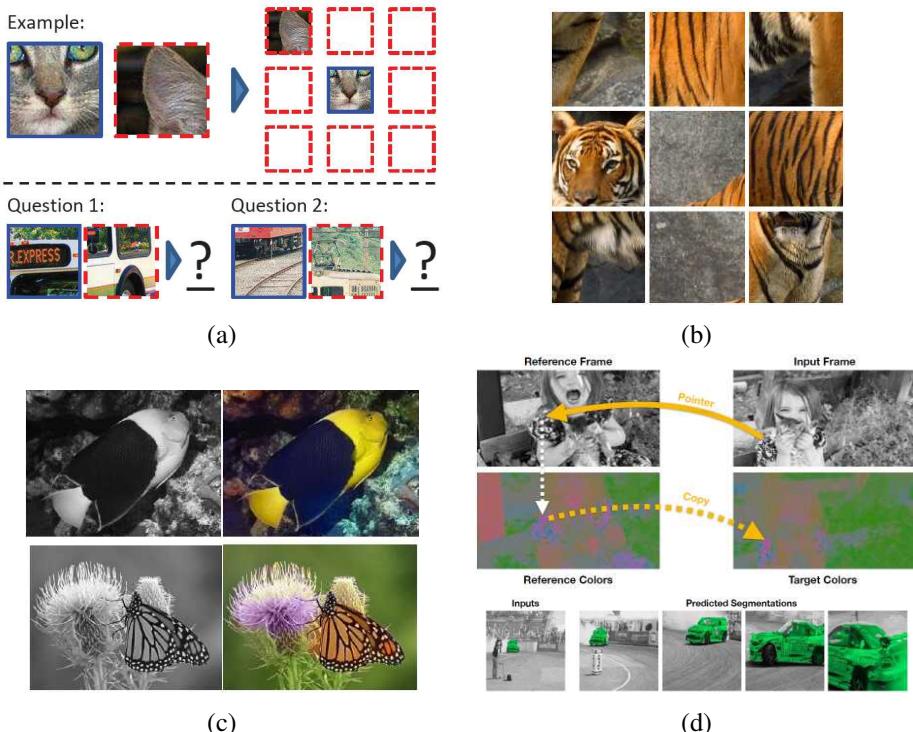


Figure 5.51 Examples of self-supervised learning tasks: (a) guessing the relative positions of image patches—can you guess the answers to Q1 and Q2? (Doersch, Gupta, and Efros 2015) © 2015 IEEE; (b) solving a nine-tile jigsaw puzzle (Noroozi and Favaro 2016) © 2016 Springer; (c) image colorization (Zhang, Isola, and Efros 2016) © 2016 Springer; (d) video color transfer for tracking (Vondrick, Shrivastava et al. 2018) © 2016 Springer.

Shrivastava *et al.* 2018), which encourages the network to learn semantic categories and correspondences. And since videos usually come with sounds, these can be used as additional cues in self-supervision, e.g., by asking a network to align visual and audio signals (Chung and Zisserman 2016; Arandjelovic and Zisserman 2018; Owens and Efros 2018), or in an unsupervised (contrastive) framework (Alwassel, Mahajan *et al.* 2020; Patrick, Asano *et al.* 2020).

Since self-supervised learning shows such great promise, an open question is whether such techniques could at some point surpass the performance of fully-supervised networks trained on smaller fully-labeled datasets.⁵⁰ Some impressive results have been shown using

⁵⁰https://people.eecs.berkeley.edu/~efros/gelato_beta.html

semi-supervised (weak) learning (Section 5.2.5) on very large (300M–3.5B) partially or noisily labeled datasets such as JFT-300M (Sun, Shrivastava *et al.* 2017) and Instagram hashtags (Mahajan, Girshick *et al.* 2018). Other researchers have tried simultaneously using supervised learning on labeled data and self-supervised pretext task learning on unlabeled data (Zhai, Oliver *et al.* 2019; Sun, Tzeng *et al.* 2019). It turns out that getting the most out of such approaches requires careful attention to dataset size, model architecture and capacity, and the extract details (and difficulty) of the pretext tasks (Kolesnikov, Zhai, and Beyer 2019; Goyal, Mahajan *et al.* 2019; Misra and Maaten 2020). At the same time, others are investigating how much real benefit pre-training actually gives in downstream tasks (He, Girshick, and Dollár 2019; Newell and Deng 2020; Feichtenhofer, Fan *et al.* 2021).

Semi-supervised training systems automatically generate ground truth labels for pretext tasks so that these can be used in a supervised manner (e.g, by minimizing classification errors). An alternative is to use unsupervised learning with a contrastive loss (Section 5.3.4) or other ranking loss (Gómez 2019) to encourage semantically similar inputs to produce similar encodings while spreading dissimilar inputs further apart. This is commonly now called *contrastive (metric) learning*.

Wu, Xiong *et al.* (2018) train a network to produce a separate embedding for each instance (training example), which they store in a moving average *memory bank* as new samples are fed through the neural network being trained. They then classify new images using nearest neighbors in the embedding space. Momentum Contrast (MoCo) replaces the memory bank with a fixed-length queue of encoded samples fed through a temporally adapted momentum encoder, which is separate from the actual network being trained (He, Fan *et al.* 2020). Pretext-invariant representation learning (PIRL) uses pretext tasks and “multi-crop” data augmentation, but then compares their outputs using a memory bank and contrastive loss (Misra and Maaten 2020). SimCLR (simple framework for contrastive learning) uses fixed mini-batches and applies a contrastive loss (normalized temperature cross-entropy, similar to (5.58)) between each sample in the batch and all the other samples, along with aggressive data augmentation (Chen, Kornblith *et al.* 2020). MoCo v2 combines ideas from MoCo and SimCLR to obtain even better results (Chen, Fan *et al.* 2020). Rather than directly comparing the generated embeddings, a fully connected network (MLP) is first applied.

Contrastive losses are a useful tool in metric learning, since they encourage distances in an embedding space to be small for semantically related inputs. An alternative is to use deep clustering to similarly encourage related inputs to produce similar outputs (Caron, Bojanowski *et al.* 2018; Ji, Henriques, and Vedaldi 2019; Asano, Rupprecht, and Vedaldi 2020; Gidaris, Bursuc *et al.* 2020; Yan, Misra *et al.* 2020). Some of the latest results using clustering for unsupervised learning now produce results competitive with contrastive metric learning

and also suggest that the kinds of data augmentation being used are even more important than the actual losses that are chosen (Caron, Misra *et al.* 2020; Tian, Chen, and Ganguli 2021). In the context of vision and language (Section 6.6), CLIP (Radford, Kim *et al.* 2021) has achieved remarkable generalization for image classification using contrastive learning and “natural-language supervision.” With a dataset of 400 million text and image pairs, their task is to take in a single image and a random sample of 32,768 text snippets and predict which text snippet is truly paired with the image.

Interestingly, it has recently been discovered that representation learning that *only* enforces similarity between semantically similar inputs also works well. This seems counter-intuitive, because without negative pairs as in contrastive learning, the representation can easily collapse to trivial solutions by predicting a constant for any input and maximizing similarity. To avoid this collapse, careful attention is often paid to the network design. Bootstrap Your Own Latent (BYOL) (Grill, Strub *et al.* 2020) shows that with a momentum encoder, an extra predictor MLP on the online network side, and a stop-gradient operation on the target network side, one can successfully remove the negatives from MoCo v2 training. SimSiam (Chen and He 2021) further show that even the momentum encoder is not required and only a stop-gradient operation is sufficient for the network to learn meaningful representations. While both systems jointly train the predictor MLP and the encoder with gradient updates, it has been even more recently shown that the predictor weights can be directly set using statistics of the input right before the predictor layer (Tian, Chen, and Ganguli 2021). Feichtenhofer, Fan *et al.* (2021) compare a number of these unsupervised representation learning techniques on a variety of video understanding tasks and find that the learned spatiotemporal representations generalize well to different tasks.

Contrastive learning and related work rely on compositions of data augmentations (e.g. color jitters, random crops, etc.) to learn representations that are *invariant* to such changes (Chen and He 2021). An alternative attempt is to use generative modeling (Chen, Radford *et al.* 2020), where the representations are pre-trained by predicting pixels either in an auto-regressive (GPT- or other language model) manner or a de-noising (BERT-, masked auto-encoder) manner. Generative modeling has the potential to bridge self-supervised learning across domains from vision to NLP, where scalable pre-trained models are now dominant.

One final variant on self-supervised learning is using a student-teacher model, where the teacher network is used to provide training examples to a student network. These kinds of architectures were originally called *model compression* (Bucilă, Caruana, and Niculescu-Mizil 2006) and *knowledge distillation* (Hinton, Vinyals, and Dean 2015) and were used to produce smaller models. However, when coupled with additional data and larger capacity networks, they can also be used to improve performance. Xie, Luong *et al.* (2020) train

an EfficientNet (Tan and Le 2019) on the labeled ImageNet training set, and then use this network to label an additional 300M unlabeled images. The true labels and pseudo-labeled images are then used to train a higher-capacity “student”, using regularization (e.g., dropout) and data augmentation to improve generalization. The process is then repeated to yield further improvements.

In all, self-supervised learning is currently one of the most exciting sub-areas in deep learning,⁵¹ and many leading researchers believe that it may hold the key to even better deep learning (LeCun and Bengio 2020). To explore implementations further, VISSL provides open-source PyTorch implementations of many state-of-the-art self-supervised learning models (with weights) that were described in this section.⁵²

5.5 More complex models

While deep neural networks started off being used in 2D image understanding and processing applications, they are now also widely used for 3D data such as medical images and video sequences. We can also chain a series of deep networks together in time by feeding the results from one time frame to the next (or even forward-backward). In this section, we look first at three-dimensional convolutions and then at recurrent models that propagate information forward or bi-directionally in time. We also look at *generative models* that can synthesize completely new images from semantic or related inputs.

5.5.1 Three-dimensional CNNs

As we just mentioned, deep neural networks in computer vision started off being used in the processing of regular two-dimensional images. However, as the amount of video being shared and analyzed increases, deep networks are also being applied to video understanding, which we discuss in more detail Section 6.5. We are also seeing applications in three-dimensional volumetric models such as occupancy maps created from range data (Section 13.5) and volumetric medical images (Section 6.4.1).

It may appear, at first glance, that the convolutional networks we have already studied, such as the ones illustrated in Figures 5.33, 5.34, and 5.39 already perform 3D convolutions, since their input receptive fields are 3D boxes in (x, y, c) , where c is the (feature) *channel* dimension. So, we could in principle fit a sliding window (say in time, or elevation) into a 2D network and be done. Or, we could use something like *grouped convolutions*. However,

⁵¹<https://sites.google.com/view/self-supervised-icml2019>

⁵²<https://github.com/facebookresearch/vissl>

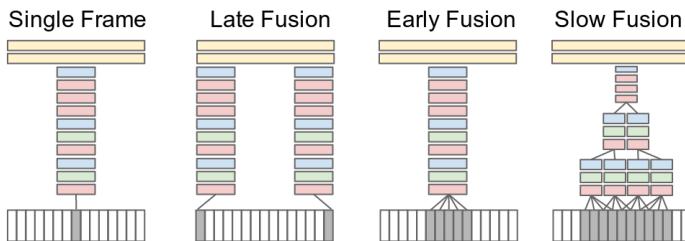


Figure 5.52 Alternative approaches to information fusion over the temporal dimensions (Karpathy, Toderici et al. 2014) © 2014 IEEE.

it's more convenient to operate on a complete 3D volume all at once, and to have *weight sharing* across the third dimension for all kernels, as well as multiple input and output feature channels at each layer.

One of the earliest applications of 3D convolutions was in the processing of video data to classify human actions (Kim, Lee, and Yang 2007; Ji, Xu et al. 2013; Baccouche, Mamalet et al. 2011). Karpathy, Toderici et al. (2014) describe a number of alternative architectures for fusing temporal information, as illustrated in Figure 5.52. The single frame approach classifies each frame independently, depending purely on that frame's content. Late fusion takes features generated from each frame and makes a per-clip classification. Early fusion groups small sets of adjacent frames into multiple channels in a 2D CNN. As mentioned before, the interactions across time do not have the convolutional aspects of weight sharing and temporal shift invariance. Finally, 3D CNNs (Ji, Xu et al. 2013) (not shown in this figure) learn 3D space and time-invariance kernels that are run over spatio-temporal windows and fused into a final score.

Tran, Bourdev et al. (2015) show how very simple $3 \times 3 \times 3$ convolutions combined with pooling in a deep network can be used to obtain even better performance. Their C3D network can be thought of as the “VGG of 3D CNNs” (Johnson 2020, Lecture 18). Carreira and Zisserman (2017) compare this architecture to alternatives that include *two-stream* models built by analyzing pixels and optical flows in parallel pathways (Figure 6.44b). Section 6.5 on video understanding discusses these and other architectures used for such problems, which have also been attacked using sequential models such as recurrent neural networks (RNNs) and LSTM, which we discuss in Section 5.5.2. Lecture 18 on video understanding by Johnson (2020) has a nice review of all these video understanding architectures.

In addition to video processing, 3D convolutional neural networks have been applied to volumetric image processing. Two examples of shape modeling and recognition from range data, i.e., 3D ShapeNets (Wu, Song et al. 2015) and VoxNet (Maturana and Scherer 2015)

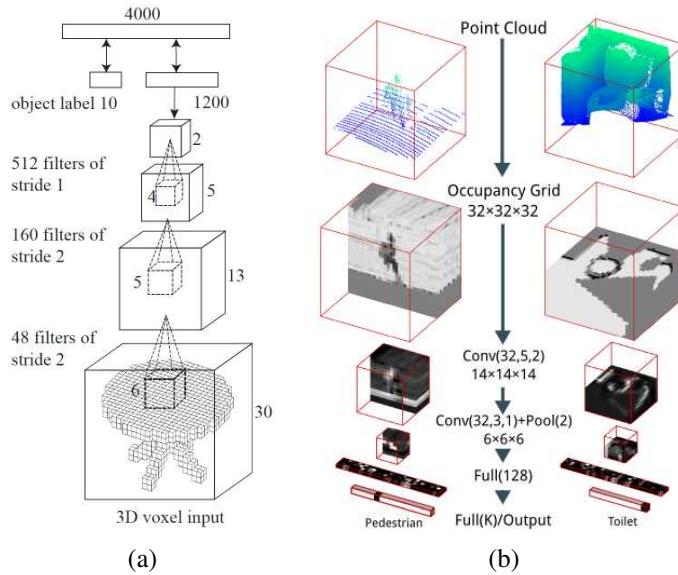


Figure 5.53 3D convolutional networks applied to volumetric data for object detection: (a) 3D ShapeNets (Wu, Song *et al.* 2015) © 2015 IEEE; (b) VoxNet (Maturana and Scherer 2015) © 2015 IEEE.

are shown in Figure 5.53. Examples of their application to medical image segmentation (Kamnitsas, Ferrante *et al.* 2016; Kamnitsas, Ledig *et al.* 2017) are discussed in Section 6.4.1. We discuss neural network approaches to 3D modeling in more detail in Sections 13.5.1 and 14.6.

Like regular 2D CNNs, 3D CNN architectures can exploit different spatial and temporal resolutions, striding, and channel depths, but they can be very computation and memory intensive. To counteract this, Feichtenhofer, Fan *et al.* (2019) develop a two-stream SlowFast architecture, where a slow pathway operates at a lower frame rate and is combined with features from a fast pathway with higher temporal sampling but fewer channels (Figure 6.44c). Video processing networks can also be made more efficient using channel-separated convolutions (Tran, Wang *et al.* 2019) and neural architecture search (Feichtenhofer 2020). Multigrid techniques (Appendix A.5.3) can also be used to accelerate the training of video recognition models (Wu, Girshick *et al.* 2020).

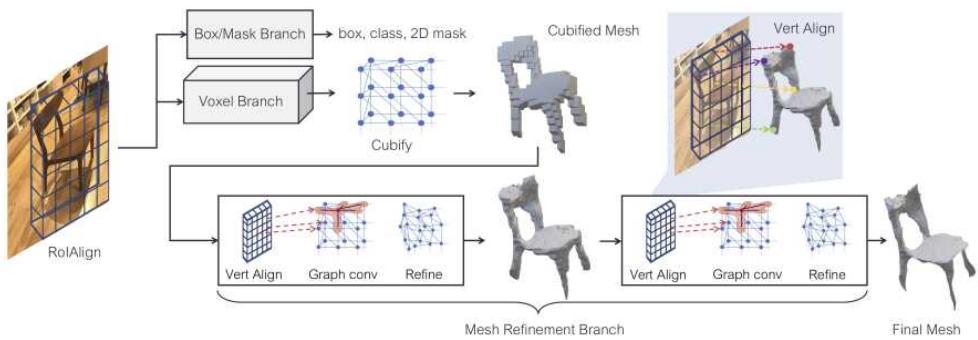


Figure 5.54 Overview of the Mesh R-CNN system (Gkioxari, Malik, and Johnson 2019) © 2019 IEEE. A Mask R-CNN backbone is augmented with two 3D shape inference branches. The voxel branch predicts a coarse shape for each detected object, which is further deformed with a sequence of refinement stages in the mesh refinement branch.

3D point clouds and meshes

In addition to processing 3D gridded data such as volumetric density, implicit distance functions, and video sequences, neural networks can be used to infer 3D models from single images. One approach is to predict per-pixel depth, which we study in Section 12.8. Another is to reconstruct full 3D models represented using volumetric density (Choy, Xu *et al.* 2016), which we study in Sections 13.5.1 and 14.6. Some more recent experiments, however, suggest that some of these 3D inference networks (Tatarchenko, Dosovitskiy, and Brox 2017; Groueix, Fisher *et al.* 2018; Richter and Roth 2018) may just be recognizing the general object category and doing a small amount of fitting (Tatarchenko, Richter *et al.* 2019).

Generating and processing 3D point clouds has also been extensively studied (Fan, Su, and Guibas 2017; Qi, Su *et al.* 2017; Wang, Sun *et al.* 2019). Guo, Wang *et al.* (2020) provide a comprehensive survey that reviews over 200 publications in this area.

A final alternative is to infer 3D triangulated meshes from either RGB-D (Wang, Zhang *et al.* 2018) or regular RGB (Gkioxari, Malik, and Johnson 2019; Wickramasinghe, Fua, and Knott 2021) images. Figure 5.54 illustrates the components of the Mesh R-CNN system, which detects images of 3D objects and turns each one into a triangulated mesh after first reconstructing a volumetric model. The primitive operations and representations needed to process such meshes using deep neural networks can be found in the PyTorch3D library.⁵³

⁵³<https://github.com/facebookresearch/pytorch3d>

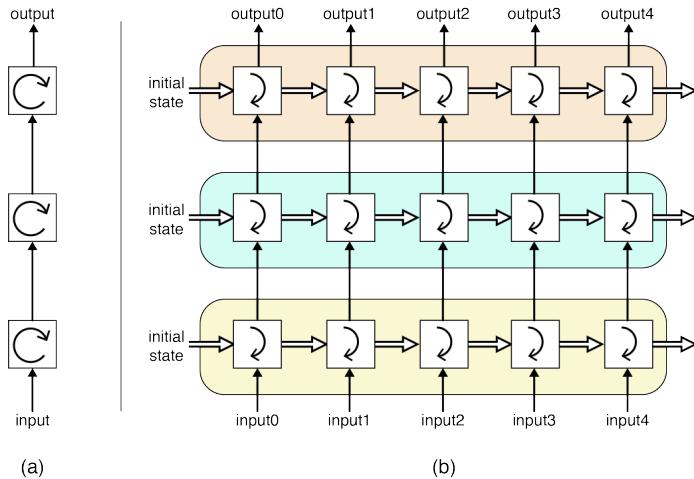


Figure 5.55 A deep recurrent neural network (RNN) uses multiple stages to process sequential data, with the output of one stage feeding the input of the next © Glassner (2018). Each stage maintains its own state and backpropagates its own gradients, although weights are shared between all stages. Column (a) shows a more compact rolled-up diagram, while column (b) shows the corresponding unrolled version.

5.5.2 Recurrent neural networks

While 2D and 3D convolutional networks are a good fit for images and volumes, sometimes we wish to process a *sequence* of images, audio signals, or text. A good way to exploit previously seen information is to pass features detected at one time instant (e.g., video frame) as input to the next frame’s processing. Such architectures are called Recurrent Neural Networks (RNNs) and are described in more detail in Goodfellow, Bengio, and Courville (2016, Chapter 10) and Zhang, Lipton *et al.* (2021, Chapter 8). Figure 5.55 shows a schematic sketch of such an architecture. Deep network layers not only pass information on to subsequent layers (and an eventual output), but also feed some of their information as input to the layer processing the next frame of data. Individual layers share weights across time (a bit like 3D convolution kernels), and backpropagation requires computing derivatives for all of the “unrolled” units (time instances) and summing these derivatives to obtain weight updates.

Because gradients can propagate for a long distance backward in time, and can therefore vanish or explode (just as in deep networks before the advent of residual networks), it is also possible to add extra *gating* units to modulate how information flows between frames. Such architectures are called *Gated Recurrent Units* (GRUs) and *Long short-term memory* (LSTM)

(Hochreiter and Schmidhuber 1997; Zhang, Lipton *et al.* 2021, Chapter 9).

RNNs and LSTMs are often used for video processing, since they can fuse information over time and model temporal dependencies (Baccouche, Mamalet *et al.* 2011; Donahue, Hendricks *et al.* 2015; Ng, Hausknecht *et al.* 2015; Srivastava, Mansimov, and Salakhudinov 2015; Ballas, Yao *et al.* 2016), as well as language modeling, image captioning, and visual question answering. We discuss these topics in more detail in Sections 6.5 and 6.6. They have also occasionally been used to merge multi-view information in stereo (Yao, Luo *et al.* 2019; Riegler and Koltun 2020a) and to simulate iterative flow algorithms in a fully differentiable (and hence trainable) manner (Hur and Roth 2019; Teed and Deng 2020b).

To propagate information forward in time, RNNs, GRUs, and LSTMs need to encode *all* of the potentially useful previous information in the hidden state being passed between time steps. In some situations, it is useful for a sequence modeling network to look further back (or even forward) in time. This kind of capability is often called *attention* and is described in more detail in Zhang, Lipton *et al.* (2021, Chapter 10), Johnson (2020, Lecture 12), and Section 5.5.3 on transformers. In brief, networks with attention store lists of *keys* and *values*, which can be probed with a *query* to return a weighted blend of values depending on the alignment between the query and each key. In this sense, they are similar to *kernel regression* (4.12–4.14), which we studied in Section 4.1.1, except that the query and the keys are multiplied (with appropriate weights) before being passed through a softmax to determine the blending weights.

Attention can either be used to look backward at the hidden states in previous time instances (which is called *self-attention*), or to look at different parts of the image (*visual attention*, as illustrated in Figure 6.46). We discuss these topics in more detail in Section 6.6 on vision and language. When recognizing or generating sequences, such as the words in a sentence, attention modules often used to work in tandem with sequential models such as RNNs or LSTMs. However, more recent works have made it possible to apply attention to the entire input sequence in one parallel step, as described in Section 5.5.3 on transformers.

The brief descriptions in this section just barely skim the broad topic of deep sequence modeling, which is usually covered in several lectures in courses on deep learning (e.g., Johnson 2020, Lectures 12–13) and several chapters in deep learning textbooks (Zhang, Lipton *et al.* 2021, Chapters 8–10). Interested readers should consult these sources for more detailed information.

5.5.3 Transformers

Transformers, which are a novel architecture that adds attention mechanisms (which we describe below) to deep neural networks, were first introduced by Vaswani, Shazeer *et al.* (2017)

in the context of neural machine translation, where the task consists of translating text from one language to another (Mikolov, Sutskever *et al.* 2013). In contrast to RNNs and their variants (Section 5.5.2), which process input tokens one at a time, transformers can to operate on the entire input sequence at once. In the years after first being introduced, transformers became the dominant paradigm for many tasks in natural language processing (NLP), enabling the impressive results produced by BERT (Devlin, Chang *et al.* 2018), RoBERTa (Liu, Ott *et al.* 2019), and GPT-3 (Brown, Mann *et al.* 2020), among many others. Transformers then began seeing success when processing the natural language component and later layers of many vision and language tasks (Section 6.6). More recently, they have gained traction in pure computer vision tasks, even outperforming CNNs on several popular benchmarks.

The motivation for applying transformers to computer vision is different than that of applying it to NLP. Whereas RNNs suffer from sequentially processing the input, convolutions do not have this problem, as their operations are already inherently parallel. Instead, the problem with convolutions has to do with their *inductive biases*, i.e., the default assumptions encoded into convolutional models.

A convolution operation assumes that nearby pixels are more important than far away pixels. Only after several convolutional layers are stacked together does the receptive field grow large enough to attend to the entire image (Araujo, Norris, and Sim 2019), unless the network is endowed with non-local operations (Wang, Girshick *et al.* 2018) similar to those used in some image denoising algorithms (Buades, Coll, and Morel 2005a). As we have seen in this chapter, convolution’s spatial locality bias has led to remarkable success across many aspects of computer vision. But as datasets, models, and computational power grow by orders of magnitude, these inductive biases may become a factor inhibiting further progress.⁵⁴

The fundamental component of a transformer is *self-attention*, which is itself built out of applying *attention* to each of N unit activations in a given layer in the network.⁵⁵ Attention is often described using an analogy to the concept of *associative maps* or *dictionaries* found as data structures in programming languages and databases. Given a set of *key-value pairs*, $\{(\mathbf{k}_i, \mathbf{v}_i)\}$ and a query \mathbf{q} , a dictionary returns the value \mathbf{v}_i corresponding to the key \mathbf{k}_i that exactly matches the query. In neural networks, the key and query values are real-valued vectors (e.g., linear projections of activations), so the corresponding operation returns a weighted sum of values where the weights depend on the pairwise distances between a query and the set of keys. This is basically the same as *scattered data interpolation*, which we studied in

⁵⁴Rich Sutton, a pioneer in reinforcement learning, believes that learning to leverage computation, instead of encoding human knowledge, is *the bitter lesson* to learn from the history of AI research (Sutton 2019). Others disagree with this view, believing that it is essential to be able to learn from small amounts of data (Lake, Salakhutdinov, and Tenenbaum 2015; Marcus 2020).

⁵⁵ N may indicate the number of words in a sentence or patches in an image

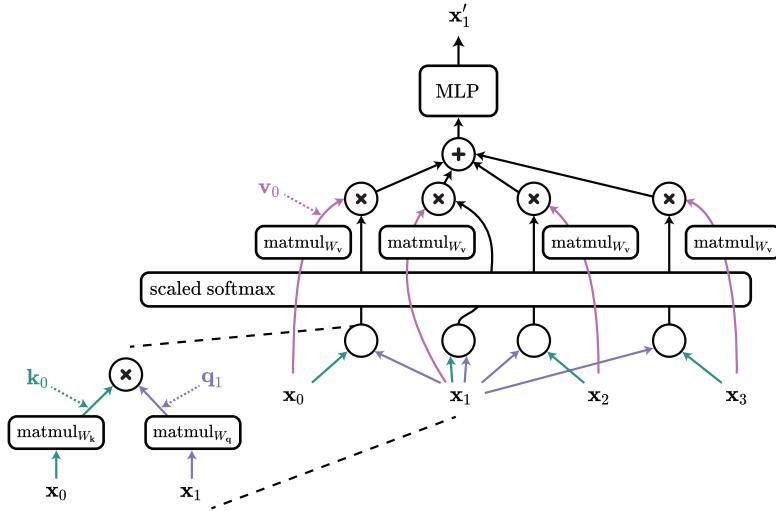


Figure 5.56 The self-attention computation graph to compute a single output vector \mathbf{x}'_1 , courtesy of Matt Deitke, adapted from Vaswani, Huang, and Manning (2019). Note that the full self-attention operation also computes outputs for \mathbf{x}'_1 , \mathbf{x}'_3 , and \mathbf{x}'_4 by shifting the input to the query (\mathbf{x}_2 in this case) between \mathbf{x}_1 , \mathbf{x}_3 , and \mathbf{x}_4 , respectively. For each of $\text{matmul}_{\mathbf{V}}$, $\text{matmul}_{\mathbf{K}}$, and $\text{matmul}_{\mathbf{Q}}$, there is a single matrix of weights that gets reused with each call.

Section 4.1.1, as pointed out in Zhang, Lipton *et al.* (2021, Section 10.2). However, instead of using radial distances as in (4.14), attention mechanisms in neural networks more commonly use *scaled dot-product attention* (Vaswani, Shazeer *et al.* 2017; Zhang, Lipton *et al.* 2021, Section 10.3.3), which involves taking the dot product between the query and key vectors, scaling down by the square root of the dimension of these embeddings D ,⁵⁶ and then applying the softmax function of (5.5), i.e.,

$$\mathbf{y} = \sum_i \alpha(\mathbf{q} \cdot \mathbf{k}_i / D) \mathbf{v}_i = \text{softmax}(\mathbf{q}^T \mathbf{K} / D)^T \mathbf{V}, \quad (5.77)$$

where \mathbf{K} and \mathbf{V} are the row-stacked matrices composed of the key and value vectors, respectively, and \mathbf{y} is the output of the attention operator.⁵⁷

Given a set of input vectors $\{\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_{N-1}\}$, the self-attention operation produces a set of output vectors $\{\mathbf{x}'_0, \mathbf{x}'_1, \dots, \mathbf{x}'_{N-1}\}$. Figure 5.56 shows the case for $N = 4$, where

⁵⁶We divide the dot product by \sqrt{D} so that the variance of the scaled dot product does not increase for larger embedding dimensions, which could result in vanishing gradients.

⁵⁷The partition of unity function α notation is borrowed from Zhang, Lipton *et al.* (2021, Section 10.3).

the self-attention computation graph is used to obtain a single output vector \mathbf{x}'_2 . As pictured, self-attention uses three learned weight matrices, \mathbf{W}_q , \mathbf{W}_k , and \mathbf{W}_v , which determine the

$$\mathbf{q}_i = \mathbf{W}_q \mathbf{x}_i, \mathbf{k}_i = \mathbf{W}_k \mathbf{x}_i, \text{ and } \mathbf{v}_i = \mathbf{W}_v \mathbf{x}_i \quad (5.78)$$

per-unit query, key, and value vectors going into each attention block. The weighted sum of values is then optionally passed through a multi-layer perceptron (MLP) to produce \mathbf{x}'_2 .

In comparison to a fully connected or convolutional layer, self-attention computes each output (e.g., \mathbf{x}'_i) based on all of the input vectors $\{\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_{N-1}\}$. In that sense, it is often compared to a fully connected layer, but instead of the weights being fixed for each input, the weights are adapted on the spot, based on the input (Khan, Naseer *et al.* 2021). Compared to convolutions, self-attention is able to attend to every part of the input from the start, instead of constraining itself to local regions of the input, which may help it introduce the kind of context information needed to disambiguate the objects shown in Figure 6.8.

There are several components that are combined with self-attention to produce a transformer block, as described in Vaswani, Shazeer *et al.* (2017). The full transformer consists of both an encoder and a decoder block, although both share many of the same components. In many applications, an encoder can be used without a decoder (Devlin, Chang *et al.* 2018; Dosovitskiy, Beyer *et al.* 2021) and vice versa (Razavi, van den Oord, and Vinyals 2019).

The right side of Figure 5.57 shows an example of a transformer encoder block. For both the encoder and decoder:

- Instead of modeling set-to-set operations, we can model sequence-to-sequence operations by adding a **positional encoding** to each input vector (Gehring, Auli *et al.* 2017). The positional encoding typically consists of a set of temporally shifted sine waves from which position information can be decoded. (Such position encodings have also recently been added to implicit neural shape representations, which we study in Sections 13.5.1 and 14.6.)
- In lieu of applying a single self-attention operation to the input, multiple self-attention operations, with different learned weight matrices to build different keys, values, and queries, are often joined together to form **multi-headed self-attention** (Vaswani, Shazeer *et al.* 2017). The result of each head is then concatenated together before everything is passed through an MLP.
- **Layer normalization** (Ba, Kiros, and Hinton 2016) is then applied to the output of the MLP. Each vector may then independently be passed through another MLP with shared weights before layer normalization is applied again.

- **Residual connections** (He, Zhang *et al.* 2016a) are employed after multi-headed attention and after the final MLP.

During training, the biggest difference in the decoder is that some of the input vectors to self-attention may be masked out, which helps support parallel training in autoregressive prediction tasks. Further exposition of the details and implementation of the transformer architecture is provided in Vaswani, Shazeer *et al.* (2017) and in the additional reading (Section 5.6).

A key challenge of applying transformers to the image domain has to do with the size of image input (Vaswani, Shazeer *et al.* 2017). Let N denote the length of the input, D denote the number of dimensions for each input entry, and K denote a convolution’s (on side) kernel size.⁵⁸ The number of floating point operations (FLOPs) required for self-attention is on the order of $O(N^2D)$, whereas the FLOPs for a convolution operation is on the order of $O(ND^2K^2)$. For instance, with an ImageNet image scaled to size $224 \times 224 \times 3$, if each pixel is treated independently, $N = 224 \times 224 = 50176$ and $D = 3$. Here, a convolution is *significantly* more efficient than self-attention. In contrast, applications like neural machine translation may only have N as the number of words in a sentence and D as the dimension for each word embedding (Mikolov, Sutskever *et al.* 2013), which makes self-attention much more efficient.

The Image Transformer (Parmar, Vaswani *et al.* 2018) was the first attempt at applying the full transformer model to the image domain, with many of the same authors that introduced the transformer. It used both an encoder and decoder to try and build an autoregressive generative model that predicts the next pixel, given a sequence of input pixels and all the previously predicted pixels. (The earlier work on non-local networks by Wang, Girshick *et al.* (2018) also used ideas inspired by transformers, but with a simpler attention block and a fully two-dimensional setup.) Each vector input to the transformer corresponded to a single pixel, which ultimately constrained them to generate small images (i.e., 32×32), since the quadratic cost of self-attention was too expensive otherwise.

Dosovitskiy, Beyer *et al.* (2021) had a breakthrough that allowed transformers to process much larger images. Figure 5.57 shows the diagram of the model, named the Vision Transformer (ViT). For the task of image recognition, instead of treating each pixel as a separate input vector to the transformer, they divide an image (of size 224×224) into 196 distinct 16×16 gridded image patches. Each patch is then flattened, and passed through a shared embedding matrix, which is equivalent to a strided 16×16 convolution, and the results are combined with a positional encoding vector and then passed to the transformer. Earlier work

⁵⁸In Section 5.4 on convolutional architectures, we use C to denote the number of channels instead of D to denote the embedding dimensions.

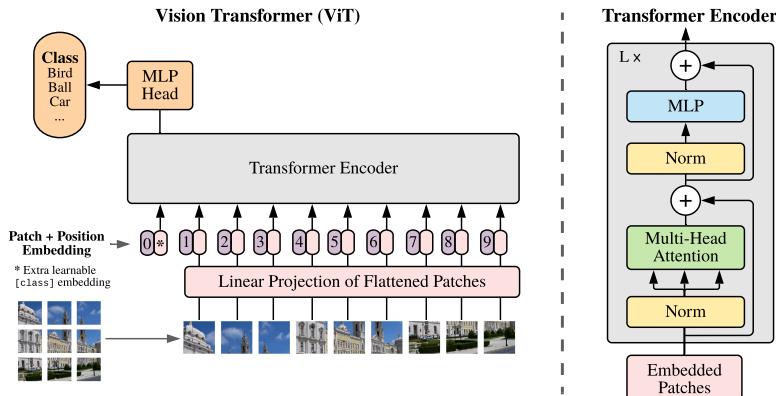


Figure 5.57 The Vision Transformer (ViT) model from (Dosovitskiy, Beyer et al. 2021) breaks an image into a 16×16 grid of patches. Each patch is then flattened, passed through a shared embedding matrix, and combined with a positional encoding vector. These inputs are then passed through a transformer encoder (right) several times before predicting an image’s class.

from Cordonnier, Loukas, and Jaggi (2019) introduced a similar patching approach, but on a smaller scale with 2×2 patches.

ViT was only able to outperform their convolutional baseline BiT (Kolesnikov, Beyer et al. 2020) when using over 100 million training images from JFT-300M (Sun, Shrivastava et al. 2017). When using ImageNet alone, or a random subset of 10 or 30 million training samples from JPT-300, the ViT model typically performed much worse than the BiT baseline. Their results suggest that in low-data domains, the inductive biases present in convolutions are typically quite useful. But, with orders of magnitude of more data, a transformer model might discover even better representations that are not representable with a CNN.

Some works have also gone into combining the inductive biases of convolutions with transformers (Srinivas, Lin et al. 2021; Wu, Xiao et al. 2021; Lu, Batra et al. 2019; Yuan, Guo et al. 2021). An influential example of such a network is DETR (Carion, Massa et al. 2020), which is applied to the task of object detection. It first processes the image with a ResNet backbone, with the output getting passed to a transformer encoder-decoder architecture. They find that the addition of a transformer improves the ability to detect large objects, which is believed to be because of its ability to reason globally about correspondences between inputted encoding vectors.

The application and usefulness of transformers in the realm of computer vision is still being widely researched. Already, however, they have achieved impressive performance on

a wide range of tasks, with new papers being published rapidly.⁵⁹ Some more notable applications include image classification (Liu, Lin *et al.* 2021; Touvron, Cord *et al.* 2020), object detection (Dai, Cai *et al.* 2020; Liu, Lin *et al.* 2021), image pre-training (Chen, Radford *et al.* 2020), semantic segmentation (Zheng, Lu *et al.* 2020), pose recognition (Li, Wang *et al.* 2021), super-resolution (Zeng, Fu, and Chao 2020), colorization (Kumar, Weissenborn, and Kalchbrenner 2021), generative modeling (Jiang, Chang, and Wang 2021; Hudson and Zitnick 2021), and video classification (Arnab, Dehghani *et al.* 2021; Fan, Xiong *et al.* 2021; Li, Zhang *et al.* 2021). Recent works have also found success extending ViT’s patch embedding to pure MLP vision architectures (Tolstikhin, Houlsby *et al.* 2021; Liu, Dai *et al.* 2021; Touvron, Bojanowski *et al.* 2021). Applications to vision and language are discussed in Section 6.6.

5.5.4 Generative models

Throughout this chapter, I have mentioned that machine learning algorithms such as logistic regression, support vector machines, random trees, and feedforward deep neural networks are all examples of *discriminative* systems that never form an explicit *generative* model of the quantities they are trying to estimate (Bishop 2006, Section 1.5; Murphy 2012, Section 8.6). In addition to the potential benefits of generative models discussed in these two textbooks, Goodfellow (2016) and Kingma and Welling (2019) list some additional ones, such as the ability to visualize our assumptions about our unknowns, training with missing or incompletely labeled data, and the ability to generate multiple, alternative, results.

In computer graphics, which is sometimes called *image synthesis* (as opposed to the *image understanding* or *image analysis* we do in computer vision), the ability to easily generate realistic random images and models has long been an essential tool. Examples of such algorithms include texture synthesis and style transfer, which we study in more detail in Section 10.5, as well as fractal terrain (Fournier, Fussel, and Carpenter 1982) and tree generation (Prusinkiewicz and Lindenmayer 1996). Examples of deep neural networks being used to generate such novel images, often under user control, are shown in Figures 5.60 and 10.58. Related techniques are also used in the nascent field of *neural rendering*, which we discuss in Section 14.6.

How can we unlock the demonstrated power of deep neural networks to capture semantics in order to visualize sample images and generate new ones? One approach could be to use the visualization techniques introduced in Section 5.4.5. But as you can see from Figure 5.49, while such techniques can give us insights into individual units, they fail to create

⁵⁹<https://github.com/dk-liang/Awesome-Visual-Transformer>

fully realistic images.

Another approach might be to construct a *decoder* network to undo the classification performed by the original (*encoder*) network. This kind of “bottleneck” architecture is widely used, as shown in Figure 5.37a, to derive semantic per-pixel labels from images. Can we use a similar idea to generate realistic looking images?

Variational autoencoders

A network that encodes an image into small compact codes and then attempts to decode it back into the same image is called an *autoencoder*. The compact codes are typically represented as a vector, which is often called the *latent* vector to emphasize that it is hidden and unknown. Autoencoders have a long history of use in neural networks, even predating today’s feedforward networks (Kingma and Welling 2019). It was once believed that this might be a good way to pre-train networks, but the more challenging proxy tasks we studied in Section 5.4.7 have proven to be more effective.

At a high level, to train an autoencoder on a dataset of images, we can use an unsupervised objective that tries to have the output image of the decoder match the training image input to the encoder. To generate a new image, we can then randomly sample a latent vector and hope that from that vector, the decoder can generate a new image that looks like it came from the distribution of training images in our dataset.

With an autoencoder, there is a deterministic, one-to-one mapping from each input to its latent vector. Hence, the number of latent vectors that are generated exactly matches the number of input data points. If the encoder’s objective is to produce a latent vector that makes it easy to decode, one possible solution would be for every latent vector to be extremely far away from every other latent vector. Here, the decoder can overfit all the latent vectors it has seen since they would all be unique with little overlap. However, as our goal is to randomly generate latent vectors that can be passed to the decoder to generate realistic images, we want the latent space to both be well explored and to encode some meaning, such as nearby vectors being semantically similar. Ghosh, Sajjadi *et al.* (2019) propose one potential solution, where they inject noise into the latent vector and empirically find that it works quite well.

Another extension of the autoencoder is the *variational autoencoder* (VAE) (Kingma and Welling 2013; Rezende, Mohamed, and Wierstra 2014; Kingma and Welling 2019). Instead of generating a single latent vector for each input, it generates the mean and covariance that define a chosen distribution of latent vectors. The distribution can then be sampled from to produce a single latent vector, which gets passed into the decoder. To avoid having the covariance matrix become the zero matrix, making the sampling process deterministic, the objective function often includes a regularization term to penalize the distribution if it is far

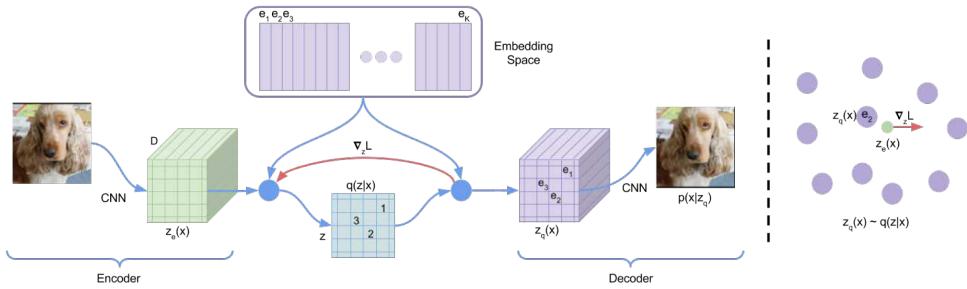


Figure 5.58 The VQ-VAE model. On the left, $z_e(x)$ represents the output of the encoder, the embedding space on top represents the codebook of K embedding vectors, and $q(z | x)$ represents the process of replacing each spatial (i.e., channel-wise) vector in the output of the encoder with its nearest vector in the codebook. On the right, we see how a $z_e(x)$ vector (green) may be rounded to e_2 , and that the gradient in the encoder network (red) may push the vector away from e_2 during backpropagation. © van den Oord, Vinyals, and Kavukcuoglu (2017)

from some chosen (e.g., Gaussian) distribution. Due to their probabilistic nature, VAEs can explore the space of possible latent vectors significantly better than autoencoders, making it harder for the decoder to overfit the training data.

Motivated by how natural language is discrete and by how images can typically be described in language (Section 6.6), the vector quantized VAE (VQ-VAE) of van den Oord, Vinyals, and Kavukcuoglu (2017) takes the approach of modeling the latent space with categorical variables. Figure 5.58 shows an outline of the VQ-VAE architecture. The encoder and decoder operate like a normal VAE, where the encoder predicts some latent representation from the input, and the decoder generates an image from the latent representation. However, in contrast to the normal VAE, the VQ-VAE replaces each spatial dimension of the predicted latent representation with its nearest vector from a discrete set of vectors (named the *codebook*). The discretized latent representation is then passed to the decoder. The vectors in the codebook are trained simultaneously with the VAE’s encoder and decoder. Here, the codebook vectors are optimized to move closer to the spatial vectors outputted by the encoder.

Although a VQ-VAE uses a discrete codebook of vectors, the number of possible images it can represent is still monstrously large. In some of their image experiments, they set the size of the codebook to $K = 512$ vectors and set the size of the latent variable to be $z = 32 \times 32 \times 1$. Here, they can represent $512^{32 \times 32 \times 1}$ possible images.

Compared to a VAE, which typically assumes a Gaussian latent distribution, the latent

distribution of a VQ-VAE is not as clearly defined, so a separate generative model is trained to sample latent variables z . The model is trained on the final latent variables outputted from the trained VQ-VAE encoder across the training data. For images, entries in z are often spatially dependent, e.g., an object may be encoded over many neighboring entries. With entries being chosen from a discrete codebook of vectors, we can use a PixelCNN (van den Oord, Kalchbrenner *et al.* 2016) to autoregressively sample new entries in the latent variable based on previously sampled neighboring entries. The PixelCNN can also be conditionally trained, which enables the ability to sample latent variables corresponding to a particular image class or feature.

A follow-up to the VQ-VAE model, named VQ-VAE-2 (Razavi, van den Oord, and Vinyals 2019), uses a two-level approach to decoding images, where with both a small and large latent vector, they can get much higher fidelity reconstructed and generated images. Section 6.6 discusses Dall·E (Ramesh, Pavlov *et al.* 2021), a model that applies VQ-VAE-2 to text-to-image generation and achieves remarkable results.

Generative adversarial networks

Another possibility for image synthesis is to use the multi-resolution features computed by pre-trained networks to match the statistics of a given texture or style image, as described in Figure 10.57. While such networks are useful for matching the *style* of a given artist and the high-level *content* (layout) of a photograph, they are not sufficient to generate completely photorealistic images.

In order to create truly photorealistic synthetic images, we want to determine if an image is real(istic) or fake. If such a loss function existed, we could use it to train networks to generate synthetic images. But, since such a loss function is incredibly difficult to write by hand, why not train a separate neural network to play the critic role? This is the main insight behind the *generative adversarial networks* introduced by Goodfellow, Pouget-Abadie *et al.* (2014). In their system, the output of the *generator* network G is fed into a separate *discriminator* network D , whose task is to tell “fake” synthetically generated images apart from real ones, as shown in Figure 5.59a. The goal of the generator is to create images that “fool” the discriminator into accepting them as real, while the goal of the discriminator is to catch the “forger” in their act. Both networks are co-trained simultaneously, using a blend of loss functions that encourage each network to do its job. The joint loss function can be written as

$$E_{\text{GAN}}(\mathbf{w}_G, \mathbf{w}_D) = \sum_n \log D(\mathbf{x}_n) + \log (1 - D(G(\mathbf{z}_n))), \quad (5.79)$$

where the $\{\mathbf{x}_n\}$ are the real-world training images, $\{\mathbf{z}_n\}$ are random vectors, which are

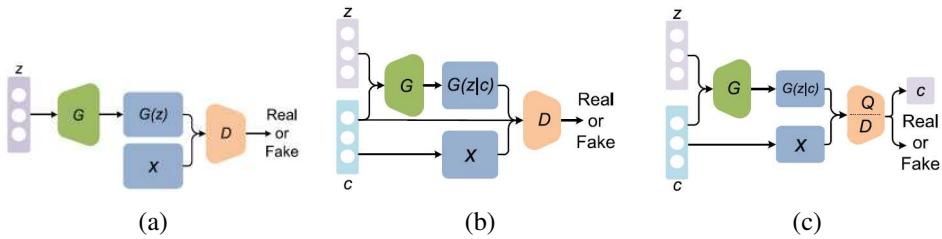


Figure 5.59 Generative adversarial network (GAN) architectures from Pan, Yu et al. (2019) © 2019 IEEE. (a) In a regular GAN, random “latent” noise vectors z are fed into a generator network G , which produces synthetic “fake” images $x' = G(z)$. The job of the discriminator D is to tell the fake images apart from real samples x . (b) In a conditional GAN (cGAN), the network iterates (during training) over all the classes that we wish to synthesize. The generator G gets both a class id c and a random noise vector z as input, and the discriminator D gets the class id as well and needs to determine if its input is a real member of the given class. (c) The discriminator in an InfoGAN does not have access to the class id, but must instead infer it from the samples it is given.

passed through the generator G to produce synthetic images \mathbf{x}'_n , and the $\{\mathbf{w}_G, \mathbf{w}_D\}$ are the weights (parameters) in the generator and discriminator.

Instead of minimizing this loss, we adjust the weights of the generator to minimize the second term (they do not affect the first), and adjust the weights of the discriminator to *maximize* both terms, i.e., minimize the discriminator’s error. This process is often called a *minimax game*.⁶⁰ More details about the formulation and how to optimize it can be found in the original paper by Goodfellow, Pouget-Abadie et al. (2014), as well as deep learning textbooks (Zhang, Lipton et al. 2021, Chapter 17), lectures (Johnson 2020, Lecture 20), tutorials (Goodfellow, Isola et al. 2018), and review articles (Creswell, White et al. 2018; Pan, Yu et al. 2019).

The original paper by Goodfellow, Pouget-Abadie et al. (2014) used a small, fully connected network to demonstrate the basic idea, so it could only generate 32×32 images such as MNIST digits and low-resolution faces. The *Deep Convolutional GAN* (DCGAN) introduced by Radford, Metz, and Chintala (2015) uses the second half of the deconvolution network shown in Figure 5.37a to map from the random latent vectors \mathbf{z} to arbitrary size images and can therefore generate a much wider variety of outputs, while LAPGAN uses

⁶⁰Note that the term *adversarial* in GANs refers to this adversarial game between the generator and the discriminator, which helps the generator create better pictures. This is distinct from the *adversarial examples* we discussed in Section 5.4.6, which are images designed to fool recognition systems.

a Laplacian pyramid of adversarial networks (Denton, Chintala *et al.* 2015). Blending between different latent vectors (or perturbing them in certain directions) generates in-between synthetic images.

GANs and DCGANs can be trained to generate new samples from a given class, but it is even more useful to generate samples from different classes using the same trained network. The *conditional GAN* (cGAN) proposed by Mirza and Osindero (2014) achieves this by feeding a class vector into both the generator, which *conditions* its output on this second input, as well as the discriminator, as shown in Figure 5.59b. It is also possible to make the discriminator predict classes that correlate with the class vector using an extra mutual information term, as shown in Figure 5.59c (Chen, Duan *et al.* 2016). This allows the resulting InfoGAN network to learn disentangled representations, such as the digit shapes and writing styles in MNIST, or pose and lighting.

While generating random images can have many useful graphics applications, such as generating textures, filling holes, and stylizing photographs, as discussed in Section 10.5, it becomes even more useful when it can be done under a person’s artistic control (Lee, Zitnick, and Cohen 2011). The iGAN interactive image editing system developed by Zhu, Krähenbühl *et al.* (2016) does this by learning a manifold of photorealistic images using a generative adversarial network and then constraining user edits (or even sketches) to produce images that lie on this manifold.

This approach was generalized by Isola, Zhu *et al.* (2017) to all kinds of other image-to-image translation tasks, as shown in Figure 5.60a. In their pix2pix system, images, which can just be sketches or semantic labels, are fed into a modified U-Net, which converts them to images with different semantic meanings or styles (e.g., photographs or road maps). When the input is a semantic label map and the output is a photorealistic image, this process is often called *semantic image synthesis*. The translation network is trained with a conditional GAN, which takes paired images from the two domains at training time and has the discriminator decide if the synthesized (translated) image together with the input image are a real or fake pair. Referring back to Figure 5.59b, the class c is now a complete image, which is fed into both G and the discriminator D , along with its paired or synthesized output. Instead of making a decision for the whole image, the discriminator looks at overlapping patches and makes decisions on a patch-by-patch basis, which requires fewer parameters and provides more training data and more discriminative feedback. In their implementation, there is no random vector z ; instead, dropout is used during both training and “test” (translation) time, which is equivalent to injecting noise at different levels in the network.

In many situations, paired images are not available, e.g., when you have collections of paintings and photographs from different locations, or pictures of animals in two different

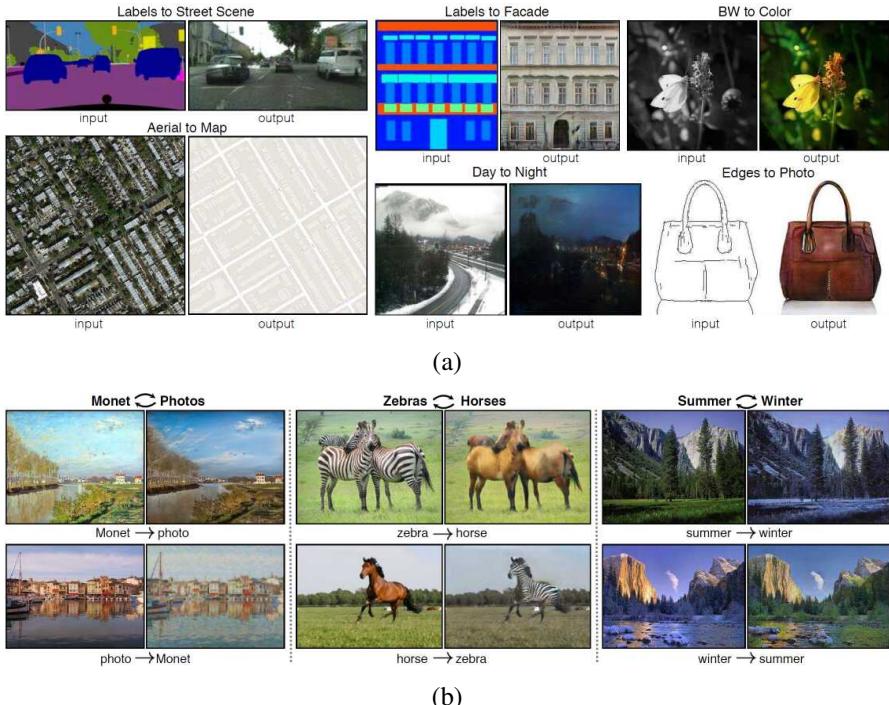


Figure 5.60 *Image-to-image translation. (a) Given paired training images, the original pix2pix system learns how to turn sketches into photos, semantic maps to images, and other pixel remapping tasks (Isola, Zhu et al. 2017) © 2017 IEEE. (b) CycleGAN does not require paired training images, just collections coming from different sources, such as painting and photographs or horses and zebras (Zhu, Park et al. 2017) © 2017 IEEE.*

classes, as shown in Figure 5.60b. In this case, a cycle-consistent adversarial network (CycleGAN) can be used to require the mappings between the two domains to encourage identity, while also ensuring that generated images are perceptually similar to the training images (Zhu, Park et al. 2017). DualGAN (Yi, Zhang et al. 2017) and DiscoGAN (Kim, Cha et al. 2017) use related ideas. The BicycleGAN system of Zhu, Zhang et al. (2017) uses a similar idea of transformation cycles to encourage encoded latent vectors to correspond to different modes in the outputs for better interpretability and control.

Since the publication of the original GAN paper, the number of extensions, applications, and follow-on papers has exploded. The GAN Zoo website⁶¹ lists over 500 GAN papers published between 2014 and mid-2018, at which point it stopped being updated. Large number

⁶¹<https://github.com/hindupuravinash/the-gan-zoo>

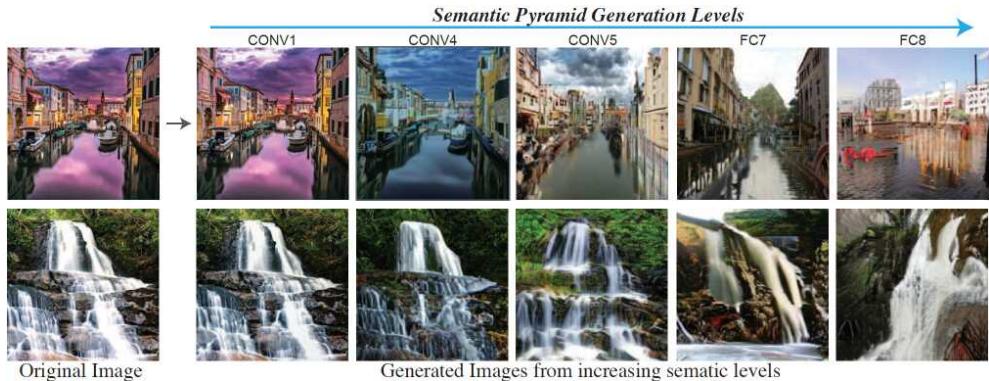


Figure 5.61 The Semantic Image Pyramid can be used to choose which semantic level in a deep network to modify when editing an image (Shocher, Gandsman et al. 2020) © 2020 IEEE.

of papers continue to appear each year in vision, machine learning, and graphics conferences.

Some of the more important papers since 2017 include Wasserstein GANs (Arjovsky, Chintala, and Bottou 2017), Progressive GANs (Karras, Aila et al. 2018), UNIT (Liu, Breuel, and Kautz 2017) and MUNIT (Huang, Liu et al. 2018), spectral normalization (Miyato, Kataoka et al. 2018), SAGAN (Zhang, Goodfellow et al. 2019), BigGAN (Brock, Donahue, and Simonyan 2019), StarGAN (Choi, Choi et al. 2018) and StyleGAN (Karras, Laine, and Aila 2019) and follow-on papers (Choi, Uh et al. 2020; Karras, Laine et al. 2020; Viazovetskyi, Ivashkin, and Kashin 2020), SPADE (Park, Liu et al. 2019), GANSpace (Härkönen, Hertzmann et al. 2020), and VQGAN (Esser, Rombach, and Ommer 2020). You can find more detailed explanations and references to many more papers in the lectures by Johnson (2020, Lecture 20), tutorials by Goodfellow, Isola et al. (2018), and review articles by Creswell, White et al. (2018), Pan, Yu et al. (2019), and Tewari, Fried et al. (2020).

In summary, generative adversarial networks and their myriad extensions continue to be an extremely vibrant and useful research area, with applications such as image super-resolution (Section 10.3), photorealistic image synthesis (Section 10.5.3), image-to-image translation, and interactive image editing. Two very recent examples of this last application are the Semantic Pyramid for Image Generation by Shocher, Gandsman et al. (2020), in which the semantic manipulation level can be controlled (from small texture changes to higher-level layout changes), as shown in Figure 5.61, and the Swapping Autoencoder by Park, Zhu et al. (2020), where structure and texture can be independently edited.

5.6 Additional reading

Machine learning and deep learning are rich, broad subjects which properly deserve their own course of study to master. Fortunately, there are a large number of good textbooks and online courses available to learn this material.

My own favorite for machine learning is the book by Bishop (2006), since it provides a broad treatment with a Bayesian flavor and excellent figures, which I have re-used in this book. The books by Glassner (2018, 2021) provide an even gentler introduction to both classic machine learning and deep learning, as well as additional figures I reference in this book. Two additional widely used textbooks for machine learning are Hastie, Tibshirani, and Friedman (2009) and Murphy (2012). Deisenroth, Faisal, and Ong (2020) provide a nice compact treatment of mathematics for machine learning, including linear and matrix algebra, probability theory, model fitting, regression, PCA, and SVMs, with a more in-depth exposition than the terse summaries I provide in this book. The book on Automated Machine Learning edited by Hutter, Kotthoff, and Vanschoren (2019) surveys automated techniques for designing and optimizing machine learning algorithms.

For deep learning, Goodfellow, Bengio, and Courville (2016) were the first to provide a comprehensive treatment, but it has not recently been revised. Glassner (2018, 2021) provides a wonderful introduction to deep learning, with lots of figures and no equations. I recommend it even to experienced practitioners since it helps develop and solidify intuitions about how learning works. An up-to-date reference on deep learning is the *Dive into Deep Learning* online textbook by Zhang, Lipton *et al.* (2021), which comes with interactive Python notebooks sprinkled throughout the text, as well as an associated course (Smola and Li 2019). Some introductory courses to deep learning use Charniak (2019).

Rawat and Wang (2017) provide a nice review article on deep learning, including a history of early and later neural networks, as well in-depth discussion of many deep learning components, such as pooling, activation functions, losses, regularization, and optimization. Additional surveys related to advances in deep learning include Sze, Chen *et al.* (2017), Elsken, Metzen, and Hutter (2019), Gu, Wang *et al.* (2018), and Choudhary, Mishra *et al.* (2020). Sejnowski (2018) provides an in-depth history of the early days of neural networks.

The Deep Learning for Computer Vision course slides by Johnson (2020) are an outstanding reference and a great way to learn the material, both for the depth of their information and how up-to-date the presentations are kept. They are based on Stanford's CS231n course (Li, Johnson, and Yeung 2019), which is also a great up-to-date source. Additional classes on deep learning with slides and/or video lectures include Grosse and Ba (2019), McAllester (2020), Leal-Taixé and Nießner (2020), Leal-Taixé and Nießner (2021), and Geiger (2021).

For transformers, Bloem (2019) provides a nice starting tutorial on implementing the standard transformer encoder and decoder block in PyTorch, from scratch. More comprehensive surveys of transformers applied to computer vision include Khan, Naseer *et al.* (2021) and Han, Wang *et al.* (2020). Tay, Dehghani *et al.* (2020) provides an overview of many attempts to reduce the quadratic cost of self-attention. Wightman (2021) makes available a fantastic collection of computer vision transformer implementations in PyTorch, with pre-trained weights and great documentation. Additional course lectures introducing transformers with videos and slides include Johnson (2020, Lecture 13), Vaswani, Huang, and Manning (2019, Lecture 14) and LeCun and Canziani (2020, Week 12).

For GANs, the new deep learning textbook by Zhang, Lipton *et al.* (2021, Chapter 17), lectures by Johnson (2020, Lecture 20), tutorials by Goodfellow, Isola *et al.* (2018), and review articles by Creswell, White *et al.* (2018), Pan, Yu *et al.* (2019), and Tewari, Fried *et al.* (2020) are all good sources. For a survey of the latest visual recognition techniques, the tutorials presented at ICCV (Xie, Girshick *et al.* 2019), CVPR (Girshick, Kirillov *et al.* 2020), and ECCV (Xie, Girshick *et al.* 2020) are excellent up-to-date sources.

5.7 Exercises

Ex 5.1: Backpropagation and weight updates. Implement the forward activation, backward gradient and error propagation, and weight update steps in a simple neural network. You can find examples of such code in HW3 of the 2020 UW CSE 576 class⁶² or the Educational Framework (EDF) developed by McAllester (2020) and used in Geiger (2021).

Ex 5.2: LeNet. Download, train, and test a simple “LeNet” (LeCun, Bottou *et al.* 1998) convolutional neural network on the CIFAR-10 (Krizhevsky 2009) or Fashion MNIST (Xiao, Rasul, and Vollgraf 2017) datasets. You can find such code in numerous places on the web, including HW4 of the 2020 UW CSE 576 class or the PyTorch beginner tutorial on Neural Networks.⁶³

Modify the network to remove the non-linearities. How does the performance change? Can you improve the performance of the original network by increasing the number of channels, layers, or convolution sizes? Do the training and testing accuracies move in the same or different directions as you modify your network?

Ex 5.3: Deep learning textbooks. Both the *Deep Learning: From Basics to Practice* book by Glassner (2018, Chapters 15, 23, and 24) and the *Dive into Deep Learning* book by Zhang,

⁶²<https://courses.cs.washington.edu/courses/cse576/20sp/calendar/>

⁶³https://pytorch.org/tutorials/beginner/blitz/neural_networks_tutorial.html

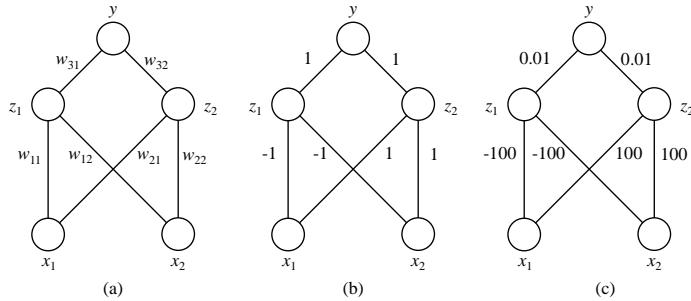


Figure 5.62 Simple two hidden unit network with a ReLU activation function and no bias parameters for regressing the function $y = |x_1 + 1.1x_2|$: (a) can you guess a set of weights that would fit this function?; (b) a reasonable set of starting weights; (c) a poorly scaled set of weights.

Lipton *et al.* (2021) contain myriad graded exercises with code samples to develop your understanding of deep neural networks. If you have the time, try to work through most of these.

Ex 5.4: Activation and weight scaling. Consider the two hidden unit network shown in Figure 5.62, which uses ReLU activation functions and has no additive bias parameters. Your task is to find a set of weights that will fit the function

$$y = |x_1 + 1.1x_2|. \quad (5.80)$$

1. Can you guess a set of weights that will fit this function?
2. Starting with the weights shown in column b, compute the activations for the hidden and final units as well as the regression loss for the nine input values $(x_1, x_2) \in \{-1, 0, 1\} \times \{-1, 0, 1\}$.
3. Now compute the gradients of the squared loss with respect to all six weights using the backpropagation chain rule equations (5.65–5.68) and sum them up across the training samples to get a final gradient.
4. What step size should you take in the gradient direction, and what would your update squared loss become?
5. Repeat this exercise for the initial weights in column (c) of Figure 5.62.
6. Given this new set of weights, how much worse is your error decrease, and how many iterations would you expect it to take to achieve a reasonable solution?

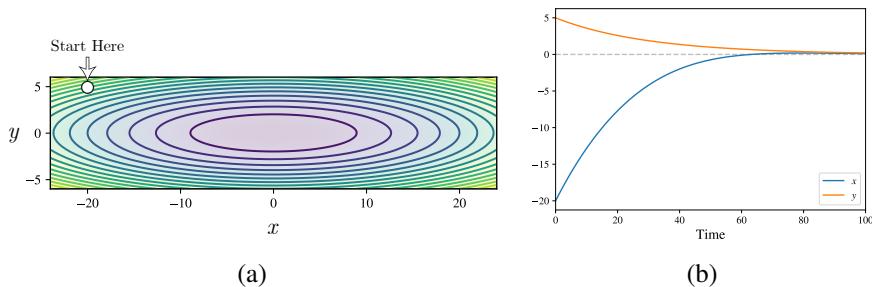


Figure 5.63 Function optimization: (a) the contour plot of $f(x, y) = x^2 + 20y^2$ with the function being minimized at $(0, 0)$; (b) ideal gradient descent optimization that quickly converges towards the minimum at $x = 0, y = 0$.

7. Would batch normalization help in this case?

Note: the following exercises were suggested by Matt Deitke.

Ex 5.5: Function optimization. Consider the function $f(x, y) = x^2 + 20y^2$ shown in Figure 5.63a. Begin by solving for the following:

1. Calculate ∇f , i.e., the gradient of f .
2. Evaluate the gradient at $x = -20, y = 5$.

Implement some of the common gradient descent optimizers, which should take you from the starting point $x = -20, y = 5$ to near the minimum at $x = 0, y = 0$. Try each of the following optimizers:

1. Standard gradient descent.
2. Gradient descent with momentum, starting with the momentum term as $\rho = 0.99$.
3. Adam, starting with decay rates of $\beta_1 = 0.9$ and $b_2 = 0.999$.

Play around with the learning rate α . For each experiment, plot how x and y change over time, as shown in Figure 5.63b.

How do the optimizers behave differently? Is there a single learning rate that makes all the optimizers converge towards $x = 0, y = 0$ in under 200 steps? Does each optimizer monotonically trend towards $x = 0, y = 0$?

Ex 5.6: Weight initialization. For an arbitrary neural network, is it possible to initialize the weights of a neural network such that it will never train on any non-trivial task, such as image classification or object detection? Explain why or why not.

Ex 5.7: Convolutions. Consider convolving a $256 \times 256 \times 3$ image with 64 separate convolution kernels. For kernels with heights and widths of $\{(3 \times 3), (5 \times 5), (7 \times 7), \text{ and } (9 \times 9)\}$, answer each of the following:

1. How many parameters (i.e., weights) make up the convolution operation?
2. What is the output size after convolving the image with the kernels?

Ex 5.8: Data augmentation. The figure below shows image augmentations that translate and scale an image.



Let CONV denote a convolution operation, f denote an arbitrary function (such as scaling or translating an image), and IMAGE denote the input image. A function f has invariance, with respect to a convolution, when $\text{CONV}(\text{IMAGE}) = \text{CONV}(f(\text{IMAGE}))$, and equivariance when $\text{CONV}(f(\text{IMAGE})) = f(\text{CONV}(\text{IMAGE}))$. Answer and explain each of the following:

1. Are convolutions translation invariant?
2. Are convolutions translation equivariant?
3. Are convolutions scale invariant?
4. Are convolutions scale equivariant?

Ex 5.9: Training vs. validation. Suppose your model is performing significantly better on the training data than it is on the validation data. What changes might be made to the loss function, training data, and network architecture to prevent such overfitting?

Ex 5.10: Cascaded convolutions. With only a single matrix multiplication, how can multiple convolutional kernel's convolve over an entire input image? Here, let the input image be of size $256 \times 256 \times 3$ and each of the 64 kernels be of size $3 \times 3 \times 3$.⁶⁴

Ex 5.11: Pooling vs. 1×1 convolutions. Pooling layers and 1×1 convolutions are both commonly used to shrink the size of the proceeding layer. When would you use one over the other?

⁶⁴*Hint:* You will need to reshape the input and each convolution's kernel size.

Ex 5.12: Inception. Why is an inception module more efficient than a residual block? What are the comparative disadvantages of using an inception module?

Ex 5.13: ResNets. Why is it easier to train a ResNet with 100 layers than a VGG network with 100 layers?

Ex 5.14: U-Nets. An alternative to the U-Net architecture is to not change the size of the height and width intermediate activations throughout the network. The final layer would then be able to output the same transformed pixel-wise representation of the input image. What is the disadvantage of this approach?

Ex 5.15: Early vs. late fusion in video processing. What are two advantages of early fusion compared to late fusion?

Ex 5.16: Video-to-video translation. Independently pass each frame in a video through a `pix2pix` model. For instance, if the video is of the day, then the output might be each frame at night. Stitch the output frames together to form a video. What do you notice? Does the video look plausible?

Ex 5.17: Vision Transformer. Using a Vision Transformer (ViT) model, pass several images through it and create a histogram of the activations after each layer normalization operation. Do the histograms tend to form of a normal distribution?

Ex 5.18: GAN training. In the GAN loss formulation, suppose the discriminator D is near-perfect, such that it correctly outputs near 1 for real images \mathbf{x}_n and near 0 for synthetically generated images $G(\mathbf{z}_n)$.

1. For both the discriminator and the generator, compute its approximate loss with

$$\mathcal{L}_{\text{GAN}}(\mathbf{x}_n, \mathbf{z}_n) = \log D(\mathbf{x}_n) + \log(1 - D(G(\mathbf{z}_n))), \quad (5.81)$$

where the discriminator tries to minimize \mathcal{L}_{GAN} and the generator tries to maximize \mathcal{L}_{GAN} .

2. How well can this discriminator be used to train the generator?
3. Can you modify the generator's loss function, $\min \log(1 - D(G(\mathbf{z}_n)))$, such that it is easier to train with both a great discriminator and a discriminator that is no better than random?⁶⁵

⁶⁵*Hint:* The loss function should suggest a relatively large change to fool a great discriminator and a relatively small change with a discriminator that is no better than random.

Ex 5.19: Colorization. Even though large amounts of unsupervised data can be collected for image colorization, it often does not train well using a pixel-wise regression loss between an image's predicted colors and its true colors. Why is that? Is there another loss function that may be better suited for the problem?

Chapter 6

Recognition

| | | |
|-------|--|-----|
| 6.1 | Instance recognition | 346 |
| 6.2 | Image classification | 349 |
| 6.2.1 | Feature-based methods | 350 |
| 6.2.2 | Deep networks | 358 |
| 6.2.3 | <i>Application:</i> Visual similarity search | 360 |
| 6.2.4 | Face recognition | 363 |
| 6.3 | Object detection | 370 |
| 6.3.1 | Face detection | 371 |
| 6.3.2 | Pedestrian detection | 376 |
| 6.3.3 | General object detection | 379 |
| 6.4 | Semantic segmentation | 387 |
| 6.4.1 | <i>Application:</i> Medical image segmentation | 390 |
| 6.4.2 | Instance segmentation | 391 |
| 6.4.3 | Panoptic segmentation | 392 |
| 6.4.4 | <i>Application:</i> Intelligent photo editing | 394 |
| 6.4.5 | Pose estimation | 395 |
| 6.5 | Video understanding | 396 |
| 6.6 | Vision and language | 400 |
| 6.7 | Additional reading | 409 |
| 6.8 | Exercises | 413 |

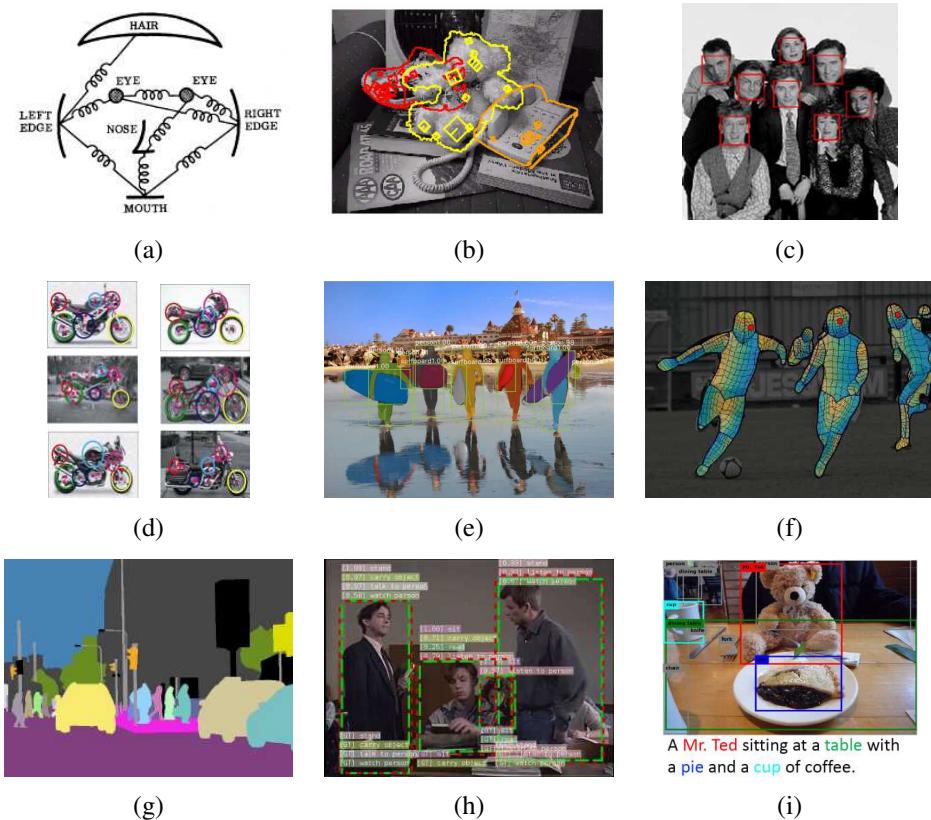


Figure 6.1 Various kinds of recognition: (a) face recognition with pictorial structures (Fischler and Elschlager 1973) © 1973 IEEE; (b) instance (known object) recognition (Lowe 1999) © 1999 IEEE; (c) real-time face detection (Viola and Jones 2004) © 2004 Springer; (d) feature-based recognition (Fergus, Perona, and Zisserman 2007) © 2007 Springer; (e) instance segmentation using Mask R-CNN (He, Gkioxari et al. 2017) © 2017 IEEE; (f) pose estimation (Güler, Neverova, and Kokkinos 2018) © 2018 IEEE; (g) panoptic segmentation (Kirillov, He et al. 2019) © 2019 IEEE; (h) video action recognition (Feichtenhofer, Fan et al. 2019); (i) image captioning (Lu, Yang et al. 2018) © 2018 IEEE.

Of all the computer vision topics covered in this book, visual recognition has undergone the largest changes and fastest development in the last decade, due in part to the availability of much larger labeled datasets as well as breakthroughs in deep learning (Figure 5.40). In the first edition of this book (Szeliski 2010), recognition was the last chapter, since it was considered a “high-level task” to be layered on top of lower-level components such as feature detection and matching. In fact, many introductory vision courses still teach recognition at the end, often covering “classic” (non-learning) vision algorithms and applications first, and then shifting to deep learning and recognition.

As I mentioned in the preface and introduction, I have now moved machine and deep learning to early in the book, since it is foundational technology widely used in other parts of computer vision. I also decided to move the recognition chapter right after deep learning, since most of the modern techniques for recognition are natural applications of deep neural networks. The majority of the old recognition chapter has been replaced with newer deep learning techniques, so you will sometimes find terse descriptions of classical recognition techniques along with pointers to the first edition and relevant surveys or seminal papers.

A good example of the classic approach is *instance recognition*, where we are trying to find exemplars of a particular manufactured object such as a stop sign or sneaker (Figure 6.1b). (An even earlier example is face recognition using relative feature locations, as shown in Figure 6.1a.) The general approach of finding distinctive features while dealing with local appearance variation (Section 7.1.2), and then checking for their co-occurrence and relative positions in an image, is still widely used for manufactured 3D object detection (Figure 6.3), 3D structure and pose recovery (Chapter 11), and location recognition (Section 11.2.3). Highly accurate and widely used feature-based approaches to instance recognition were developed in the 2000s (Figure 7.27) and, despite more recent deep learning-based alternatives, are often still the preferred method (Sattler, Zhou *et al.* 2019). We review instance recognition in Section 6.1, although some of the needed components, such as feature detection, description, and matching (Chapter 7), as well as 3D pose estimation and verification (Chapter 11), will not be introduced until later.

The more difficult problem of *category* or *class recognition* (e.g., recognizing members of highly variable categories such as cats, dogs, or motorcycles) was also initially attacked using feature-based approaches and relative locations (*part-based models*), such as the one depicted in Figure 6.1d. We begin our discussion of *image classification* (another name for whole-image category recognition) in Section 6.2 with a review of such “classic” (though now rarely used) techniques. We then show how the deep neural networks described in the previous chapter are ideally suited to these kinds of classification problems. Next, we cover visual similarity search, where instead of categorizing an image into a predefined number of

categories, we retrieve other images that are semantically similar. Finally, we focus on face recognition, which is one of the longest studied topics in computer vision.

In Section 6.3, we turn to the topic of *object detection*, where we categorize not just whole images but delineate (with bounding boxes) where various objects are located. This topic includes more specialized variants such as *face detection* and *pedestrian detection*, as well as the detection of objects in generic categories. In Section 6.4, we study *semantic segmentation*, where the task is now to delineate various objects and materials in a pixel-accurate manner, i.e., to label each pixel with an object identity and class. Variants on this include *instance segmentation*, where each separate object gets a unique label, *panoptic segmentation*, where both objects and stuff (e.g., grass, sky) get labeled, and *pose estimation*, where pixels get labeled with people’s body parts and orientations. The last two sections of this chapter briefly touch on *video understanding* (Section 6.5) and *vision and language* (Section 6.6).

Before starting to describe individual recognition algorithms and variants, I should briefly mention the critical role that large-scale datasets and benchmarks have played in the rapid advancement of recognition systems. While small datasets such as Xerox 10 (Csurka, Dance *et al.* 2006) and Caltech-101 (Fei-Fei, Fergus, and Perona 2006) played an early role in evaluating object recognition systems, the PASCAL Visual Object Class (VOC) challenge (Everingham, Van Gool *et al.* 2010; Everingham, Eslami *et al.* 2015) was the first dataset large and challenging enough to significantly propel the field forward. However, PASCAL VOC only contained 20 classes. The introduction of the ImageNet dataset (Deng, Dong *et al.* 2009; Russakovsky, Deng *et al.* 2015), which had 1,000 classes and over one million labeled images, finally provided enough data to enable end-to-end learning systems to break through. The Microsoft COCO (Common Objects in Context) dataset spurred further development (Lin, Maire *et al.* 2014), especially in accurate per-object segmentation, which we study in Section 6.4. A nice review of crowdsourcing methods to construct such datasets is presented in (Kovashka, Russakovsky *et al.* 2016). We will mention additional, sometimes more specialized, datasets throughout this chapter. A listing of the most popular and active datasets and benchmarks is provided in Tables 6.1–6.4.

6.1 Instance recognition

General object recognition falls into two broad categories, namely *instance recognition* and *class recognition*. The former involves re-recognizing a known 2D or 3D rigid object, potentially being viewed from a novel viewpoint, against a cluttered background, and with partial



Figure 6.2 Recognizing objects in a cluttered scene (Lowe 2004) © 2004 Springer. Two of the training images in the database are shown on the left. They are matched to the cluttered scene in the middle using SIFT features, shown as small squares in the right image. The affine warp of each recognized database image onto the scene is shown as a larger parallelogram in the right image.

occlusions.¹ The latter, which is also known as *category-level* or *generic* object recognition (Ponce, Hebert *et al.* 2006), is the much more challenging problem of recognizing any instance of a particular general class, such as “cat”, “car”, or “bicycle”.

Over the years, many different algorithms have been developed for instance recognition. Mundy (2006) surveys earlier approaches, which focused on extracting lines, contours, or 3D surfaces from images and matching them to known 3D object models. Another popular approach was to acquire images from a large set of viewpoints and illuminations and to represent them using an eigenspace decomposition (Murase and Nayar 1995). More recent approaches (Lowe 2004; Lepetit and Fua 2005; Rothganger, Lazebnik *et al.* 2006; Ferrari, Tuytelaars, and Van Gool 2006b; Gordon and Lowe 2006; Obdržálek and Matas 2006; Sivic and Zisserman 2009; Zheng, Yang, and Tian 2018) tend to use viewpoint-invariant 2D features, such as those we will discuss in Section 7.1.2. After extracting informative sparse 2D features from both the new image and the images in the database, image features are matched against the object database, using one of the sparse feature matching strategies described in Section 7.1.3. Whenever a sufficient number of matches have been found, they are verified by finding a geometric transformation that aligns the two sets of features (Figure 6.2).

¹The Microsoft COCO dataset paper (Lin, Maire *et al.* 2014) introduced the newer concept of *instance segmentation*, which is the pixel-accurate delineation of different objects drawn from a set of generic classes (Section 6.4.2). This now sometimes leads to confusion, unless you look at these two terms (instance recognition vs. segmentation) carefully.

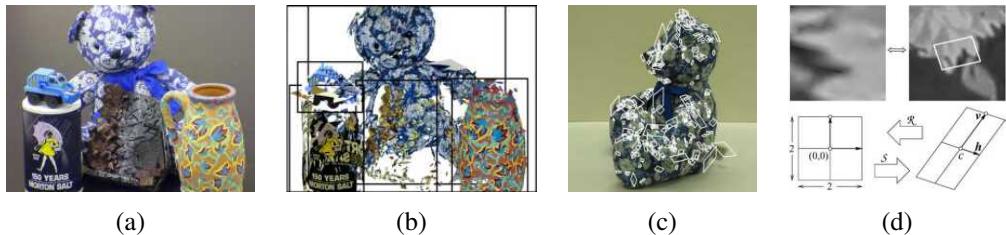


Figure 6.3 3D object recognition with affine regions (Rothganger, Lazebnik et al. 2006) © 2006 Springer: (a) sample input image; (b) five of the recognized (reprojected) objects along with their bounding boxes; (c) a few of the local affine regions; (d) local affine region (patch) reprojected into a canonical (square) frame, along with its geometric affine transformations.

Geometric alignment

To recognize one or more instances of some known objects, such as those shown in the left column of Figure 6.2, the recognition system first extracts a set of interest points in each database image and stores the associated descriptors (and original positions) in an indexing structure such as a search tree (Section 7.1.3). At recognition time, features are extracted from the new image and compared against the stored object features. Whenever a sufficient number of matching features (say, three or more) are found for a given object, the system then invokes a *match verification* stage, whose job is to determine whether the spatial arrangement of matching features is consistent with those in the database image.

Because images can be highly cluttered and similar features may belong to several objects, the original set of feature matches can have a large number of outliers. For this reason, Lowe (2004) suggests using a Hough transform (Section 7.4.2) to accumulate votes for likely geometric transformations. In his system, he uses an affine transformation between the database object and the collection of scene features, which works well for objects that are mostly planar, or where at least several corresponding features share a quasi-planar geometry.²

Another system that uses local affine frames is the one developed by Rothganger, Lazebnik et al. (2006). In their system, the affine region detector of Mikolajczyk and Schmid (2004) is used to rectify local image patches (Figure 6.3d), from which both a SIFT descriptor and a 10×10 UV color histogram are computed and used for matching and recognition. Corresponding patches in different views of the same object, along with their local affine deformations, are used to compute a 3D affine model for the object using an extension of

²When a larger number of features is available, a full fundamental matrix can be used (Brown and Lowe 2002; Gordon and Lowe 2006). When image stitching is being performed (Brown and Lowe 2007), the motion models discussed in Section 8.2.1 can be used instead.

the factorization algorithm of Section 11.4.1, which can then be upgraded to a Euclidean reconstruction (Tomasi and Kanade 1992). At recognition time, local Euclidean neighborhood constraints are used to filter potential matches, in a manner analogous to the affine geometric constraints used by Lowe (2004) and Obdržálek and Matas (2006). Figure 6.3 shows the results of recognizing five objects in a cluttered scene using this approach.

While feature-based approaches are normally used to detect and localize known objects in scenes, it is also possible to get pixel-level segmentations of the scene based on such matches. Ferrari, Tuytelaars, and Van Gool (2006b) describe such a system for simultaneously recognizing objects and segmenting scenes, while Kannala, Rahtu *et al.* (2008) extend this approach to non-rigid deformations. Section 6.4 re-visits this topic of joint recognition and segmentation in the context of generic class (category) recognition.

While instance recognition in the early to mid-2000s focused on the problem of locating a known 3D object in an image, as shown in Figures 6.2–6.3, attention shifted to the more challenging problem of *instance retrieval* (also known as *content-based image retrieval*), in which the number of images being searched can be very large. Section 7.1.4 reviews such techniques, a snapshot of which can be seen in Figure 7.27 and the survey by Zheng, Yang, and Tian (2018). This topic is also related to visual similarity search (Section 6.2.3) and 3D pose estimation (Section 11.2).

6.2 Image classification

While instance recognition techniques are relatively mature and are used in commercial applications such as traffic sign recognition (Stallkamp, Schlipsing *et al.* 2012), generic category (class) recognition is still a rapidly evolving research area. Consider for example the set of photographs in Figure 6.4a, which shows objects taken from 10 different visual categories. (I'll leave it up to you to name each of the categories.) How would you go about writing a program to categorize each of these images into the appropriate class, especially if you were also given the choice “none of the above”?

As you can tell from this example, visual category recognition is an extremely challenging problem. However, the progress in the field has been quite dramatic, if judged by how much better today's algorithms are compared to those of a decade ago.

In this section, we review the main classes of algorithms used for whole-image classification. We begin with classic feature-based approaches that rely on handcrafted features and their statistics, optionally using machine learning to do the final classification (Figure 5.2b). Since such techniques are no longer widely used, we present a fairly terse description of the most important techniques. More details can be found in the first edition of this book

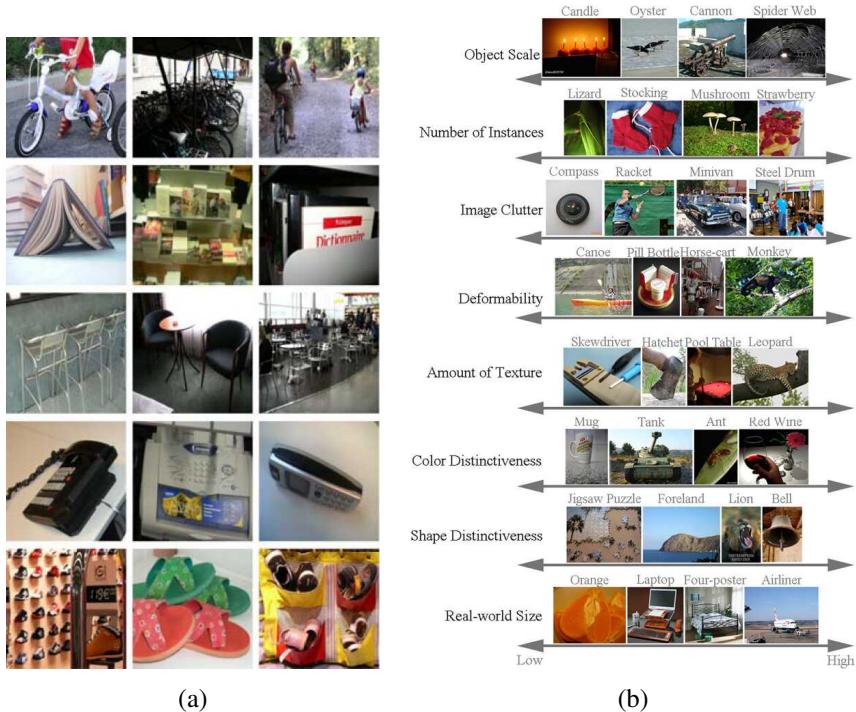


Figure 6.4 Challenges in image recognition: (a) sample images from the Xerox 10 class dataset (Csurka, Dance et al. 2006) © 2007 Springer; (b) axes of difficulty and variation from the ImageNet dataset (Russakovsky, Deng et al. 2015) © 2015 Springer.

(Szeliski 2010, Chapter 14) and in the cited journal papers and surveys. Next, we describe modern image classification systems, which are based on the deep neural networks we introduced in the previous chapter. We then describe visual similarity search, where the task is to find visually and semantically similar images, rather than classification into a fixed set of categories. Finally, we look at face recognition, since this topic has its own long history and set of techniques.

6.2.1 Feature-based methods

In this section, we review “classic” feature-based approaches to category recognition (image classification). While, historically, *part-based* representations and recognition algorithms (Section 6.2.1) were the preferred approach (Fischler and Elschlager 1973; Felzenszwalb and Huttenlocher 2005; Fergus, Perona, and Zisserman 2007), we begin by describing sim-

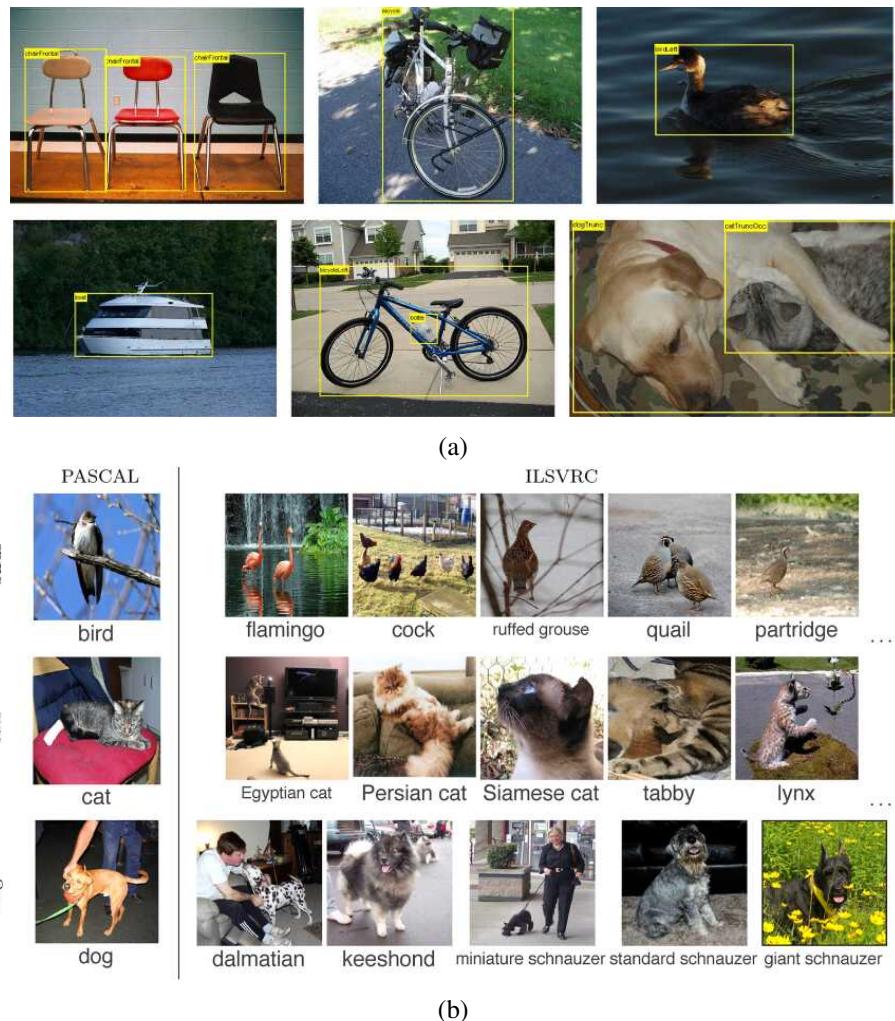


Figure 6.5 Sample images from two widely used image classification datasets: (a) Pascal Visual Object Categories (VOC) (Everingham, Eslami et al. 2015) © 2015 Springer; (b) ImageNet (Russakovsky, Deng et al. 2015) © 2015 Springer.

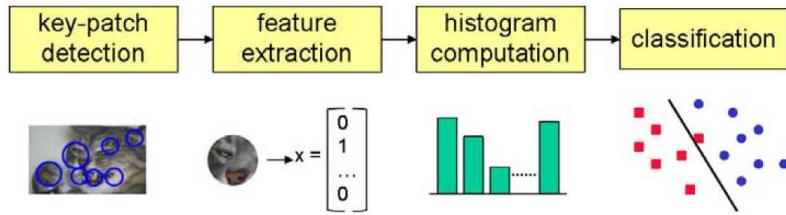


Figure 6.6 A typical processing pipeline for a bag-of-words category recognition system (Csurka, Dance *et al.* 2006) © 2007 Springer. Features are first extracted at keypoints and then quantized to get a distribution (histogram) over the learned visual words (feature cluster centers). The feature distribution histogram is used to learn a decision surface using a classification algorithm, such as a support vector machine.

pler *bag-of-features* approaches that represent objects and images as unordered collections of feature descriptors. We then review more complex systems constructed with part-based models, and then look at how context and scene understanding, as well as machine learning, can improve overall recognition results. Additional details on the techniques presented in this section can be found in older survey articles, paper collections, and courses (Pinz 2005; Ponce, Hebert *et al.* 2006; Dickinson, Leonardis *et al.* 2007; Fei-Fei, Fergus, and Torralba 2009), as well as two review articles on the PASCAL and ImageNet recognition challenges (Everingham, Van Gool *et al.* 2010; Everingham, Eslami *et al.* 2015; Russakovsky, Deng *et al.* 2015) and the first edition of this book (Szeliski 2010, Chapter 14).

Bag of words

One of the simplest algorithms for category recognition is the *bag of words* (also known as *bag of features* or *bag of keypoints*) approach (Csurka, Dance *et al.* 2004; Lazebnik, Schmid, and Ponce 2006; Csurka, Dance *et al.* 2006; Zhang, Marszalek *et al.* 2007). As shown in Figure 6.6, this algorithm simply computes the distribution (histogram) of visual words found in the query image and compares this distribution to those found in the training images. We will give more details of this approach in Section 7.1.4. The biggest difference from instance recognition is the absence of a geometric verification stage (Section 6.1), since individual instances of generic visual categories, such as those shown in Figure 6.4a, have relatively little spatial coherence to their features (but see the work by Lazebnik, Schmid, and Ponce (2006)).

Csurka, Dance *et al.* (2004) were the first to use the term *bag of keypoints* to describe such approaches and among the first to demonstrate the utility of frequency-based techniques for

category recognition. Their original system used affine covariant regions and SIFT descriptors, k-means visual vocabulary construction, and both a naïve Bayesian classifier and support vector machines for classification. (The latter was found to perform better.) Their newer system (Csurka, Dance *et al.* 2006) uses regular (non-affine) SIFT patches and boosting instead of SVMs and incorporates a small amount of geometric consistency information.

Zhang, Marszalek *et al.* (2007) perform a more detailed study of such bag of features systems. They compare a number of feature detectors (Harris–Laplace (Mikolajczyk and Schmid 2004) and Laplacian (Lindeberg 1998b)), descriptors (SIFT, RIFT, and SPIN (Lazebnik, Schmid, and Ponce 2005)), and SVM kernel functions.

Instead of quantizing feature vectors to visual words, Grauman and Darrell (2007b) develop a technique for directly computing an approximate distance between two variably sized collections of feature vectors. Their approach is to bin the feature vectors into a multi-resolution pyramid defined in feature space and count the number of features that land in corresponding bins B_{il} and B'_{il} . The distance between the two sets of feature vectors (which can be thought of as points in a high-dimensional space) is computed using histogram intersection between corresponding bins, while discounting matches already found at finer levels and weighting finer matches more heavily. In follow-on work, Grauman and Darrell (2007a) show how an explicit construction of the pyramid can be avoided using hashing techniques.

Inspired by this work, Lazebnik, Schmid, and Ponce (2006) show how a similar idea can be employed to augment bags of keypoints with loose notions of 2D spatial location analogous to the pooling performed by SIFT (Lowe 2004) and “gist” (Torralba, Murphy *et al.* 2003). In their work, they extract affine region descriptors (Lazebnik, Schmid, and Ponce 2005) and quantize them into visual words. (Based on previous results by Fei-Fei and Perona (2005), the feature descriptors are extracted densely (on a regular grid) over the image, which can be helpful in describing textureless regions such as the sky.) They then form a spatial pyramid of bins containing word counts (histograms) and use a similar pyramid match kernel to combine histogram intersection counts in a hierarchical fashion.

The debate about whether to use quantized feature descriptors or continuous descriptors and also whether to use sparse or dense features went on for many years. Boiman, Shechtman, and Irani (2008) show that if query images are compared to *all* the features representing a given class, rather than just each class image individually, nearest-neighbor matching followed by a naïve Bayes classifier outperforms quantized visual words. Instead of using generic feature detectors and descriptors, some authors have been investigating *learning* class-specific features (Ferencz, Learned-Miller, and Malik 2008), often using randomized forests (Philbin, Chum *et al.* 2007; Moosmann, Nowak, and Jurie 2008; Shotton, Johnson, and Cipolla 2008) or combining the feature generation and image classification stages (Yang,

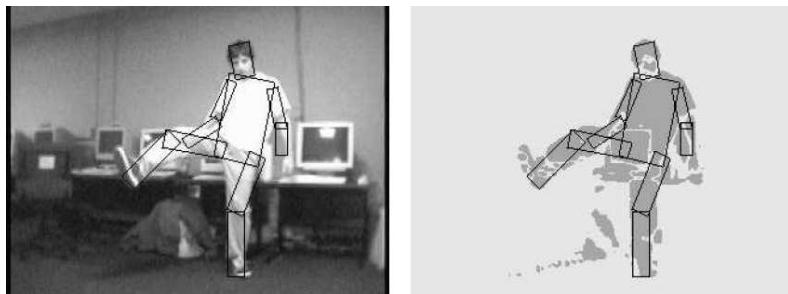


Figure 6.7 Using pictorial structures to locate and track a person (Felzenszwalb and Huttenlocher 2005) © 2005 Springer. The structure consists of articulated rectangular body parts (torso, head, and limbs) connected in a tree topology that encodes relative part positions and orientations. To fit a pictorial structure model, a binary silhouette image is first computed using background subtraction.

Jin *et al.* 2008). Others, such as Serre, Wolf, and Poggio (2005) and Mutch and Lowe (2008) use hierarchies of dense feature transforms inspired by biological (visual cortical) processing combined with SVMs for final classification.

Part-based models

Recognizing an object by finding its constituent parts and measuring their geometric relationships is one of the oldest approaches to object recognition (Fischler and Elschlager 1973; Kanade 1977; Yuille 1991). Part-based approaches were often used for face recognition (Moghaddam and Pentland 1997; Heisele, Ho *et al.* 2003; Heisele, Serre, and Poggio 2007) and continue being used for pedestrian detection (Figure 6.24) (Felzenszwalb, McAllester, and Ramanan 2008) and pose estimation (Güler, Neverova, and Kokkinos 2018).

In this overview, we discuss some of the central issues in part-based recognition, namely, the representation of geometric relationships, the representation of individual parts, and algorithms for learning such descriptions and recognizing them at run time. More details on part-based models for recognition can be found in the course notes by Fergus (2009).

The earliest approaches to representing geometric relationships were dubbed *pictorial structures* by Fischler and Elschlager (1973) and consisted of spring-like connections between different feature locations (Figure 6.1a). To fit a pictorial structure to an image, an energy function of the form

$$E = \sum_i V_i(\mathbf{l}_i) + \sum_{ij \in E} V_{ij}(\mathbf{l}_i, \mathbf{l}_j) \quad (6.1)$$

is minimized over all potential part locations or poses $\{\mathbf{l}_i\}$ and pairs of parts (i, j) for which an edge (geometric relationship) exists in E . Note how this energy is closely related to that used with Markov random fields (4.35–4.38), which can be used to embed pictorial structures in a probabilistic framework that makes parameter learning easier (Felzenszwalb and Huttenlocher 2005).

Part-based models can have different topologies for the geometric connections between the parts (Carneiro and Lowe 2006). For example, Felzenszwalb and Huttenlocher (2005) restrict the connections to a tree, which makes learning and inference more tractable. A tree topology enables the use of a recursive Viterbi (dynamic programming) algorithm (Pearl 1988; Bishop 2006), in which leaf nodes are first optimized as a function of their parents, and the resulting values are then plugged in and eliminated from the energy function. To further increase the efficiency of the inference algorithm, Felzenszwalb and Huttenlocher (2005) restrict the pairwise energy functions $V_{ij}(\mathbf{l}_i, \mathbf{l}_j)$ to be Mahalanobis distances on functions of location variables and then use fast distance transform algorithms to minimize each pairwise interaction in time that is closer to linear in N .

Figure 6.7 shows the results of using their pictorial structures algorithm to fit an articulated body model to a binary image obtained by background segmentation. In this application of pictorial structures, parts are parameterized by the locations, sizes, and orientations of their approximating rectangles. Unary matching potentials $V_i(\mathbf{l}_i)$ are determined by counting the percentage of foreground and background pixels inside and just outside the tilted rectangle representing each part.

A large number of different graphical models have been proposed for part-based recognition. Carneiro and Lowe (2006) discuss a number of these models and propose one of their own, which they call a *sparse flexible model*; it involves ordering the parts and having each part's location depend on at most k of its ancestor locations.

The simplest models are bags of words, where there are no geometric relationships between different parts or features. While such models can be very efficient, they have a very limited capacity to express the spatial arrangement of parts. Trees and stars (a special case of trees where all leaf nodes are directly connected to a common root) are the most efficient in terms of inference and hence also learning (Felzenszwalb and Huttenlocher 2005; Fergus, Perona, and Zisserman 2005; Felzenszwalb, McAllester, and Ramanan 2008). Directed acyclic graphs come next in terms of complexity and can still support efficient inference, although at the cost of imposing a causal structure on the part model (Bouchard and Triggs 2005; Carneiro and Lowe 2006). k -fans, in which a clique of size k forms the root of a star-shaped model have inference complexity $O(N^{k+1})$, although with distance transforms and Gaussian priors, this can be lowered to $O(N^k)$ (Crandall, Felzenszwalb, and Huttenlocher

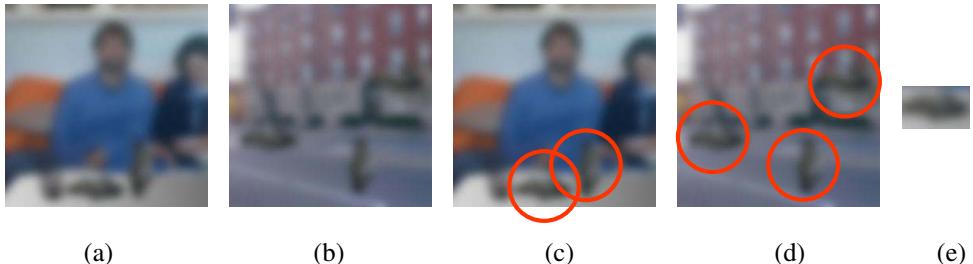


Figure 6.8 *The importance of context (images courtesy of Antonio Torralba). Can you name all of the objects in images (a–b), especially those that are circled in (c–d). Look carefully at the circled objects. Did you notice that they all have the same shape (after being rotated), as shown in column (e)?*

2005; Crandall and Huttenlocher 2006). Finally, fully connected *constellation* models are the most general, but the assignment of features to parts becomes intractable for moderate numbers of parts P , since the complexity of such an assignment is $O(N^P)$ (Fergus, Perona, and Zisserman 2007).

The original constellation model was developed by Burl, Weber, and Perona (1998) and consists of a number of parts whose relative positions are encoded by their mean locations and a full covariance matrix, which is used to denote not only positional uncertainty but also potential correlations between different parts. Weber, Welling, and Perona (2000) extended this technique to a weakly supervised setting, where both the appearance of each part and its locations are automatically learned given whole image labels. Fergus, Perona, and Zisserman (2007) further extend this approach to simultaneous learning of appearance and shape models from scale-invariant keypoint detections.

The part-based approach to recognition has also been extended to learning new categories from small numbers of examples, building on recognition components developed for other classes (Fei-Fei, Fergus, and Perona 2006). More complex hierarchical part-based models can be developed using the concept of grammars (Bouchard and Triggs 2005; Zhu and Mumford 2006). A simpler way to use parts is to have keypoints that are recognized as being part of a class vote for the estimated part locations (Leibe, Leonardis, and Schiele 2008). Parts can also be a useful component of fine-grained category recognition systems, as shown in Figure 6.9.

Context and scene understanding

Thus far, we have mostly considered the task of recognizing and localizing objects in isolation from that of understanding the scene (context) in which the object occur. This is a big

limitation, as context plays a very important role in human object recognition (Oliva and Torralba 2007). Context can greatly improve the performance of object recognition algorithms (Divvala, Hoiem *et al.* 2009), as well as providing useful semantic clues for general scene understanding (Torralba 2008).

Consider the two photographs in Figure 6.8a–b. Can you name all of the objects, especially those circled in images (c–d)? Now have a closer look at the circled objects. Do see any similarity in their shapes? In fact, if you rotate them by 90°, they are all the same as the “blob” shown in Figure 6.8e. So much for our ability to recognize object by their shape!

Even though we have not addressed context explicitly earlier in this chapter, we have already seen several instances of this general idea being used. A simple way to incorporate spatial information into a recognition algorithm is to compute feature statistics over different regions, as in the spatial pyramid system of Lazebnik, Schmid, and Ponce (2006). Part-based models (Figure 6.7) use a kind of local context, where various parts need to be arranged in a proper geometric relationship to constitute an object.

The biggest difference between part-based and context models is that the latter combine objects into scenes and the number of constituent objects from each class is not known in advance. In fact, it is possible to combine part-based and context models into the same recognition architecture (Murphy, Torralba, and Freeman 2003; Suderth, Torralba *et al.* 2008; Crandall and Huttenlocher 2007).

Consider an image database consisting of street and office scenes. If we have enough training images with labeled regions, such as buildings, cars, and roads, or monitors, keyboards, and mice, we can develop a geometric model for describing their relative positions. Suderth, Torralba *et al.* (2008) develop such a model, which can be thought of as a two-level constellation model. At the top level, the distributions of objects relative to each other (say, buildings with respect to cars) is modeled as a Gaussian. At the bottom level, the distribution of parts (affine covariant features) with respect to the object center is modeled using a mixture of Gaussians. However, since the number of objects in the scene and parts in each object are unknown, a *latent Dirichlet process* (LDP) is used to model object and part creation in a generative framework. The distributions for all of the objects and parts are learned from a large labeled database and then later used during inference (recognition) to label the elements of a scene.

Another example of context is in simultaneous segmentation and recognition (Section 6.4 and Figure 6.33), where the arrangements of various objects in a scene are used as part of the labeling process. Torralba, Murphy, and Freeman (2004) describe a conditional random field where the estimated locations of building and roads influence the detection of cars, and where boosting is used to learn the structure of the CRF. Rabinovich, Vedaldi *et al.* (2007)

use context to improve the results of CRF segmentation by noting that certain adjacencies (relationships) are more likely than others, e.g., a person is more likely to be on a horse than on a dog. Galleguillos and Belongie (2010) review various approaches proposed for adding context to object categorization, while Yao and Fei-Fei (2012) study human-object interactions. (For a more recent take on this problem, see Gkioxari, Girshick *et al.* (2018).)

Context also plays an important role in 3D inference from single images (Figure 6.41), using computer vision techniques for labeling pixels as belonging to the ground, vertical surfaces, or sky (Hoiem, Efros, and Hebert 2005a). This line of work has been extended to a more holistic approach that simultaneously reasons about object identity, location, surface orientations, occlusions, and camera viewing parameters (Hoiem, Efros, and Hebert 2008).

A number of approaches use the *gist* of a scene (Torralba 2003; Torralba, Murphy *et al.* 2003) to determine where instances of particular objects are likely to occur. For example, Murphy, Torralba, and Freeman (2003) train a regressor to predict the vertical locations of objects such as pedestrians, cars, and buildings (or screens and keyboards for indoor office scenes) based on the gist of an image. These location distributions are then used with classic object detectors to improve the performance of the detectors. Gists can also be used to directly match complete images, as we saw in the scene completion work of Hays and Efros (2007).

Finally, some of the work in scene understanding exploits the existence of large numbers of labeled (or even unlabeled) images to perform matching directly against whole images, where the images themselves implicitly encode the expected relationships between objects (Russell, Torralba *et al.* 2007; Malisiewicz and Efros 2008; Galleguillos and Belongie 2010). This, of course, is one of the central benefits of using deep neural networks, which we discuss in the next section.

6.2.2 Deep networks

As we saw in Section 5.4.3, deep networks started outperforming “shallow” learning-based approaches on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) with the introduction of the “AlexNet” SuperVision system of Krizhevsky, Sutskever, and Hinton (2012). Since that time, recognition accuracy has continued to improve dramatically (Figure 5.40) driven to a large degree by deeper networks and better training algorithms. More recently, more efficient networks have become the focus of research (Figure 5.45) as well as larger (unlabeled) training datasets (Section 5.4.7). There are now open-source frameworks such as Classy Vision³ for training and fine tuning your own image and video classification models. Users can also upload custom images on the web to the Computer Vision Explorer⁴

³<https://classyvision.ai>

⁴<https://vision-explorer.allenai.org>

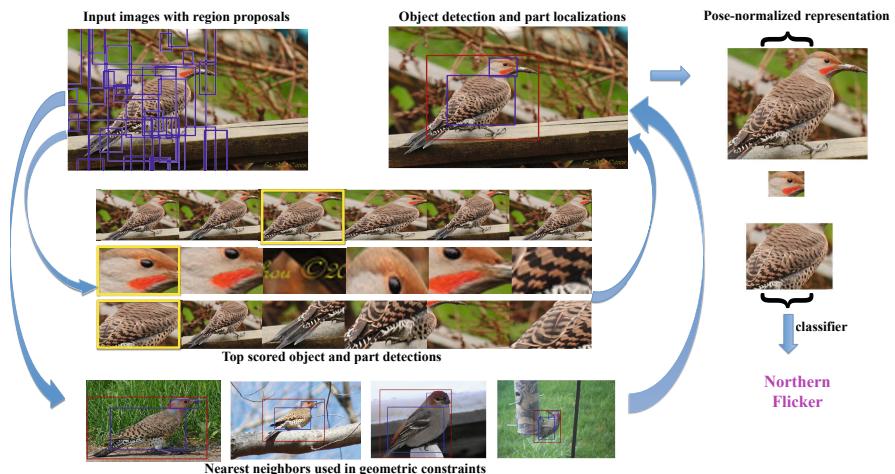


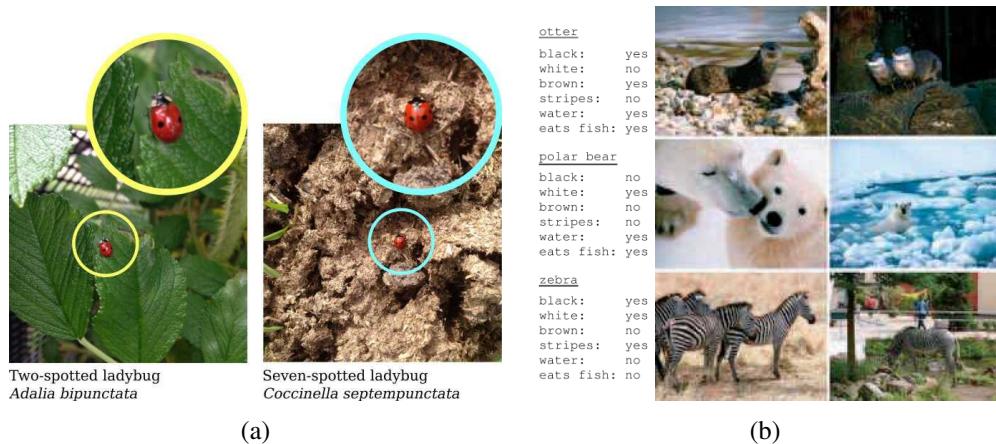
Figure 6.9 Fine-grained category recognition using parts (Zhang, Donahue et al. 2014) © 2014 Springer. Deep neural network object and part detectors are trained and their outputs are combined using geometric constraints. A classifier trained on features from the extracted parts is used for the final categorization.

to see how well many popular computer vision models perform on their own images.

In addition to recognizing commonly occurring categories such as those found in the ImageNet and COCO datasets, researchers have studied the problem of *fine-grained* category recognition (Duan, Parikh et al. 2012; Zhang, Donahue et al. 2014; Krause, Jin et al. 2015), where the differences between sub-categories can be subtle and the number of exemplars is quite low (Figure 6.9). Examples of categories with fine-grained sub-classes include flowers (Nilsback and Zisserman 2006), cats and dogs (Parkhi, Vedaldi et al. 2012), birds (Wah, Branson et al. 2011; Van Horn, Branson et al. 2015), and cars (Yang, Luo et al. 2015). A recent example of fine-grained categorization is the iNaturalist system (Van Horn, Mac Aodha et al. 2018),⁵ which allows both specialists and citizen scientists to photograph and label biological species, using a fine-grained category recognition system to label new images (Figure 6.10a).

Fine-grained categorization is often attacked using *attributes* of images and classes (Lampert, Nickisch, and Harmeling 2009; Parikh and Grauman 2011; Lampert, Nickisch, and Harmeling 2014), as shown in Figure 6.10b. Extracting attributes can enable *zero-shot learning* (Xian, Lampert et al. 2019), where previously unseen categories can be described using combinations of such attributes. However, some caution must be used in order not to

⁵<https://www.inaturalist.org>



(a)

(b)

Figure 6.10 Fine-grained category recognition. (a) The iNaturalist website and app allows citizen scientists to collect and classify images on their phones (Van Horn, Mac Aodha et al. 2018) © 2018 IEEE. (b) Attributes can be used for fine-grained categorization and zero-shot learning (Lampert, Nickisch, and Harmeling 2014) © 2014 Springer. These images are part of the Animals with Attributes dataset.

learn spurious correlations between different attributes (Jayaraman, Sha, and Grauman 2014) or between objects and their common contexts (Singh, Mahajan et al. 2020). Fine-grained recognition can also be tackled using metric learning (Wu, Manmatha et al. 2017) or nearest-neighbor visual similarity search (Touvron, Sablayrolles et al. 2020), which we discuss next.

6.2.3 Application: Visual similarity search

Automatically classifying images into categories and tagging them with attributes using computer vision algorithms makes it easier to find them in catalogues and on the web. This is commonly used in *image search* or *image retrieval* engines, which find likely images based on keywords, just as regular web search engines find relevant documents and pages.

Sometimes, however, it's easier to find the information you need from an image, i.e., using *visual search*. Examples of this include fine-grained categorization, which we have just seen, as well as instance retrieval, i.e., finding the exact same object (Section 6.1) or location (Section 11.2.3). Another variant is finding visually similar images (often called *visual similarity search* or *reverse image search*), which is useful when the search intent cannot be succinctly captured in words.⁶

⁶Some authors use the term *image retrieval* to denote visual similarity search, (e.g., Jégou, Perronnin et al. 2012;



Figure 6.11 The GrokNet product recognition service is used for product tagging, visual search, and recommendations © Bell, Liu et al. (2020): (a) recognizing all the products in a photo; (b) automatically sourcing data for metric learning using weakly supervised data augmentation.

The topic of searching by visual similarity has a long history and goes by a variety of names, including query by image content (QBIC) (Flickner, Sawhney *et al.* 1995) and content-based image retrieval (CBIR) (Smeulders, Worring *et al.* 2000; Lew, Sebe *et al.* 2006; Vasconcelos 2007; Datta, Joshi *et al.* 2008). Early publications in these fields were based primarily on simple whole-image similarity metrics, such as color and texture (Swain and Ballard 1991; Jacobs, Finkelstein, and Salesin 1995; Manjunathi and Ma 1996).

Later architectures, such as that by Fergus, Perona, and Zisserman (2004), use a feature-based learning and recognition algorithm to re-rank the outputs from a traditional keyword-based image search engine. In follow-on work, Fergus, Fei-Fei *et al.* (2005) cluster the results returned by image search using an extension of probabilistic latent semantic analysis (PLSA) (Hofmann 1999) and then select the clusters associated with the highest ranked results as the representative images for that category. Other approaches rely on carefully annotated image databases such as LabelMe (Russell, Torralba *et al.* 2008). For example, Malisiewicz and Efros (2008) describe a system that, given a query image, can find similar LabelMe images, whereas Liu, Yuen, and Torralba (2009) combine feature-based correspondence algorithms with the labeled database to perform simultaneous recognition and segmentation.

Newer approaches to visual similarity search use whole-image descriptors such as Fisher kernels and the Vector of Locally Aggregated Descriptors (VLAD) (Jégou, Perronnin *et al.* 2012) or pooled CNN activations (Babenko and Lempitsky 2015a; Tolias, Sicre, and Jégou 2016; Cao, Araujo, and Sim 2020; Ng, Balntas *et al.* 2020; Tolias, Jenicek, and Chum 2020) combined with metric learning (Bell and Bala 2015; Song, Xiang *et al.* 2016; Gordo, Almazán *et al.* 2017; Wu, Manmatha *et al.* 2017; Berman, Jégou *et al.* 2019) to represent each image with a compact descriptor that can be used to measure similarity in large databases (Johnson, Douze, and Jégou 2021). It is also possible to combine several techniques, such as deep networks with VLAD (Arandjelovic, Gronat *et al.* 2016), generalized mean (GeM)

Radenović, Tolias, and Chum 2019).

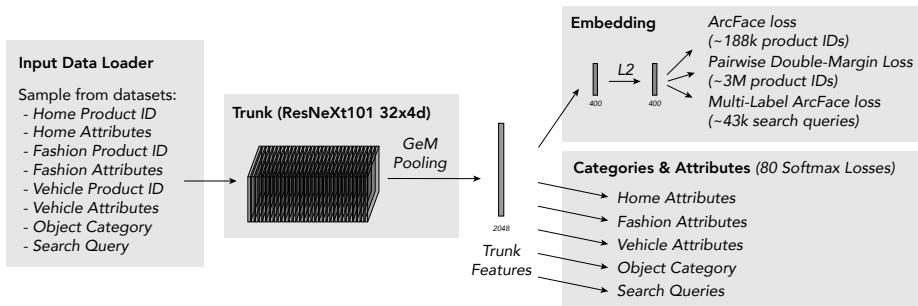


Figure 6.12 The GrokNet training architecture uses seven datasets, a common DNN trunk, two branches, and 83 loss functions (80 categorical losses + 3 embedding losses) © Bell, Liu et al. (2020).

pooling (Radenović, Tolias, and Chum 2019), or dynamic mean (DAME) pooling (Yang, Kien Nguyen *et al.* 2019) into complete systems that are end-to-end tunable. Gordo, Almazán *et al.* (2017) provide a comprehensive review and experimental comparison of many of these techniques, which we also discuss in Section 7.1.4 on large-scale matching and retrieval. Some of the latest techniques for image retrieval use combinations of local and global descriptors to obtain state-of-the art performance on the landmark recognition tasks (Cao, Araujo, and Sim 2020; Ng, Balntas *et al.* 2020; Tolias, Jenicek, and Chum 2020). The ECCV 2020 Workshop on Instance-Level Recognition⁷ has pointers to some of the latest work in this area, while the upcoming NeurIPS’21 Image Similarity Challenge⁸ has new datasets for detecting content manipulation.

A recent example of a commercial system that uses visual similarity search, in addition to category recognition, is the GrokNet product recognition service described by Bell, Liu *et al.* (2020). GrokNet takes as input user images and shopping queries and returns indexed items similar to the ones in the query image (Figure 6.11a). The reason for needing a similarity search component is that the world contains too many “long-tail” items such as “a fur sink, an electric dog polisher, or a gasoline powered turtleneck sweater”,⁹ to make full categorization practical.

At training time, GrokNet takes both weakly labeled images, with category and/or attribute labels, and unlabeled images, where features in objects are detected and then used for metric learning, using a modification of ArcFace loss (Deng, Guo *et al.* 2019) and a novel

⁷<https://ilr-workshop.github.io/ECCVW2020>

⁸<https://www.drivendata.org/competitions/79/>

⁹<https://www.google.com/search?q=gasoline+powered+turtleneck+sweater>



Figure 6.13 Humans can recognize low-resolution faces of familiar people (Sinha, Balas et al. 2006) © 2006 IEEE.

pairwise margin loss (Figure 6.11b). The overall system takes in large collections of unlabeled and weakly labeled images and trains a ResNeXt101 trunk using a combination of category and attribute softmax losses and three different metric losses on the embeddings (Figure 6.12). GrokNet is just one example of a large number of commercial visual product search systems that have recently been developed. Others include systems from Amazon (Wu, Manmatha et al. 2017), Pinterest (Zhai, Wu et al. 2019), and Facebook (Tang, Borisyuk et al. 2019). In addition to helping people find items they may wish to purchase, large-scale similarity search can also speed the search for harmful content on the web, as exemplified in Facebook’s SimSearchNet.¹⁰

6.2.4 Face recognition

Among the various recognition tasks that computers are asked to perform, face recognition is the one where they have arguably had the most success.¹¹ While even people cannot readily distinguish between similar people with whom they are not familiar (O’Toole, Jiang et al. 2006; O’Toole, Phillips et al. 2009), computers’ ability to distinguish among a small number of family members and friends has found its way into consumer-level photo applications. Face recognition can be used in a variety of additional applications, including human-computer interaction (HCI), identity verification (Kirovski, Jovic, and Jancke 2004), desktop login, parental controls, and patient monitoring (Zhao, Chellappa et al. 2003), but it also has the potential for misuse (Chokshi 2019; Ovide 2020).

Face recognizers work best when they are given images of faces under a wide variety of

¹⁰<https://ai.facebook.com/blog/using-ai-to-detect-covid-19-misinformation-and-exploitative-content>

¹¹Instance recognition, i.e., the re-recognition of known objects such as locations or planar objects, is the other most successful application of general image recognition. In the general domain of *biometrics*, i.e., identity recognition, specialized images such as irises and fingerprints perform even better (Jain, Bolle, and Pankanti 1999; Daugman 2004).

| Name/URL | Contents/Reference |
|--|---|
| CMU Multi-PIE database http://www.cs.cmu.edu/afs/cs/project/PIE/MultiPie | 337 people's faces in various poses Gross, Matthews <i>et al.</i> (2010) |
| Faces in the Wild http://vis-www.cs.umass.edu/lfw | 5,749 internet celebrities Huang, Ramesh <i>et al.</i> (2007) |
| YouTube Faces (YTF) https://www.cs.tau.ac.il/~wolf/ytfaces | 1,595 people in 3,425 YouTube videos Wolf, Hassner, and Maoz (2011) |
| MegaFace https://megaface.cs.washington.edu | 1M internet faces Nech and Kemelmacher-Shlizerman (2017) |
| IARPA Janus Benchmark (IJB) https://www.nist.gov/programs-projects/face-challenges | 31,334 faces of 3,531 people in videos Maze, Adams <i>et al.</i> (2018) |
| WIDER FACE http://shuoyang1213.me/WIDERFACE | 32,203 images for face <i>detection</i> Yang, Luo <i>et al.</i> (2016) |

Table 6.1 Face recognition and detection datasets, adapted from Maze, Adams *et al.* (2018).

pose, illumination, and expression (PIE) conditions (Phillips, Moon *et al.* 2000; Sim, Baker, and Bsat 2003; Gross, Shi, and Cohn 2005; Huang, Ramesh *et al.* 2007; Phillips, Scruggs *et al.* 2010). More recent widely used datasets include labeled Faces in the Wild (LFW) (Huang, Ramesh *et al.* 2007; Learned-Miller, Huang *et al.* 2016), YouTube Faces (YTF) (Wolf, Hassner, and Maoz 2011), MegaFace (Kemelmacher-Shlizerman, Seitz *et al.* 2016; Nech and Kemelmacher-Shlizerman 2017), and the IARPA Janus Benchmark (IJB) (Klare, Klein *et al.* 2015; Maze, Adams *et al.* 2018), as tabulated in Table 6.1. (See Masi, Wu *et al.* (2018) for additional datasets used for training.)

Some of the earliest approaches to face recognition involved finding the locations of distinctive image features, such as the eyes, nose, and mouth, and measuring the distances between these feature locations (Fischler and Elschlager 1973; Kanade 1977; Yuille 1991). Other approaches relied on comparing gray-level images projected onto lower dimensional subspaces called *eigenfaces* (Section 5.2.3) and jointly modeling shape and appearance variations (while discounting pose variations) using *active appearance models* (Section 6.2.4). Descriptions of “classic” (pre-DNN) face recognition systems can be found in a number of surveys and books on this topic (Chellappa, Wilson, and Sirohey 1995; Zhao, Chellappa *et al.* 2003; Li and Jain 2005) as well as the Face Recognition website.¹² The survey on face recognition by humans by Sinha, Balas *et al.* (2006) is also well worth reading; it includes a number of surprising results, such as humans’ ability to recognize low-resolution images of familiar

¹²<https://www.face-rec.org>

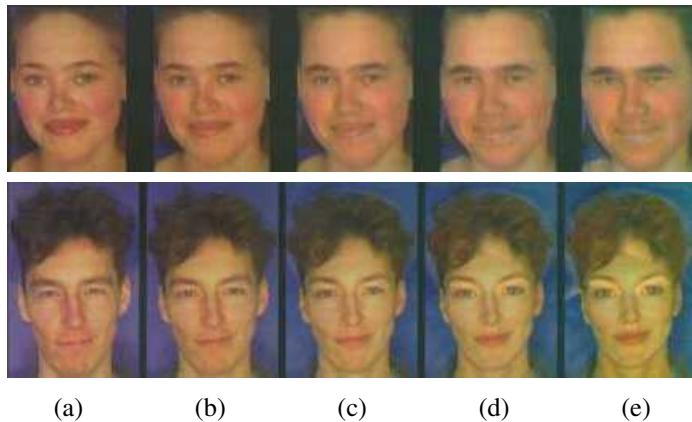


Figure 6.14 Manipulating facial appearance through shape and color (Rowland and Pentreath 1995) © 1995 IEEE. By adding or subtracting gender-specific shape and color characteristics to an input image (b), different amounts of gender variation can be induced. The amounts added (from the mean) are: (a) +50% (gender enhancement), (b) 0% (original image), (c) -50% (near “androgyny”), (d) -100% (gender switched), and (e) -150% (opposite gender attributes enhanced).

faces (Figure 6.13) and the importance of eyebrows in recognition. Researchers have also studied the automatic recognition of facial expressions. See Chang, Hu *et al.* (2006), Shan, Gong, and McOwan (2009), and Li and Deng (2020) for some representative papers.

Active appearance and 3D shape models

The need to use modular or view-based eigenspaces for face recognition, which we discussed in Section 5.2.3, is symptomatic of a more general observation, i.e., that facial appearance and identifiability depend as much on *shape* as they do on color or texture (which is what eigenfaces capture). Furthermore, when dealing with 3D head rotations, the *pose* of a person’s head should be discounted when performing recognition.

In fact, the earliest face recognition systems, such as those by Fischler and Elschlager (1973), Kanade (1977), and Yuille (1991), found distinctive feature points on facial images and performed recognition on the basis of their relative positions or distances. Later techniques such as *local feature analysis* (Penev and Atick 1996) and *elastic bunch graph matching* (Wiskott, Fellous *et al.* 1997) combined local filter responses (jets) at distinctive feature locations together with shape models to perform recognition.

A visually compelling example of why both shape and texture are important is the work

of Rowland and Perrett (1995), who manually traced the contours of facial features and then used these contours to normalize (warp) each image to a canonical shape. After analyzing both the shape and color images for deviations from the mean, they were able to associate certain shape and color deformations with personal characteristics such as age and gender (Figure 6.14). Their work demonstrates that both shape and color have an important influence on the perception of such characteristics.

Around the same time, researchers in computer vision were beginning to use simultaneous shape deformations and texture interpolation to model the variability in facial appearance caused by identity or expression (Beymer 1996; Vetter and Poggio 1997), developing techniques such as Active Shape Models (Lanitis, Taylor, and Cootes 1997), 3D Morphable Models (Blanz and Vetter 1999; Egger, Smith *et al.* 2020), and Elastic Bunch Graph Matching (Wiskott, Fellous *et al.* 1997).¹³

The *active appearance models* (AAMs) of Cootes, Edwards, and Taylor (2001) model both the variation in the shape of an image \mathbf{s} , which is normally encoded by the location of key feature points on the image, as well as the variation in texture \mathbf{t} , which is normalized to a canonical shape before being analyzed. Both shape and texture are represented as deviations from a mean shape $\bar{\mathbf{s}}$ and texture $\bar{\mathbf{t}}$,

$$\mathbf{s} = \bar{\mathbf{s}} + \mathbf{U}_s \mathbf{a} \quad (6.2)$$

$$\mathbf{t} = \bar{\mathbf{t}} + \mathbf{U}_t \mathbf{a}, \quad (6.3)$$

where the eigenvectors in \mathbf{U}_s and \mathbf{U}_t have been pre-scaled (whitened) so that unit vectors in \mathbf{a} represent one standard deviation of variation observed in the training data. In addition to these principal deformations, the shape parameters are transformed by a global similarity to match the location, size, and orientation of a given face. Similarly, the texture image contains a scale and offset to best match novel illumination conditions.

As you can see, the same appearance parameters \mathbf{a} in (6.2–6.3) simultaneously control both the shape and texture deformations from the mean, which makes sense if we believe them to be correlated. Figure 6.15 shows how moving three standard deviations along each of the first four principal directions ends up changing several correlated factors in a person’s appearance, including expression, gender, age, and identity.

Although active appearance models are primarily designed to accurately capture the variability in appearance and deformation that are characteristic of faces, they can be adapted to face recognition by computing an identity subspace that separates variation in identity from other sources of variability such as lighting, pose, and expression (Costen, Cootes *et al.*

¹³We will look at the application of PCA to 3D head and face modeling and animation in Section 13.6.3.

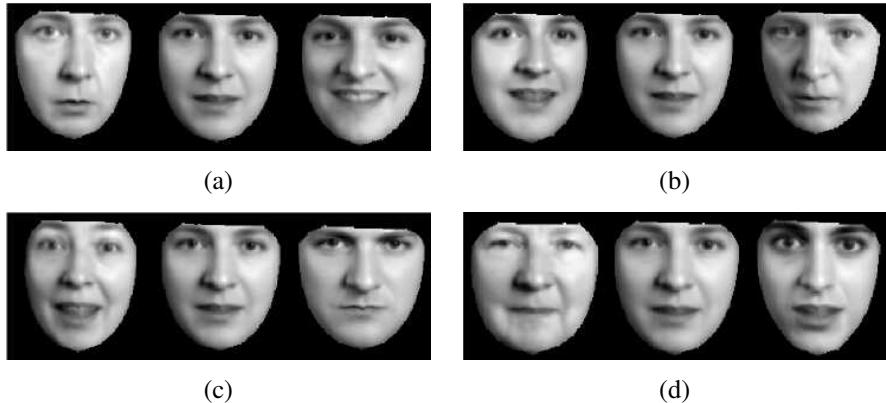


Figure 6.15 Principal modes of variation in active appearance models (Cootes, Edwards, and Taylor 2001) © 2001 IEEE. The four images show the effects of simultaneously changing the first four modes of variation in both shape and texture by $\pm\sigma$ from the mean. You can clearly see how the shape of the face and the shading are simultaneously affected.

1999). The basic idea, which is modeled after similar work in eigenfaces (Belhumeur, Hespanha, and Kriegman 1997; Moghaddam, Jebara, and Pentland 2000), is to compute separate statistics for intrapersonal and extrapersonal variation and then find discriminating directions in these subspaces. While AAMs have sometimes been used directly for recognition (Blanz and Vetter 2003), their main use in the context of recognition is to align faces into a canonical pose (Liang, Xiao *et al.* 2008; Ren, Cao *et al.* 2014) so that more traditional methods of face recognition (Penev and Atick 1996; Wiskott, Fellous *et al.* 1997; Ahonen, Hadid, and Pietikäinen 2006; Zhao and Pietikäinen 2007; Cao, Yin *et al.* 2010) can be used.

Active appearance models have been extended to deal with illumination and viewpoint variation (Gross, Baker *et al.* 2005) as well as occlusions (Gross, Matthews, and Baker 2006). One of the most significant extensions is to construct 3D models of shape (Matthews, Xiao, and Baker 2007), which are much better at capturing and explaining the full variability of facial appearance across wide changes in pose. Such models can be constructed either from monocular video sequences (Matthews, Xiao, and Baker 2007), as shown in Figure 6.16a, or from multi-view video sequences (Ramnath, Koterba *et al.* 2008), which provide even greater reliability and accuracy in reconstruction and tracking (Murphy-Chutorian and Trivedi 2009).

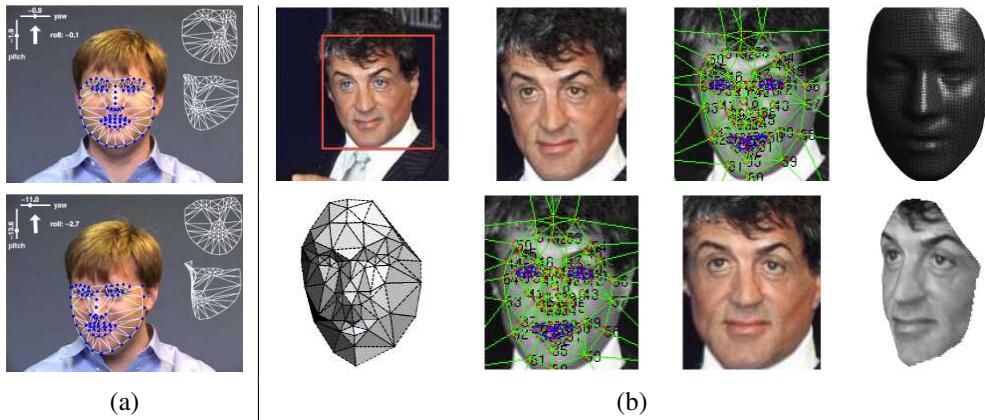


Figure 6.16 Head tracking and frontalization: (a) using 3D active appearance models (AAMs) (Matthews, Xiao, and Baker 2007) © 2007 Springer, showing video frames along with the estimated yaw, pitch, and roll parameters and the fitted 3D deformable mesh; (b) using six and then 67 fiducial points in the DeepFace system (Taigman, Yang et al. 2014) © 2014 IEEE, used to frontalize the face image (bottom row).

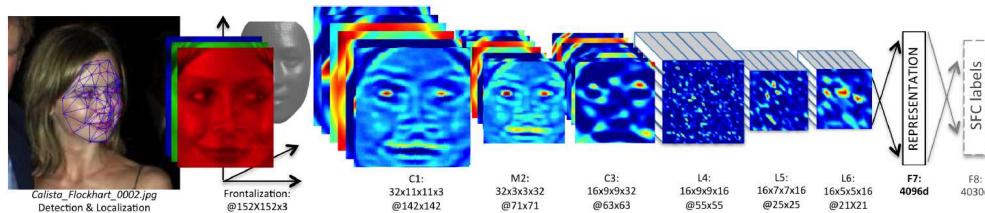


Figure 6.17 The DeepFace architecture (Taigman, Yang et al. 2014) © 2014 IEEE, starts with a frontalization stage, followed by several locally connected (non-convolutional) layers, and then two fully connected layers with a K -class softmax.

Facial recognition using deep learning

Prompted by the dramatic success of deep networks in whole-image categorization, face recognition researchers started using deep neural network backbones as part of their systems. Figures 6.16b–6.17 shows two stages in the DeepFace system of Taigman, Yang *et al.* (2014), which was one of the first systems to realize large gains using deep networks. In their system, a landmark-based pre-processing *frontalization* step is used to convert the original color image into a well-cropped front-looking face. Then, a deep locally connected network (where the convolution kernels can vary spatially) is fed into two final fully connected layers before classification.

Some of the more recent deep face recognizers omit the frontalization stage and instead use data augmentation (Section 5.3.3) to create synthetic inputs with a larger variety of poses (Schroff, Kalenichenko, and Philbin 2015; Parkhi, Vedaldi, and Zisserman 2015). Masi, Wu *et al.* (2018) provide an excellent tutorial and survey on deep face recognition, including a list of widely used training and testing datasets, a discussion of frontalization and dataset augmentation, and a section on training losses (Figure 6.18). This last topic is central to the ability to scale to larger and larger numbers of people. Schroff, Kalenichenko, and Philbin (2015) and Parkhi, Vedaldi, and Zisserman (2015) use triplet losses to construct a low-dimensional embedding space that is independent of the number of subjects. More recent systems use contrastive losses inspired by the softmax function, which we discussed in Section 5.3.4. For example, the ArcFace paper by Deng, Guo *et al.* (2019) measures angular distances on the unit hypersphere in the embedding space and adds an extra margin to get identities to clump together. This idea has been further extended for visual similarity search (Bell, Liu *et al.* 2020) and face recognition (Huang, Shen *et al.* 2020; Deng, Guo *et al.* 2020a).

Personal photo collections

In addition to digital cameras automatically finding faces to aid in auto-focusing and video cameras finding faces in video conferencing to center on the speaker (either mechanically or digitally), face detection has found its way into most consumer-level photo organization packages and photo sharing sites. Finding faces and allowing users to tag them makes it easier to find photos of selected people at a later date or to automatically share them with friends. In fact, the ability to tag friends in photos is one of the more popular features on Facebook.

Sometimes, however, faces can be hard to find and recognize, especially if they are small, turned away from the camera, or otherwise occluded. In such cases, combining face recognition with person detection and clothes recognition can be very effective, as illustrated in Figure 6.19 (Sivic, Zitnick, and Szeliski 2006). Combining person recognition with other

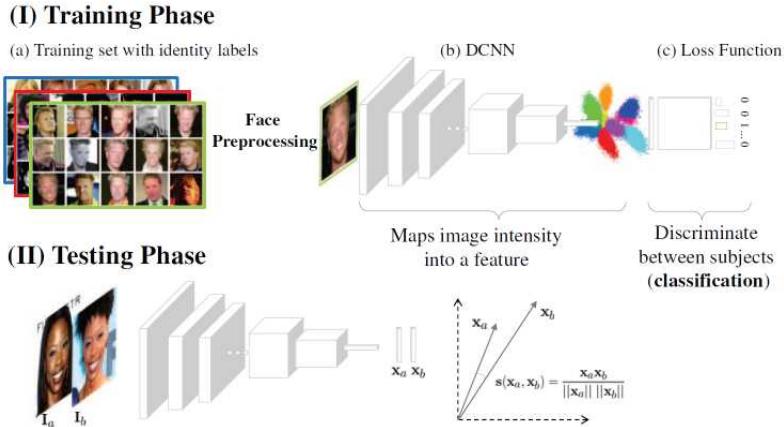


Figure 6.18 A typical modern deep face recognition architecture, from the survey by Masi, Wu et al. (2018) © 2018 IEEE. At training time, a huge labeled face set (a) is used to constrain the weights of a DCNN (b), optimizing a loss function (c) for a classification task. At test time, the classification layer is often discarded, and the DCNN is used as a feature extractor for comparing face descriptors.

kinds of context, such as location recognition (Section 11.2.3) or activity or event recognition, can also help boost performance (Lin, Kapoor *et al.* 2010).

6.3 Object detection

If we are given an image to analyze, such as the group portrait in Figure 6.20, we could try to apply a recognition algorithm to every possible sub-window in this image. Such algorithms are likely to be both slow and error-prone. Instead, it is more effective to construct special-purpose *detectors*, whose job it is to rapidly find likely regions where particular objects might occur.

We begin this section with face detectors, which were some of the earliest successful examples of recognition. Such algorithms are built into most of today's digital cameras to enhance auto-focus and into video conferencing systems to control panning and zooming. We then look at pedestrian detectors, as an example of more general methods for object detection. Finally, we turn to the problem of multi-class object detection, which today is solved using deep neural networks.

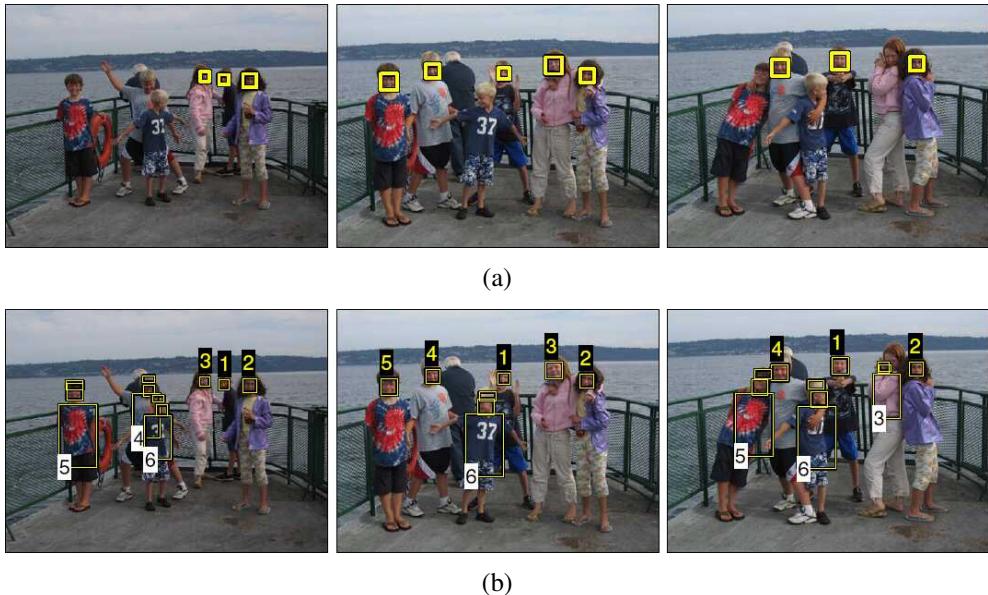


Figure 6.19 Person detection and re-recognition using a combined face, hair, and torso model (Sivic, Zitnick, and Szeliski 2006) © 2006 Springer. (a) Using face detection alone, several of the heads are missed. (b) The combined face and clothing model successfully re-finds all the people.

6.3.1 Face detection

Before face recognition can be applied to a general image, the locations and sizes of any faces must first be found (Figures 6.1c and 6.20). In principle, we could apply a face recognition algorithm at every pixel and scale (Moghaddam and Pentland 1997) but such a process would be too slow in practice.

Over the last four decades, a wide variety of fast face detection algorithms have been developed. Yang, Kriegman, and Ahuja (2002) and Zhao, Chellappa *et al.* (2003) provide comprehensive surveys of earlier work in this field. According to their taxonomy, face detection techniques can be classified as feature-based, template-based, or appearance-based. Feature-based techniques attempt to find the locations of distinctive image features such as the eyes, nose, and mouth, and then verify whether these features are in a plausible geometrical arrangement. These techniques include some of the early approaches to face recognition (Fischler and Elschlager 1973; Kanade 1977; Yuille 1991), as well as later approaches based on modular eigenspaces (Moghaddam and Pentland 1997), local filter jets (Leung, Burl, and



Figure 6.20 Face detection results produced by Rowley, Baluja, and Kanade (1998) © 1998 IEEE. Can you find the one false positive (a box around a non-face) among the 57 true positive results?

Perona 1995; Penev and Atick 1996; Wiskott, Fellous *et al.* 1997), support vector machines (Heisele, Ho *et al.* 2003; Heisele, Serre, and Poggio 2007), and boosting (Schneiderman and Kanade 2004).

Template-based approaches, such as active appearance models (AAMs) (Section 6.2.4), can deal with a wide range of pose and expression variability. Typically, they require good initialization near a real face and are therefore not suitable as fast face detectors.

Apearance-based approaches scan over small overlapping rectangular patches of the image searching for likely face candidates, which can then be refined using a *cascade* of more expensive but selective detection algorithms (Sung and Poggio 1998; Rowley, Baluja, and Kanade 1998; Romdhani, Torr *et al.* 2001; Fleuret and Geman 2001; Viola and Jones 2004). To deal with scale variation, the image is usually converted into a sub-octave pyramid and a separate scan is performed on each level. Most appearance-based approaches rely heavily on training classifiers using sets of labeled face and non-face patches.

Sung and Poggio (1998) and Rowley, Baluja, and Kanade (1998) present two of the earliest appearance-based face detectors and introduce a number of innovations that are widely used in later work by others. To start with, both systems collect a set of labeled face patches (Figure 6.20) as well as a set of patches taken from images that are known not to contain faces, such as aerial images or vegetation. The collected face images are augmented by artificially mirroring, rotating, scaling, and translating the images by small amounts to make the face detectors less sensitive to such effects.

The next few paragraphs provide quick reviews of a number of early appearance-based

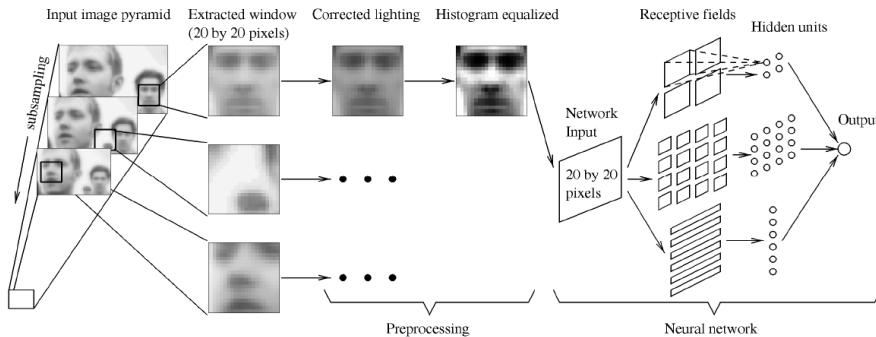


Figure 6.21 A neural network for face detection (Rowley, Baluja, and Kanade 1998) © 1998 IEEE. Overlapping patches are extracted from different levels of a pyramid and then pre-processed. A three-layer neural network is then used to detect likely face locations.

face detectors, keyed by the machine algorithms they are based on. These systems provide an interesting glimpse into the gradual adoption and evolution of machine learning in computer vision. More detailed descriptions can be found in the original papers, as well as the first edition of this book (Szeliski 2010).

Clustering and PCA. Once the face and non-face patterns have been pre-processed, Sung and Poggio (1998) cluster each of these datasets into six separate clusters using k-means and then fit PCA subspaces to each of the resulting 12 clusters. At detection time, the DIFS and DFFS metrics first developed by Moghaddam and Pentland (1997) are used to produce 24 Mahalanobis distance measurements (two per cluster). The resulting 24 measurements are input to a multi-layer perceptron (MLP), i.e., a fully connected neural network.

Neural networks. Instead of first clustering the data and computing Mahalanobis distances to the cluster centers, Rowley, Baluja, and Kanade (1998) apply a neural network (MLP) directly to the 20×20 pixel patches of gray-level intensities, using a variety of differently sized hand-crafted “receptive fields” to capture both large-scale and smaller scale structure (Figure 6.21). The resulting neural network directly outputs the likelihood of a face at the center of every overlapping patch in a multi-resolution pyramid. Since several overlapping patches (in both space and resolution) may fire near a face, an additional merging network is used to merge overlapping detections. The authors also experiment with training several networks and merging their outputs. Figure 6.20 shows a sample result from their face detector.

Support vector machines. Instead of using a neural network to classify patches, Osuna, Freund, and Girosi (1997) use *support vector machines* (SVMs), which we discussed in Section 5.1.4, to classify the same preprocessed patches as Sung and Poggio (1998). An SVM searches for a series of *maximum margin* separating planes in feature space between different classes (in this case, face and non-face patches). In those cases where linear classification boundaries are insufficient, the feature space can be lifted into higher-dimensional features using *kernels* (5.29). SVMs have been used by other researchers for both face detection and face recognition (Heisele, Ho *et al.* 2003; Heisele, Serre, and Poggio 2007) as well as general object recognition (Lampert 2008).

Boosting. Of all the face detectors developed in the 2000s, the one introduced by Viola and Jones (2004) is probably the best known. Their technique was the first to introduce the concept of *boosting* to the computer vision community, which involves training a series of increasingly discriminating simple classifiers and then blending their outputs (Bishop 2006, Section 14.3; Hastie, Tibshirani, and Friedman 2009, Chapter 10; Murphy 2012, Section 16.4; Glassner 2018, Section 14.7).

In more detail, boosting involves constructing a *classifier* $h(\mathbf{x})$ as a sum of simple *weak learners*,

$$h(\mathbf{x}) = \text{sign} \left[\sum_{j=0}^{m-1} \alpha_j h_j(\mathbf{x}) \right], \quad (6.4)$$

where each of the weak learners $h_j(\mathbf{x})$ is an extremely simple function of the input, and hence is not expected to contribute much (in isolation) to the classification performance.

In most variants of boosting, the weak learners are threshold functions,

$$h_j(\mathbf{x}) = a_j[f_j < \theta_j] + b_j[f_j \geq \theta_j] = \begin{cases} a_j & \text{if } f_j < \theta_j \\ b_j & \text{otherwise,} \end{cases} \quad (6.5)$$

which are also known as *decision stumps* (basically, the simplest possible version of *decision trees*). In most cases, it is also traditional (and simpler) to set a_j and b_j to ± 1 , i.e., $a_j = -s_j$, $b_j = +s_j$, so that only the feature f_j , the threshold value θ_j , and the polarity of the threshold $s_j \in \pm 1$ need to be selected.¹⁴

In many applications of boosting, the features are simply coordinate axes x_k , i.e., the boosting algorithm selects one of the input vector components as the best one to threshold. In Viola and Jones' face detector, the features are differences of rectangular regions in the input patch, as shown in Figure 6.22. The advantage of using these features is that, while they are

¹⁴Some variants, such as that of Viola and Jones (2004), use $(a_j, b_j) \in [0, 1]$ and adjust the learning algorithm accordingly.

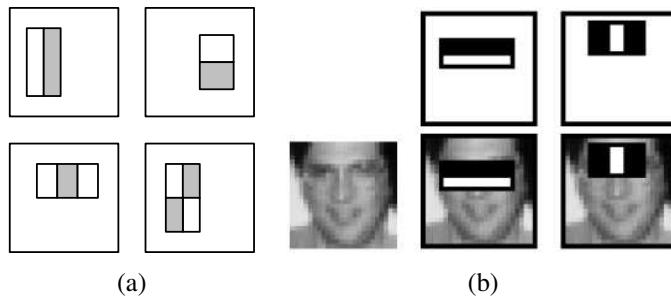


Figure 6.22 Simple features used in boosting-based face detector (Viola and Jones 2004) © 2004 Springer: (a) difference of rectangle feature composed of 2–4 different rectangles (pixels inside the white rectangles are subtracted from the gray ones); (b) the first and second features selected by AdaBoost. The first feature measures the differences in intensity between the eyes and the cheeks, the second one between the eyes and the bridge of the nose.

more discriminating than single pixels, they are extremely fast to compute once a summed area table has been precomputed, as described in Section 3.2.3 (3.31–3.32). Essentially, for the cost of an $O(N)$ precomputation phase (where N is the number of pixels in the image), subsequent differences of rectangles can be computed in $4r$ additions or subtractions, where $r \in \{2, 3, 4\}$ is the number of rectangles in the feature.

The key to the success of boosting is the method for incrementally selecting the weak learners and for re-weighting the training examples after each stage. The AdaBoost (Adaptive Boosting) algorithm (Bishop 2006; Hastie, Tibshirani, and Friedman 2009; Murphy 2012) does this by re-weighting each sample as a function of whether it is correctly classified at each stage, and using the stage-wise average classification error to determine the final weightings α_j among the weak classifiers.

To further increase the speed of the detector, it is possible to create a *cascade* of classifiers, where each classifier uses a small number of tests (say, a two-term AdaBoost classifier) to reject a large fraction of non-faces while trying to pass through all potential face candidates (Fleuret and Geman 2001; Viola and Jones 2004; Brubaker, Wu *et al.* 2008).

Deep networks. Since the initial burst of face detection research in the early 2000s, face detection algorithms have continued to evolve and improve (Zafeiriou, Zhang, and Zhang 2015). Researchers have proposed using cascades of features (Li and Zhang 2013), deformable parts models (Mathias, Benenson *et al.* 2014), aggregated channel features (Yang, Yan *et al.* 2014), and neural networks (Li, Lin *et al.* 2015; Yang, Luo *et al.* 2015). The WIDER FACE bench-

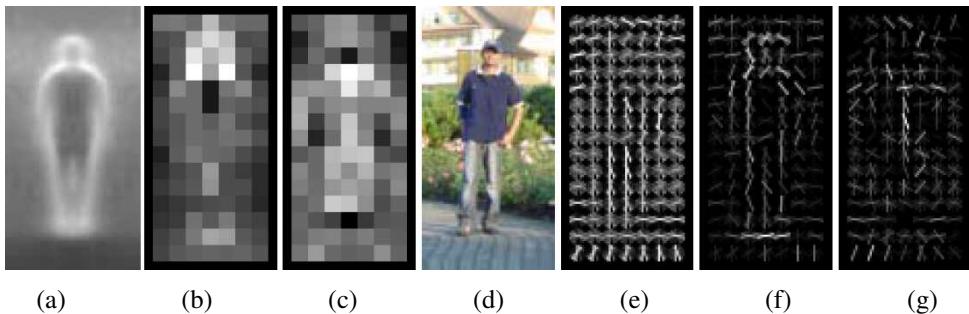


Figure 6.23 Pedestrian detection using histograms of oriented gradients (Dalal and Triggs 2005) © 2005 IEEE: (a) the average gradient image over the training examples; (b) each “pixel” shows the maximum positive SVM weight in the block centered on the pixel; (c) likewise, for the negative SVM weights; (d) a test image; (e) the computed R-HOG (rectangular histogram of gradients) descriptor; (f) the R-HOG descriptor weighted by the positive SVM weights; (g) the R-HOG descriptor weighted by the negative SVM weights.

mark^{15,16} (Yang, Luo *et al.* 2016) contains results from, and pointers to, more recent papers, including RetinaFace (Deng, Guo *et al.* 2020b), which combines ideas from other recent neural networks and object detectors such as Feature Pyramid Networks (Lin, Dollár *et al.* 2017) and RetinaNet (Lin, Goyal *et al.* 2017), and also has a nice review of other recent face detectors.

6.3.2 Pedestrian detection

While a lot of the early research on object detection focused on faces, the detection of other objects, such as pedestrians and cars, has also received widespread attention (Gavrila and Philomin 1999; Gavrila 1999; Papageorgiou and Poggio 2000; Mohan, Papageorgiou, and Poggio 2001; Schneiderman and Kanade 2004). Some of these techniques maintained the same focus as face detection on speed and efficiency. Others, however, focused on accuracy, viewing detection as a more challenging variant of generic class recognition (Section 6.3.3) in which the locations and extents of objects are to be determined as accurately as possible (Everingham, Van Gool *et al.* 2010; Everingham, Eslami *et al.* 2015; Lin, Maire *et al.* 2014).

An example of a well-known pedestrian detector is the algorithm developed by Dalal and Triggs (2005), who use a set of overlapping *histogram of oriented gradients* (HOG)

¹⁵<http://shuoyang1213.me/WIDERFACE>

¹⁶The WIDER FACE benchmark has expanded to a larger set of detection challenges and workshops: <https://wider-challenge.org/2019.html>.

descriptors fed into a support vector machine (Figure 6.23). Each HOG has cells to accumulate magnitude-weighted votes for gradients at particular orientations, just as in the scale invariant feature transform (SIFT) developed by Lowe (2004), which we will describe in Section 7.1.2 and Figure 7.16. Unlike SIFT, however, which is only evaluated at interest point locations, HOGs are evaluated on a regular overlapping grid and their descriptor magnitudes are normalized using an even coarser grid; they are only computed at a single scale and a fixed orientation. To capture the subtle variations in orientation around a person’s outline, a large number of orientation bins are used and no smoothing is performed in the central difference gradient computation—see Dalal and Triggs (2005) for more implementation details. Figure 6.23d shows a sample input image, while Figure 6.23e shows the associated HOG descriptors.

Once the descriptors have been computed, a support vector machine (SVM) is trained on the resulting high-dimensional continuous descriptor vectors. Figures 6.23b–c show a diagram of the (most) positive and negative SVM weights in each block, while Figures 6.23f–g show the corresponding weighted HOG responses for the central input image. As you can see, there are a fair number of positive responses around the head, torso, and feet of the person, and relatively few negative responses (mainly around the middle and the neck of the sweater).

Much like face detection, the fields of pedestrian and general object detection continued to advance rapidly in the 2000s (Belongie, Malik, and Puzicha 2002; Mikolajczyk, Schmid, and Zisserman 2004; Dalal and Triggs 2005; Leibe, Seemann, and Schiele 2005; Opelt, Pinz, and Zisserman 2006; Torralba 2007; Andriluka, Roth, and Schiele 2009; Maji and Berg 2009; Andriluka, Roth, and Schiele 2010; Dollár, Belongie, and Perona 2010).

A significant advance in the field of person detection was the work of Felzenszwalb, McAllester, and Ramanan (2008), who extend the histogram of oriented gradients person detector to incorporate flexible parts models (Section 6.2.1). Each part is trained and detected on HOGs evaluated at two pyramid levels below the overall object model and the locations of the parts relative to the parent node (the overall bounding box) are also learned and used during recognition (Figure 6.24b). To compensate for inaccuracies or inconsistencies in the training example bounding boxes (dashed white lines in Figure 6.24c), the “true” location of the parent (blue) bounding box is considered a latent (hidden) variable and is inferred during both training and recognition. Since the locations of the parts are also latent, the system can be trained in a semi-supervised fashion, without needing part labels in the training data. An extension to this system (Felzenszwalb, Girshick *et al.* 2010), which includes among its improvements a simple contextual model, was among the two best object detection systems in the 2008 Visual Object Classes detection challenge (Everingham, Van Gool *et al.* 2010).

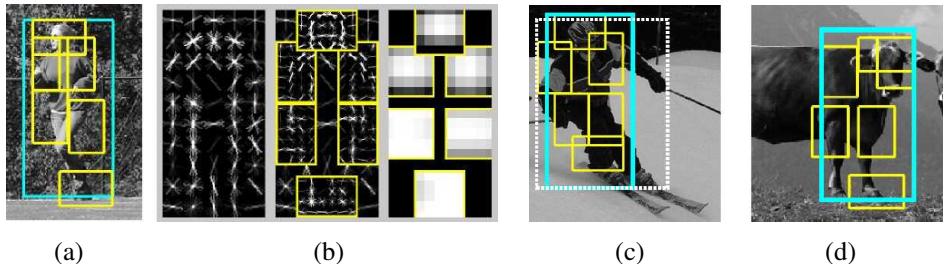


Figure 6.24 Part-based object detection (Felzenszwalb, McAllester, and Ramanan 2008) © 2008 IEEE: (a) An input photograph and its associated person (blue) and part (yellow) detection results. (b) The detection model is defined by a coarse template, several higher resolution part templates, and a spatial model for the location of each part. (c) True positive detection of a skier and (d) false positive detection of a cow (labeled as a person).

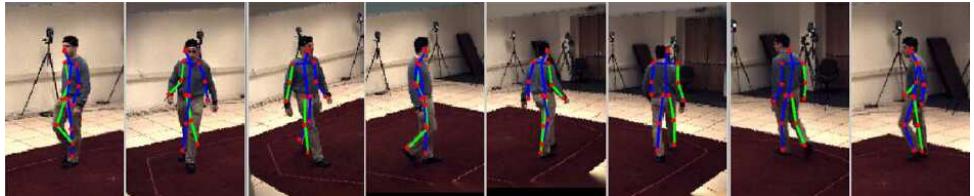


Figure 6.25 Pose detection using random forests (Rogez, Rihan et al. 2008) © 2008 IEEE. The estimated pose (state of the kinematic model) is drawn over each input frame.

Improvements to part-based person detection and pose estimation include work by Andriluka, Roth, and Schiele (2009) and Kumar, Zisserman, and Torr (2009).

An even more accurate estimate of a person's pose and location is presented by Rogez, Rihan *et al.* (2008), who compute both the phase of a person in a walk cycle and the locations of individual joints, using random forests built on top of HOGs (Figure 6.25). Since their system produces full 3D pose information, it is closer in its application domain to 3D person trackers (Sidenbladh, Black, and Fleet 2000; Andriluka, Roth, and Schiele 2010), which we will discuss in Section 13.6.4. When video sequences are available, the additional information present in the optical flow and motion discontinuities can greatly aid in the detection task, as discussed by Efros, Berg *et al.* (2003), Viola, Jones, and Snow (2003), and Dalal, Triggs, and Schmid (2006).

Since the 2000s, pedestrian and general person detection have continued to be actively developed, often in the context of more general multi-class object detection (Everingham, Van Gool *et al.* 2010; Everingham, Eslami *et al.* 2015; Lin, Maire *et al.* 2014). The Cal-

tech pedestrian detection benchmark¹⁷ and survey by Dollár, Belongie, and Perona (2010) introduces a new dataset and provides a nice review of algorithms through 2012, including Integral Channel Features (Dollár, Tu *et al.* 2009), the Fastest Pedestrian Detector in the West (Dollár, Belongie, and Perona 2010), and 3D pose estimation algorithms such as Poselets (Bourdev and Malik 2009). Since its original construction, this benchmark continues to tabulate and evaluate more recent detectors, including Dollár, Appel, and Kienzle (2012), Dollár, Appel *et al.* (2014), and more recent algorithms based on deep neural networks (Sermanet, Kavukcuoglu *et al.* 2013; Ouyang and Wang 2013; Tian, Luo *et al.* 2015; Zhang, Lin *et al.* 2016). The CityPersons dataset (Zhang, Benenson, and Schiele 2017) and WIDER Face and Person Challenge¹⁸ also report results on recent algorithms.

6.3.3 General object detection

While face and pedestrian detection algorithms were the earliest to be extensively studied, computer vision has always been interested in solving the general object detection and labeling problem, in addition to whole-image classification. The PASCAL Visual Object Classes (VOC) Challenge (Everingham, Van Gool *et al.* 2010), which contained 20 classes, had both classification and detection challenges. Early entries that did well on the detection challenge include a feature-based detector and spatial pyramid matching SVM classifier by Chum and Zisserman (2007), a star-topology deformable part model by Felzenszwalb, McAllester, and Ramanan (2008), and a sliding window SVM classifier by Lampert, Blaschko, and Hofmann (2008). The competition was re-run annually, with the two top entries in the 2012 detection challenge (Everingham, Eslami *et al.* 2015) using a sliding window spatial pyramid matching (SPM) SVM (de Sande, Uijlings *et al.* 2011) and a University of Oxford re-implementation of a deformable parts model (Felzenszwalb, Girshick *et al.* 2010).

The ImageNet Large Scale Visual Recognition Challenge (ILSVRC), released in 2010, scaled up the dataset from around 20 thousand images in PASCAL VOC 2010 to over 1.4 million in ILSVRC 2010, and from 20 object classes to 1,000 object classes (Russakovsky, Deng *et al.* 2015). Like PASCAL, it also had an object detection task, but it contained a much wider range of challenging images (Figure 6.4). The Microsoft COCO (Common Objects in Context) dataset (Lin, Maire *et al.* 2014) contained even more objects per image, as well as pixel-accurate segmentations of multiple objects, enabling the study of not only *semantic segmentation* (Section 6.4), but also individual object *instance segmentation* (Section 6.4.2). Table 6.2 list some of the datasets used for training and testing general object detection algorithms.

¹⁷http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians

¹⁸<https://wider-challenge.org/2019.html>

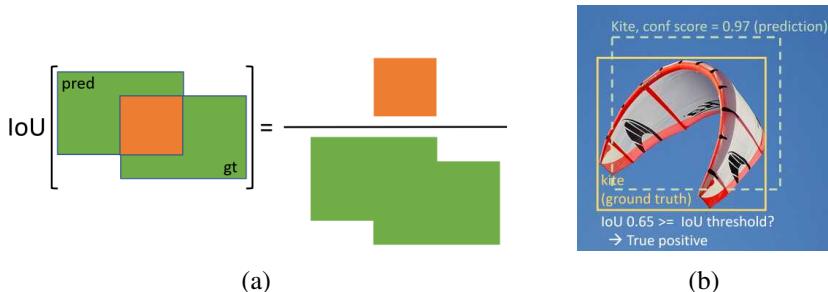


Figure 6.26 Intersection over union (IoU): (a) schematic formula, (b) real-world example
 © 2020 Ross Girshick.

The release of COCO coincided with a wholesale shift to deep networks for image classification, object detection, and segmentation (Jiao, Zhang *et al.* 2019; Zhao, Zheng *et al.* 2019). Figure 6.29 shows the rapid improvements in average precision (AP) on the COCO object detection task, which correlates strongly with advances in deep neural network architectures (Figure 5.40).

Precision vs. recall

Before we describe the elements of modern object detectors, we should first discuss what metrics they are trying to optimize. The main task in object detection, as illustrated in Figures 6.5a and 6.26b, is to put accurate bounding boxes around all the objects of interest and to correctly label such objects. To measure the accuracy of each bounding box (not too small and not too big), the common metric is *intersection over union* (IoU), which is also known as the *Jaccard index* or *Jaccard similarity coefficient* (Rezatofighi, Tsoi *et al.* 2019). The IoU is computed by taking the predicted and ground truth bounding boxes B_{pr} and B_{gt} for an object and computing the ratio of their area of intersection and their area of union,

$$IoU = \frac{B_{\text{pr}} \cap B_{\text{gt}}}{B_{\text{pr}} \cup B_{\text{gt}}}, \quad (6.6)$$

as shown in Figure 6.26a.

As we will shortly see, object detectors operate by first proposing a number of plausible rectangular regions (detections) and then classifying each detection while also producing a confidence score (Figure 6.26b). These regions are then run through some kind of *non-maximal suppression* (NMS) stage, which removes weaker detections that have too much overlap with stronger detections, using a greedy most-confident-first algorithm.

To evaluate the performance of an object detector, we run through all of the detections,

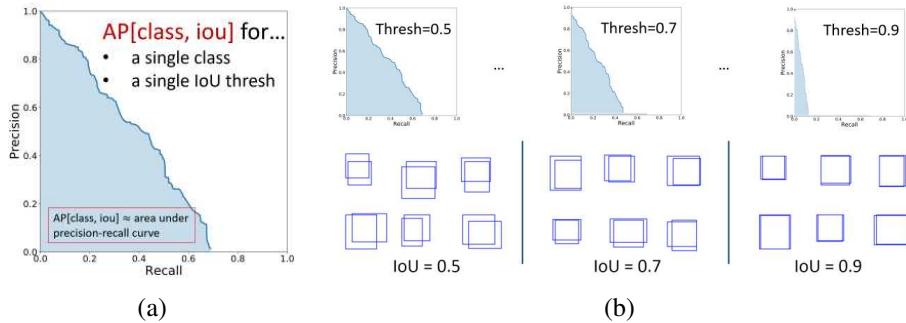


Figure 6.27 Object detector average precision © 2020 Ross Girshick: (a) a precision-recall curve for a single class and IoU threshold, with the AP being the area under the P-R curve; (b) average precision averaged over several IoU thresholds (from looser to tighter).

from most confident to least, and classify them as *true positive* TP (correct label and sufficiently high IoU) or *false positive* FP (incorrect label or ground truth object already matched). For each new decreasing confidence threshold, we can compute the *precision* and *recall* as

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (6.7)$$

$$\text{recall} = \frac{\text{TP}}{\text{P}}, \quad (6.8)$$

where P is the number of positive examples, i.e., the number of labeled ground truth detections in the test image.¹⁹ (See Section 7.1.3 on feature matching for additional terms that are often used in measuring and describing error rates.)

Computing the precision and recall at every confidence threshold allows us to populate a precision-recall curve, such as the one in Figure 6.27a. The area under this curve is called *average precision* (AP). A separate AP score can be computed for each class being detected, and the results averaged to produce a *mean average precision* (mAP). Another widely used measure if the While earlier benchmarks such as PASCAL VOC determined the mAP using a single IoU threshold of 0.5 (Everingham, Eslami *et al.* 2015), the COCO benchmark (Lin, Maire *et al.* 2014) averages the mAP over a set of IoU thresholds, $\text{IoU} \in \{0.50, 0.55, \dots, 0.95\}$, as shown in Figure 6.27a. While this AP score continues to be widely used, an alternative *probability-based detection quality* (PDQ) score has recently been proposed (Hall, Dayoub *et al.* 2020). A smoother version of average precision called Smooth-AP has also been proposed and shown to have benefits on large-scale image retrieval tasks (Brown, Xie *et al.* 2020).

¹⁹Another widely reported measure is the *F-score*, which is the harmonic mean of the precision and recall.

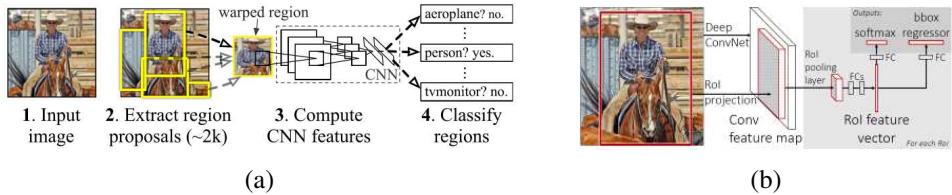


Figure 6.28 The R-CNN and Fast R-CNN object detectors. (a) R-CNN rescales pixels inside each proposal region and performs a CNN + SVM classification (Girshick, Donahue et al. 2015) © 2015 IEEE. (b) Fast R-CNN resamples convolutional features and uses fully connected layers to perform classification and bounding box regression (Girshick 2015) © 2015 IEEE.

Modern object detectors

The first stage in detecting objects in an image is to propose a set of plausible rectangular regions in which to run a classifier. The development of such *region proposal* algorithms was an active research area in the early 2000s (Alexe, Deselaers, and Ferrari 2012; Uijlings, Van De Sande et al. 2013; Cheng, Zhang et al. 2014; Zitnick and Dollár 2014).

One of the earliest object detectors based on neural networks is R-CNN, the Region-based Convolutional Network developed by Girshick, Donahue et al. (2014). As illustrated in Figure 6.28a, this detector starts by extracting about 2,000 region proposals using the selective search algorithm of Uijlings, Van De Sande et al. (2013). Each proposed regions is then rescaled (warped) to a 224 square image and passed through an AlexNet or VGG neural network with a support vector machine (SVM) final classifier.

The follow-on Fast R-CNN paper by Girshick (2015) interchanges the convolutional neural network and region extraction stages and replaces the SVM with some fully connected (FC) layers, which compute both an object class and a bounding box refinement (Figure 6.28b). This reuses the CNN computations and leads to much faster training and test times, as well as dramatically better accuracy compared to previous networks (Figure 6.29). As you can see from Figure 6.28b, Fast R-CNN is an example of a deep network with a shared *backbone* and two separate *heads*, and hence two different loss functions, although these terms were not introduced until the Mask R-CNN paper by He, Gkioxari et al. (2017).

The Faster R-CNN system, introduced a few month later by Ren, He et al. (2015), replaces the relatively slow selective search stage with a convolutional region proposal network (RPN), resulting in much faster inference. After computing convolutional features, the RPN suggests at each coarse location a number of potential *anchor boxes*, which vary in shape and size

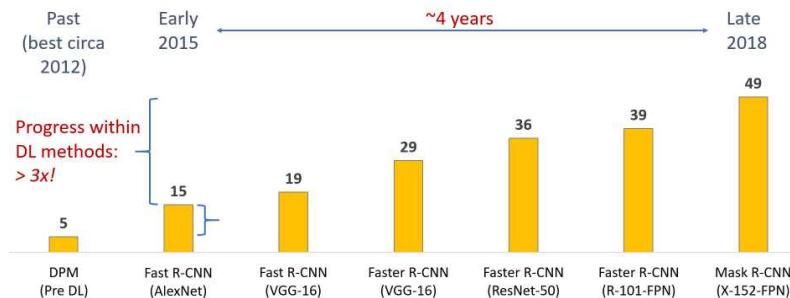


Figure 6.29 Best average precision (AP) results by year on the COCO object detection task (Lin, Maire et al. 2014) © 2020 Ross Girshick.

to accommodate different potential objects. Each proposal is then classified and refined by an instance of the Fast R-CNN heads and the final detections are ranked and merged using non-maximal suppression.

R-CNN, Fast R-CNN, and Faster R-CNN all operate on a single resolution convolutional feature map (Figure 6.30b). To obtain better scale invariance, it would be preferable to operate on a range of resolutions, e.g., by computing a feature map at each image pyramid level, as shown in Figure 6.30a, but this is computationally expensive. We could, instead, simply start with the various levels inside the convolutional network (Figure 6.30c), but these levels have different degrees of semantic abstraction, i.e., higher/smaller levels are attuned to more abstract constructs. The best solution is to construct a *Feature Pyramid Network* (FPN), as shown in Figure 6.30d, where top-down connections are used to endow higher-resolution (lower) pyramid levels with the semantics inferred at higher levels (Lin, Dollár et al. 2017).²⁰ This additional information significantly enhances the performance of object detectors (and other downstream tasks) and makes their behavior much less sensitive to object size.

DETR (Carion, Massa et al. 2020) uses a simpler architecture that eliminates the use of non-maximum suppression and anchor generation. Their model consists of a ResNet backbone that feeds into a transformer encoder-decoder. At a high level, it makes N bounding box predictions, some of which may include the “no object class”. The ground truth bounding boxes are also padded with “no object class” bounding boxes to obtain N total bounding boxes. During training, *bipartite matching* is then used to build a one-to-one mapping from every predicted bounding box to a ground truth bounding box, with the chosen mapping lead-

²⁰It's interesting to note that the human visual system is full of such re-entrant or feedback pathways (Gilbert and Li 2013), although the extent to which cognition influences perception is still being debated (Firestone and Scholl 2016).

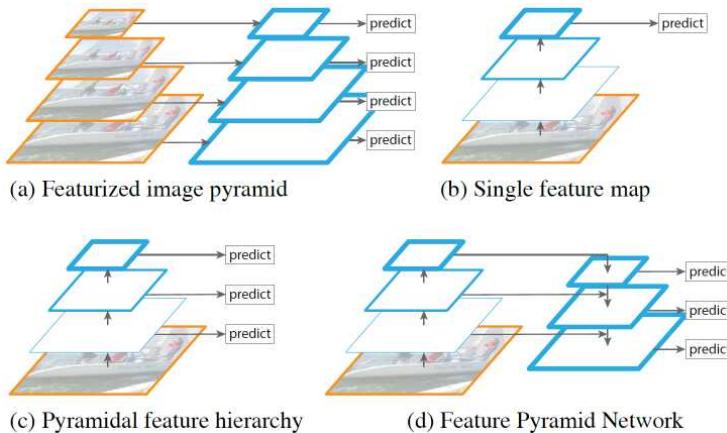


Figure 6.30 A Feature Pyramid Network and its precursors (Lin, Dollár et al. 2017) © 2017 IEEE: (a) deep features extracted at each level in an image pyramid; (b) a single low-resolution feature map; (c) a deep feature pyramid, with higher levels having greater abstraction; (d) a Feature Pyramid Network, with top-down context for all levels.

ing to the lowest possible cost. The overall training loss is then the sum of the losses between the matched bounding boxes. They find that their approach is competitive with state-of-the-art object detection performance on COCO.

Single-stage networks

In the architectures we've looked at so far, a region proposal algorithm or network selects the locations and shapes of the detections to be considered, and a second network is then used to classify and regress the pixels or features inside each region. An alternative is to use a *single-stage network*, which uses a single neural network to output detections at a variety of locations. Two examples of such detectors are SSD (Single Shot MultiBox Detector) from Liu, Anguelov et al. (2016) and the family of YOLO (You Only Look Once) detectors described in Redmon, Divvala et al. (2016), Redmon and Farhadi (2017), and Redmon and Farhadi (2018). RetinaNet (Lin, Goyal et al. 2017) is also a single-stage detector built on top of a feature pyramid network. It uses a *focal loss* to focus the training on hard examples by downweighting the loss on well-classified samples, thus preventing the larger number of easy negatives from overwhelming the training. These and more recent convolutional object detectors are described in the recent survey by Jiao, Zhang et al. (2019). Figure 6.31 shows the speed and accuracy of detectors published up through early 2017.

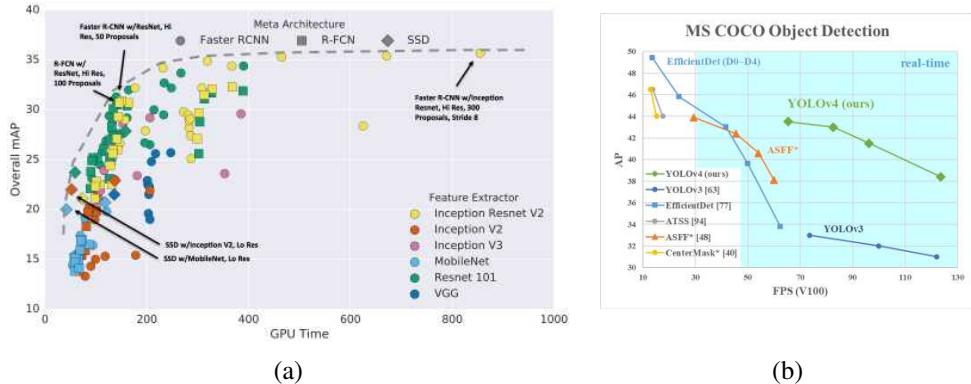


Figure 6.31 Speed/accuracy trade-offs for convolutional object detectors: (a) (Huang, Rathod et al. 2017) © 2017 IEEE; (b) YOLOv4 © Bochkovskiy, Wang, and Liao (2020).

The latest in the family of YOLO detectors is YOLOv4 by Bochkovskiy, Wang, and Liao (2020). In addition to outperforming other recent fast detectors such as EfficientDet (Tan, Pang, and Le 2020), as shown in Figure 6.31b, the paper breaks the processing pipeline into several stages, including a *neck*, which performs the top-down feature enhancement found in the feature pyramid network. The paper also evaluates many different components, which they categorize into a “bag of freebies” that can be used during training and a “bag of specials” that can be used at detection time with minimal additional cost.

While most bounding box object detectors continue to evaluate their results on the COCO dataset (Lin, Maire et al. 2014),²¹ newer datasets such as Open Images (Kuznetsova, Rom et al. 2020), and LVIS: Large Vocabulary Instance Segmentation (Gupta, Dollár, and Girshick 2019) are now also being used (see Table 6.2). Two recent workshops that highlight the latest results using these datasets are Zendel et al. (2020) and Kirillov, Lin et al. (2020) and also have challenges related to instance segmentation, panoptic segmentation, keypoint estimation, and dense pose estimation, which are topics we discuss later in this chapter. Open-source frameworks for training and fine-tuning object detectors include the TensorFlow Object Detection API²² and PyTorch’s Detectron2.²³

| Name/URL | Extents | Contents/Reference |
|--|--|---|
| <i>Object recognition</i> | | |
| Oxford buildings dataset | Pictures of buildings https://www.robots.ox.ac.uk/~vgg/data/oxbuildings | 5,062 images Philbin, Chum <i>et al.</i> (2007) |
| INRIA Holidays | Holiday scenes https://lear.inrialpes.fr/people/jegou/data.php | 1,491 images Jégou, Douze, and Schmid (2008) |
| PASCAL | Segmentations, boxes http://host.robots.ox.ac.uk/pascal/VOC | 11k images (2.9k with segmentations) Everingham, Eslami <i>et al.</i> (2015) |
| ImageNet | Complete images https://www.image-net.org | 21k (WordNet) classes, 14M images Deng, Dong <i>et al.</i> (2009) |
| Fashion MNIST | Complete images https://github.com/zalandoresearch/fashion-mnist | 70k fashion products Xiao, Rasul, and Vollgraf (2017) |
| <i>Object detection and segmentation</i> | | |
| Caltech Pedestrian Dataset | Bounding boxes http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians | Pedestrians Dollár, Wojek <i>et al.</i> (2009) |
| MSR Cambridge | Per-pixel segmentations https://www.microsoft.com/en-us/research/project/image-understanding | 23 classes Shotton, Winn <i>et al.</i> (2009) |
| LabelMe dataset | Polygonal boundaries http://labelme.csail.mit.edu | >500 categories Russell, Torralba <i>et al.</i> (2008) |
| Microsoft COCO | Segmentations, boxes https://cocodataset.org | 330k images Lin, Maire <i>et al.</i> (2014) |
| Cityscapes | Polygonal boundaries https://www.cityscapes-dataset.com | 30 classes, 25,000 images Cordts, Omran <i>et al.</i> (2016) |
| Broden | Segmentation masks http://netdissect.csail.mit.edu | A variety of visual concepts Bau, Zhou <i>et al.</i> (2017) |
| Broden+ | Segmentation masks https://github.com/CSAILVision/unifiedparsing | A variety of visual concepts Xiao, Liu <i>et al.</i> (2018) |
| LVIS | Instance segmentations https://www.lvisdataset.org | 1,000 categories, 2.2M images Gupta, Dollár, and Girshick (2019) |
| Open Images | Segs., relationships https://g.co/dataset/openimages | 478k images, 3M relationships Kuznetsova, Rom <i>et al.</i> (2020) |

Table 6.2 Image databases for classification, detection, and localization.



Figure 6.32 Examples of image segmentation (Kirillov, He et al. 2019) © 2019 IEEE: (a) original image; (b) semantic segmentation (per-pixel classification); (c) instance segmentation (delineate each object); (d) panoptic segmentation (label all things and stuff).

6.4 Semantic segmentation

A challenging version of general object recognition and scene understanding is to simultaneously perform recognition and accurate boundary segmentation (Fergus 2007). In this section, we examine a number of related problems, namely *semantic segmentation* (per-pixel class labeling), *instance segmentation* (accurately delineating each separate object), *panoptic segmentation* (labeling both objects and stuff), and dense pose estimation (labeling pixels belonging to people and their body parts). Figures 6.32 and 6.43 show some of these kinds of segmentations.

The basic approach to simultaneous recognition and segmentation is to formulate the problem as one of labeling every pixel in an image with its class membership. Older approaches often did this using energy minimization or Bayesian inference techniques, i.e., conditional random fields (Section 4.3.1). The TextronBoost system of Shotton, Winn et al. (2009) uses unary (pixel-wise) potentials based on image-specific color distributions (Section 4.3.2), location information (e.g., foreground objects are more likely to be in the middle of the image, sky is likely to be higher, and road is likely to be lower), and novel texture-layout classifiers trained using shared boosting. It also uses traditional pairwise potentials that look at image color gradients. The texton-layout features first filter the image with a series of 17 oriented filter banks and then cluster the responses to classify each pixel into 30 different texton classes (Malik, Belongie et al. 2001). The responses are then filtered using offset rectangular regions trained with joint boosting (Viola and Jones 2004) to produce the texton-layout features used as unary potentials. Figure 6.33 shows some examples of images successfully labeled and segmented using TextronBoost

The TextronBoost conditional random field framework has been extended to LayoutCRFs

²¹See <https://codalab.org> for the latest competitions and leaderboards.

²²https://github.com/tensorflow/models/tree/master/research/object_detection

²³<https://github.com/facebookresearch/detectron2>

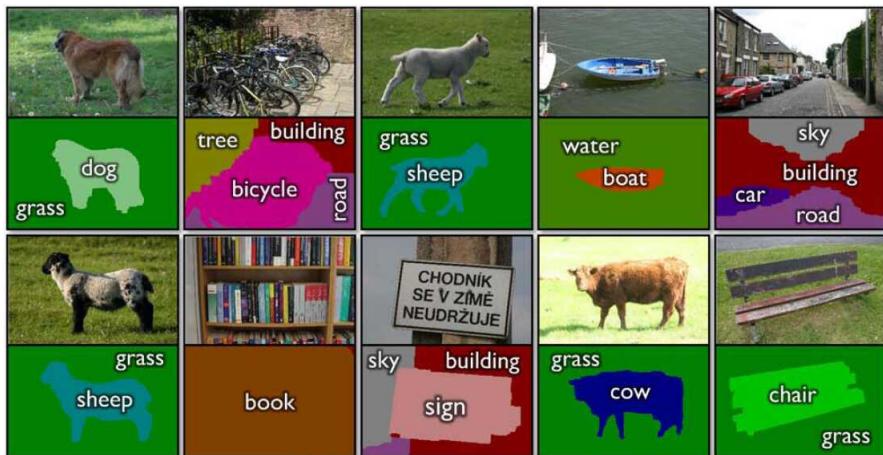


Figure 6.33 Simultaneous recognition and segmentation using TextonBoost (Shotton, Winn et al. 2009) © 2009 Springer.

by Winn and Shotton (2006), who incorporate additional constraints to recognize multiple object instances and deal with occlusions, and by Hoiem, Rother, and Winn (2007) to incorporate full 3D models. Conditional random fields continued to be widely used and extended for simultaneous recognition and segmentation applications, as described in the first edition of this book (Szeliski 2010, Section 14.4.3), along with approaches that first performed low-level or hierarchical segmentations (Section 7.5).

The development of fully convolutional networks (Long, Shelhamer, and Darrell 2015), which we described in Section 5.4.1, enabled per-pixel semantic labeling using a single neural network. While the first networks suffered from poor resolution (very loose boundaries), the addition of conditional random fields at a final stage (Chen, Papandreou *et al.* 2018; Zheng, Jayasumana *et al.* 2015), deconvolutional upsampling (Noh, Hong, and Han 2015), and fine-level connections in U-nets (Ronneberger, Fischer, and Brox 2015), all helped improve accuracy and resolution.

Modern semantic segmentation systems are often built on architectures such as the feature pyramid network (Lin, Dollár *et al.* 2017), which have top-down connections to help percolate semantic information down to higher-resolution maps. For example, the Pyramid Scene Parsing Network (PSPNet) of Zhao, Shi *et al.* (2017) uses spatial pyramid pooling (He, Zhang *et al.* 2015) to aggregate features at various resolution levels. The Unified Perceptual Parsing network (UPerNet) of Xiao, Liu *et al.* (2018) uses both a feature pyramid network and a pyramid pooling module to label image pixels not only with object categories but also

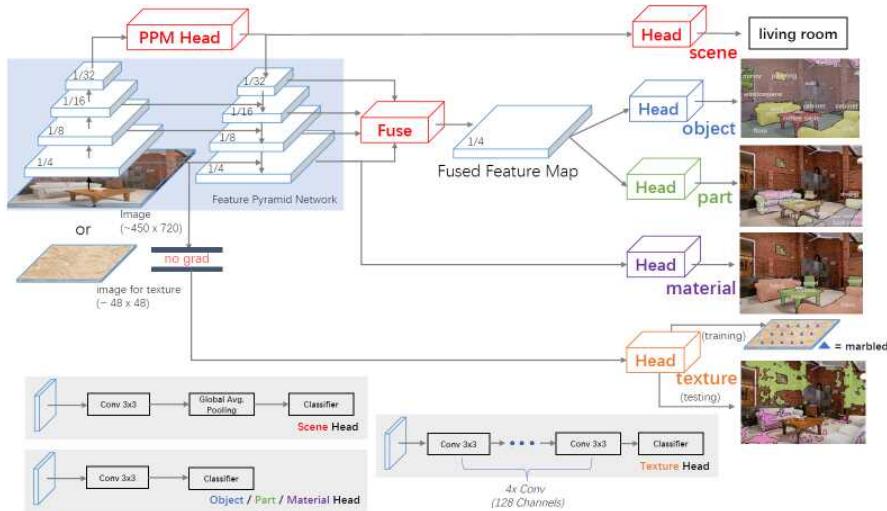


Figure 6.34 The UPerNet framework for Unified Perceptual Parsing (Xiao, Liu et al. 2018) © 2018 Springer. A Feature Pyramid Network (FPN) backbone is appended with a Pyramid Pooling Module (PPM) before feeding it into the top-down branch of the FPN. Various layers of the FPN and/or PPM are fed into different heads, including a scene head for image classification, object and part heads from the fused FPN features, a material head operating on the finest level of the FPN, and a texture head that does not participate in the FPN fine tuning. The bottom gray squares give more details into some of the heads.

materials, parts, and textures, as shown in Figure 6.34. HRNet (Wang, Sun et al. 2020) keeps high-resolution versions of feature maps throughout the pipeline with occasional interchange of information between channels at different resolution layers. Such networks can also be used to estimate surface normals and depths in an image (Huang, Zhou et al. 2019; Wang, Geraghty et al. 2020).

Semantic segmentation algorithms were initially trained and tested on datasets such as MSRC (Shotton, Winn et al. 2009) and PASCAL VOC (Everingham, Eslami et al. 2015). More recent datasets include the Cityscapes dataset for urban scene understanding (Cordts, Omran et al. 2016) and ADE20K (Zhou, Zhao et al. 2019), which labels pixels in a wider variety of indoor and outdoor scenes with 150 different category and part labels. The Broadly and Densely Labeled Dataset (Broden) created by Bau, Zhou et al. (2017) federates a number of such densely labeled datasets, including ADE20K, Pascal-Context, Pascal-Part, OpenSurfaces, and Describable Textures to obtain a wide range of labels such as materials and textures

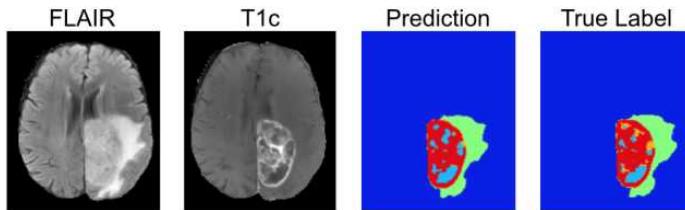


Figure 6.35 3D volumetric medical image segmentation using a deep network (Kamnitsas, Ferrante et al. 2016) © 2016 Springer.

in addition to basic object semantics. While this dataset was originally developed to aid in the interpretability of deep networks, it has also proven useful (with extensions) for training unified multi-task labeling systems such as UPerNet (Xiao, Liu et al. 2018). Table 6.2 lists some of the datasets used for training and testing semantic segmentation algorithms.

One final note. While semantic image segmentation and labeling have widespread applications in image understanding, the converse problem of going from a semantic sketch or painting of a scene to a photorealistic image has also received widespread attention (Johnson, Gupta, and Fei-Fei 2018; Park, Liu et al. 2019; Bau, Strobelt et al. 2019; Ntavelis, Romero et al. 2020b). We look at this topic in more detail in Section 10.5.3 on semantic image synthesis.

6.4.1 Application: Medical image segmentation

One of the most promising applications of image segmentation is in the medical imaging domain, where it can be used to segment anatomical tissues for later quantitative analysis. Figure 4.21 shows a binary graph cut with directed edges being used to segment the liver tissue (light gray) from its surrounding bone (white) and muscle (dark gray) tissue. Figure 6.35 shows the segmentation of a brain scan for the detection of brain tumors. Before the development of the mature optimization and deep learning techniques used in modern image segmentation algorithms, such processing required much more laborious manual tracing of individual X-ray slices.

Initially, optimization techniques such as Markov random fields (Section 4.3.2) and discriminative classifiers such as random forests (Section 5.1.5) were used for medical image segmentation (Criminisi, Robertson et al. 2013). More recently, the field has shifted to deep learning approaches (Kamnitsas, Ferrante et al. 2016; Kamnitsas, Ledig et al. 2017; Havaei, Davy et al. 2017).

The fields of medical image segmentation (McInerney and Terzopoulos 1996) and medical image registration (Kybic and Unser 2003) (Section 9.2.3) are rich research fields with

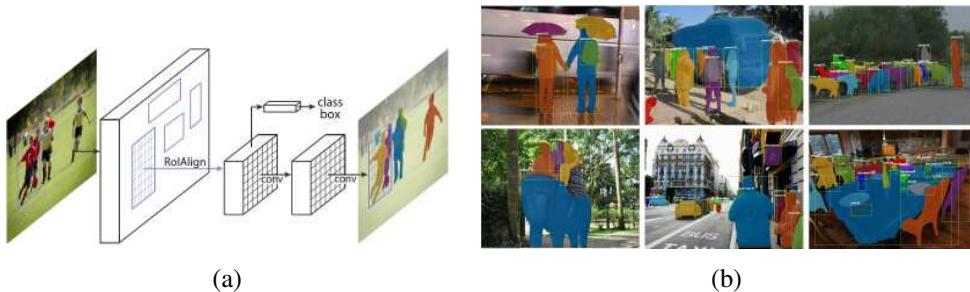


Figure 6.36 *Instance segmentation using Mask R-CNN (He, Gkioxari et al. 2017)* © 2017 IEEE: (a) system architecture, with an additional segmentation branch; (b) sample results.

their own specialized conferences, such as *Medical Imaging Computing and Computer Assisted Intervention (MICCAI)*, and journals, such as *Medical Image Analysis* and *IEEE Transactions on Medical Imaging*. These can be great sources of references and ideas for research in this area.

6.4.2 Instance segmentation

Instance segmentation is the task of finding all of the relevant objects in an image and producing pixel-accurate masks for their visible regions (Figure 6.36b). One potential approach to this task is to perform known object instance recognition (Section 6.1) and to then back-project the object model into the scene (Lowe 2004), as shown in Figure 6.1d, or matching portions of the new scene to pre-learned (segmented) object models (Ferrari, Tuytelaars, and Van Gool 2006b; Kannala, Rahtu *et al.* 2008). However, this approach only works for known rigid 3D models.

For more complex (flexible) object models, such as those for humans, a different approach is to pre-segment the image into larger or smaller pieces (Section 7.5) and to then match such pieces to portions of the model (Mori, Ren *et al.* 2004; Mori 2005; He, Zemel, and Ray 2006; Gu, Lim *et al.* 2009). For general highly variable classes, a related approach is to vote for potential object locations and scales based on feature correspondences and to then infer the object extents (Leibe, Leonardis, and Schiele 2008).

With the advent of deep learning, researchers started combining region proposals or image pre-segmentations with convolutional second stages to infer the final instance segmentations (Hariharan, Arbeláez *et al.* 2014; Hariharan, Arbeláez *et al.* 2015; Dai, He, and Sun 2015; Pinheiro, Lin *et al.* 2016; Dai, He, and Sun 2016; Li, Qi *et al.* 2017).

A breakthrough in instance segmentation came with the introduction of Mask R-CNN



Figure 6.37 Person keypoint detection and segmentation using Mask R-CNN (He, Gkioxari et al. 2017) © 2017 IEEE

(He, Gkioxari *et al.* 2017). As shown in Figure 6.36a, Mask R-CNN uses the same region proposal network as Faster R-CNN (Ren, He *et al.* 2015), but then adds an additional *branch* for predicting the object *mask*, in addition to the existing branch for bounding box refinement and classification.²⁴ As with other networks that have multiple branches (or heads) and outputs, the training losses corresponding to each supervised output need to be carefully balanced. It is also possible to add additional branches, e.g., branches trained to detect human keypoint locations (implemented as per-keypoint mask images), as shown in Figure 6.37.

Since its introduction, the performance of Mask R-CNN and its extensions has continued to improve with advances in backbone architectures (Liu, Qi *et al.* 2018; Chen, Pang *et al.* 2019). Two recent workshops that highlight the latest results in this area are the COCO + LVIS Joint Recognition Challenge (Kirillov, Lin *et al.* 2020) and the Robust Vision Challenge (Zendel *et al.* 2020).²⁵ It is also possible to replace the pixel masks produced by most instance segmentation techniques with time-evolving closed contours, i.e., “snakes” (Section 7.3.1), as in Peng, Jiang *et al.* (2020). In order to encourage higher-quality segmentation boundaries, Cheng, Girshick *et al.* (2021) propose a new Boundary Intersection-over-Union (Boundary IoU) metric to replace the commonly used Mask IoU metric.

6.4.3 Panoptic segmentation

As we have seen, semantic segmentation classifies each pixel in an image into its semantic category, i.e., what *stuff* does each pixel correspond to. Instance segmentation associates pixels with individual objects, i.e., how many *objects* are there and what are their extents

²⁴Mask R-CNN was the first paper to introduce the terms *backbone* and *head* to describe the common deep convolutional feature extraction front end and the specialized back end branches.

²⁵You can find the leaderboards for instance segmentation and other COCO recognition tasks at <https://cocodataset.org>.



Figure 6.38 Panoptic segmentation results produced using a Panoptic Feature Pyramid Network (Kirillov, Girshick et al. 2019) © 2019 IEEE.

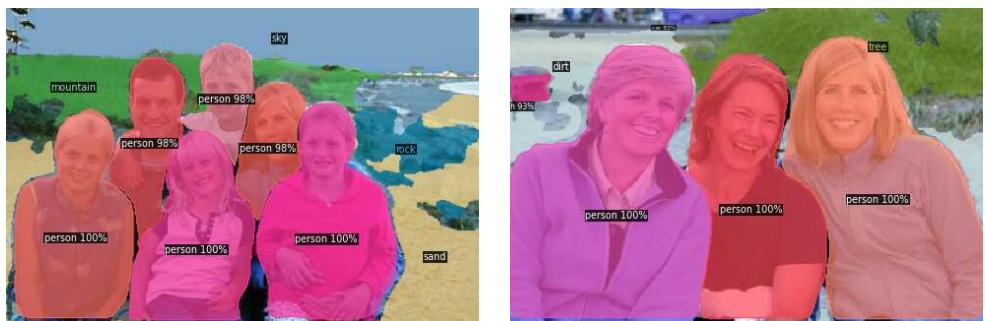


Figure 6.39 Detectron2 panoptic segmentation results on some of my personal photos. (Click on the “Colab Notebook” link at <https://github.com/facebookresearch/detectron2> and then edit the input image URL to try your own.)

(Figure 6.32). Putting both of these systems together has long been a goal of semantic scene understanding (Yao, Fidler, and Urtasun 2012; Tighe and Lazebnik 2013; Tu, Chen *et al.* 2005). Doing this on a per-pixel level results in a *panoptic segmentation* of the scene, where all of the objects are correctly segmented and the remaining stuff is correctly labeled (Kirillov, He *et al.* 2019). Producing a sensible *panoptic quality* (PQ) metric that simultaneously balances the accuracy on both tasks takes some careful design. In their paper, Kirillov, He *et al.* (2019) describe their proposed metric and analyze the performance of both humans (in terms of consistency) and recent algorithms on three different datasets.

The COCO dataset has now been extended to include a panoptic segmentation task, on which some recent results can be found in the ECCV 2020 workshop on this topic (Kirillov, Lin *et al.* 2020). Figure 6.38 show some segmentations produced by the panoptic feature pyramid network described by Kirillov, Girshick *et al.* (2019), which adds two branches for instance segmentation and semantic segmentation to a feature pyramid network.

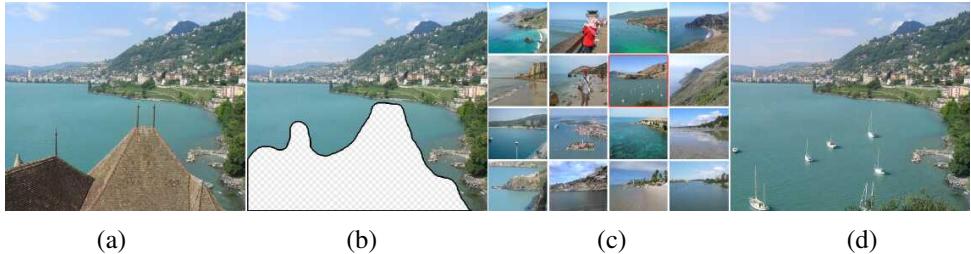


Figure 6.40 *Scene completion using millions of photographs (Hays and Efros 2007)* © 2007 ACM: (a) original image; (b) after unwanted foreground removal; (c) plausible scene matches, with the one the user selected highlighted in red; (d) output image after replacement and blending.

6.4.4 Application: Intelligent photo editing

Advances in object recognition and scene understanding have greatly increased the power of intelligent (semi-automated) photo editing applications. One example is the Photo Clip Art system of Lalonde, Hoiem *et al.* (2007), which recognizes and segments objects of interest, such as pedestrians, in internet photo collections and then allows users to paste them into their own photos. Another is the scene completion system of Hays and Efros (2007), which tackles the same *inpainting* problem we will study in Section 10.5. Given an image in which we wish to erase and fill in a large section (Figure 6.40a–b), where do you get the pixels to fill in the gaps in the edited image? Traditional approaches either use smooth continuation (Bertalmio, Sapiro *et al.* 2000) or borrow pixels from other parts of the image (Efros and Leung 1999; Criminisi, Pérez, and Toyama 2004; Efros and Freeman 2001). With the availability of huge numbers of images on the web, it often makes more sense to find a *different* image to serve as the source of the missing pixels.

In their system, Hays and Efros (2007) compute the *gist* of each image (Oliva and Torralba 2001; Torralba, Murphy *et al.* 2003) to find images with similar colors and composition. They then run a graph cut algorithm that minimizes image gradient differences and composite the new replacement piece into the original image using Poisson image blending (Section 8.4.4) (Pérez, Gangnet, and Blake 2003). Figure 6.40d shows the resulting image with the erased foreground rooftops region replaced with sailboats. Additional examples of photo editing and computational photography applications enabled by what has been dubbed “internet computer vision” can be found in the special journal issue edited by Avidan, Baker, and Shan (2010).

A different application of image recognition and segmentation is to infer 3D structure from a single photo by recognizing certain scene structures. For example, Criminisi, Reid,

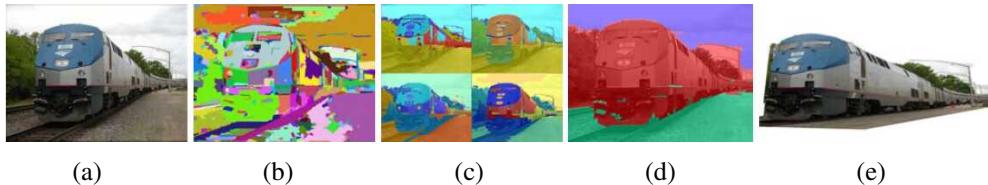


Figure 6.41 Automatic photo pop-up (Hoiem, Efros, and Hebert 2005a) © 2005 ACM: (a) input image; (b) superpixels are grouped into (c) multiple regions; (d) labels indicating ground (green), vertical (red), and sky (blue); (e) novel view of resulting piecewise-planar 3D model.

and Zisserman (2000) detect vanishing points and have the user draw basic structures, such as walls, to infer the 3D geometry (Section 11.1.2). Hoiem, Efros, and Hebert (2005a), on the other hand, work with more “organic” scenes such as the one shown in Figure 6.41. Their system uses a variety of classifiers and statistics learned from labeled images to classify each pixel as either ground, vertical, or sky (Figure 6.41d). To do this, they begin by computing superpixels (Figure 6.41b) and then group them into plausible regions that are likely to share similar geometric labels (Figure 6.41c). After all the pixels have been labeled, the boundaries between the vertical and ground pixels can be used to infer 3D lines along which the image can be folded into a “pop-up” (after removing the sky pixels), as shown in Figure 6.41e. In related work, Saxena, Sun, and Ng (2009) develop a system that directly infers the depth and orientation of each pixel instead of using just three geometric class labels. We will examine techniques to infer depth from single images in more detail in Section 12.8.

6.4.5 Pose estimation

The inference of human pose (head, body, and limb locations and attitude) from a single images can be viewed as yet another kind of segmentation task. We have already discussed some pose estimation techniques in Section 6.3.2 on pedestrian detection section, as shown in Figure 6.25. Starting with the seminal work by Felzenszwalb and Huttenlocher (2005), 2D and 3D pose detection and estimation rapidly developed as an active research area, with important advances and datasets (Sigal and Black 2006a; Rogez, Rihan *et al.* 2008; Andriluka, Roth, and Schiele 2009; Bourdev and Malik 2009; Johnson and Everingham 2011; Yang and Ramanan 2011; Pishchulin, Andriluka *et al.* 2013; Sapp and Taskar 2013; Andriluka, Pishchulin *et al.* 2014).

More recently, deep networks have become the preferred technique to identify human body keypoints in order to convert these into pose estimates (Tompson, Jain *et al.* 2014;

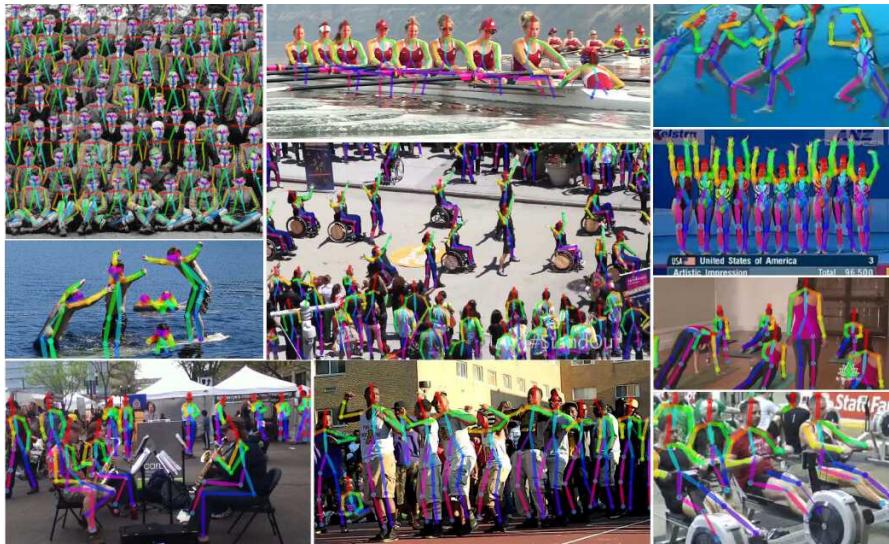


Figure 6.42 *OpenPose real-time multi-person 2D pose estimation (Cao, Simon et al. 2017)*
 © 2017 IEEE.

Toshev and Szegedy 2014; Pishchulin, Insafutdinov *et al.* 2016; Wei, Ramakrishna *et al.* 2016; Cao, Simon *et al.* 2017; He, Gkioxari *et al.* 2017; Hidalgo, Raaj *et al.* 2019; Huang, Zhu *et al.* 2020).²⁶ Figure 6.42 shows some of the impressive real-time multi-person 2D pose estimation results produced by the OpenPose system (Cao, Hidalgo *et al.* 2019).

The latest, most challenging, task in human pose estimation is the DensePose task introduced by Güler, Neverova, and Kokkinos (2018), where the task is to associate each pixel in RGB images of people with 3D points on a surface-based model, as shown in Figure 6.43. The authors provide dense annotations for 50,000 people appearing in COCO images and evaluate a number of correspondence networks, including their own DensePose-RCNN with several extensions. A more in-depth discussion on 3D human body modeling and tracking can be found in Section 13.6.4.

6.5 Video understanding

As we've seen in the previous sections of this chapter, image understanding mostly concerns itself with naming and delineating the objects and stuff in an image, although the relationships between objects and people are also sometimes inferred (Yao and Fei-Fei 2012; Gupta

²⁶You can find the leaderboards for human keypoint detection at <https://cocodataset.org>.

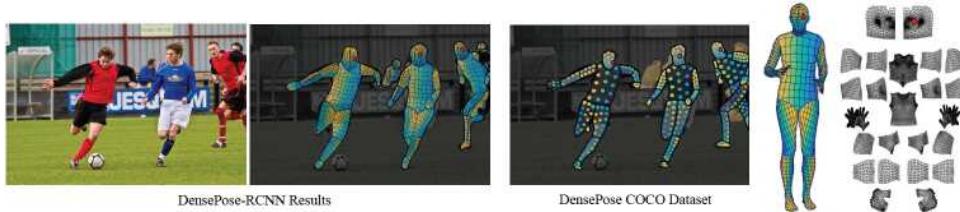


Figure 6.43 *Dense pose estimation aims at mapping all human pixels of an RGB image to the 3D surface of the human body (Güler, Neverova, and Kokkinos 2018) © 2018 IEEE. The paper describes DensePose-COCO, a large-scale ground-truth dataset containing manually annotated image-to-surface correspondences for 50K persons and a DensePose-RCNN trained to densely regress UV coordinates at multiple frames per second.*

and Malik 2015; Yatskar, Zettlemoyer, and Farhadi 2016; Gkioxari, Girshick *et al.* 2018). (We will look at the topic of describing complete images in the next section on vision and language.)

What, then, is video understanding? For many researchers, it starts with the detection and description of human actions, which are taken as the basic atomic units of videos. Of course, just as with images, these basic primitives can be chained into more complete descriptions of longer video sequences.

Human activity recognition began being studied in the 1990s, along with related topics such as human motion tracking, which we discuss in Sections 9.4.4 and 13.6.4. Aggarwal and Cai (1999) provide a comprehensive review of these two areas, which they call *human motion analysis*. Some of the techniques they survey use point and mesh tracking, as well as spatio-temporal signatures.

In the 2000s, attention shifted to spatio-temporal features, such as the clever use of optical flow in small patches to recognize sports activities (Efros, Berg *et al.* 2003) or spatio-temporal feature detectors for classifying actions in movies (Laptev, Marszalek *et al.* 2008), later combined with image context (Marszalek, Laptev, and Schmid 2009) and tracked feature trajectories (Wang and Schmid 2013). Poppe (2010), Aggarwal and Ryoo (2011), and Weinland, Ronfard, and Boyer (2011) provide surveys of algorithms from this decade. Some of the datasets used in this research include the KTH human motion dataset (Schüldt, Laptev, and Caputo 2004), the UCF sports action dataset (Rodriguez, Ahmed, and Shah 2008), the Hollywood human action dataset (Marszalek, Laptev, and Schmid 2009), UCF-101 (Soomro, Zamir, and Shah 2012), and the HMDB human motion database (Kuehne, Jhuang *et al.* 2011).

In the last decade, video understanding techniques have shifted to using deep networks

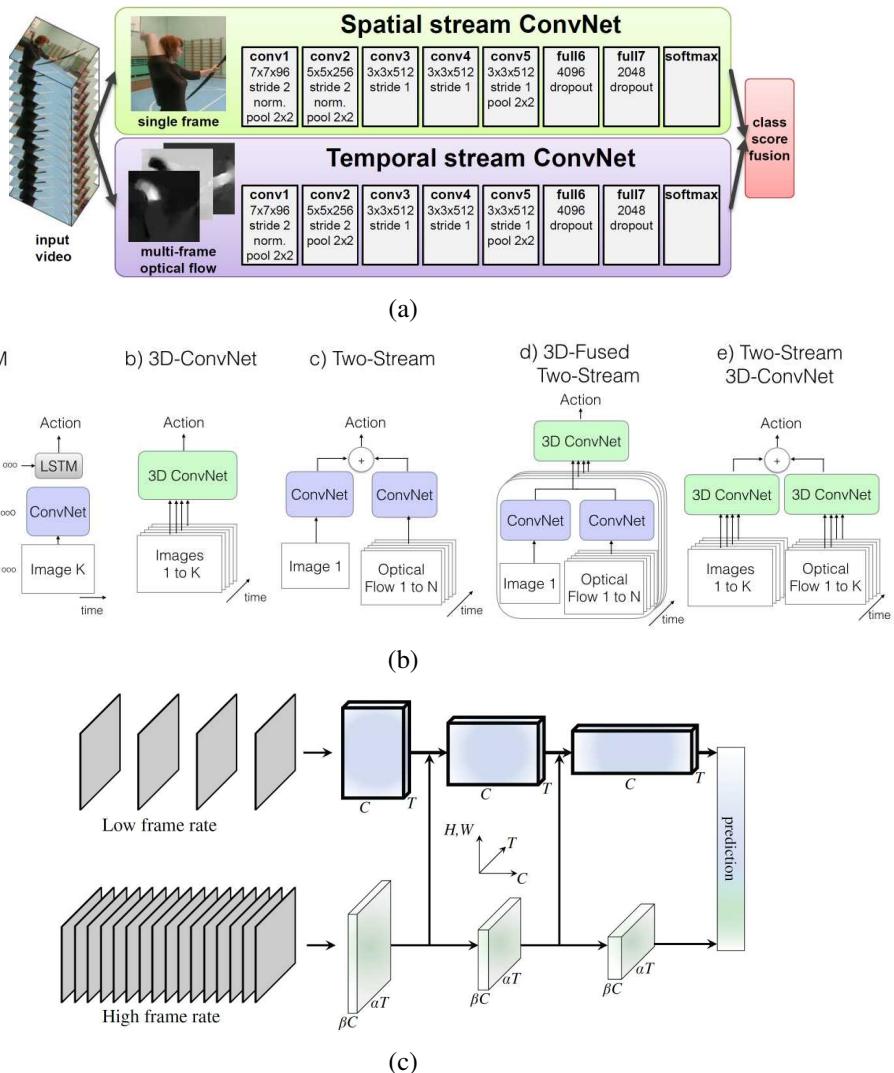


Figure 6.44 Video understanding using neural networks: (a) two-stream architecture for video classification © Simonyan and Zisserman (2014a); (b) some alternative video processing architectures (Carreira and Zisserman 2017) © 2017 IEEE; (c) a SlowFast network with a low frame rate, low temporal resolution Slow pathway and a high frame rate, higher temporal resolution Fast pathway (Feichtenhofer, Fan et al. 2019) © 2019 IEEE.

(Ji, Xu *et al.* 2013; Karpathy, Toderici *et al.* 2014; Simonyan and Zisserman 2014a; Tran, Bourdev *et al.* 2015; Feichtenhofer, Pinz, and Zisserman 2016; Carreira and Zisserman 2017; Varol, Laptev, and Schmid 2017; Wang, Xiong *et al.* 2019; Zhu, Li *et al.* 2020), sometimes combined with temporal models such as LSTMs (Baccouche, Mamalet *et al.* 2011; Donahue, Hendricks *et al.* 2015; Ng, Hausknecht *et al.* 2015; Srivastava, Mansimov, and Salakhudinov 2015).

While it is possible to apply these networks directly to the pixels in the video stream, e.g., using 3D convolutions (Section 5.5.1), researchers have also investigated using optical flow (Chapter 9.3) as an additional input. The resulting *two-stream architecture* was proposed by Simonyan and Zisserman (2014a) and is shown in Figure 6.44a. A later paper by Carreira and Zisserman (2017) compares this architecture to alternatives such as 3D convolutions on the pixel stream as well as hybrids of two streams and 3D convolutions (Figure 6.44b).

The latest architectures for video understanding have gone back to using 3D convolutions on the raw pixel stream (Tran, Wang *et al.* 2018, 2019; Kumawat, Verma *et al.* 2021). Wu, Feichtenhofer *et al.* (2019) store 3D CNN features into what they call a *long-term feature bank* to give a broader temporal context for action recognition. Feichtenhofer, Fan *et al.* (2019) propose a two-stream SlowFast architecture, where a slow pathway operates at a lower frame rate and is combined with features from a fast pathway with higher temporal sampling but fewer channels (Figure 6.44c). Some widely used datasets used for evaluating these algorithms are summarized in Table 6.3. They include Charades (Sigurdsson, Varol *et al.* 2016), YouTube8M (Abu-El-Haija, Kothari *et al.* 2016), Kinetics (Carreira and Zisserman 2017), “Something-something” (Goyal, Kahou *et al.* 2017), AVA (Gu, Sun *et al.* 2018), EPIC-KITCHENS (Damen, Doughty *et al.* 2018), and AVA-Kinetics (Li, Thotakuri *et al.* 2020). A nice exposition of these and other video understanding algorithms can be found in Johnson (2020, Lecture 18).

As with image recognition, researchers have also started using self-supervised algorithms to train video understanding systems. Unlike images, video clips are usually *multi-modal*, i.e., they contain audio tracks in addition to the pixels, which can be an excellent source of unlabeled supervisory signals (Alwassel, Mahajan *et al.* 2020; Patrick, Asano *et al.* 2020). When available at inference time, audio signals can improve the accuracy of such systems (Xiao, Lee *et al.* 2020).

Finally, while action recognition is the main focus of most recent video understanding work, it is also possible to classify videos into different scene categories such as “beach”, “fireworks”, or “snowing.” This problem is called *dynamic scene recognition* and can be addressed using spatio-temporal CNNs (Feichtenhofer, Pinz, and Wildes 2017).

| Name/URL | Metadata | Contents/Reference |
|--|--------------------------------|---|
| Charades https://prior.allenai.org/projects/charades | Actions, objects, descriptions | 9.8k videos Sigurdsson, Varol <i>et al.</i> (2016) |
| YouTube8M https://research.google.com/youtube8m | Entities | 4.8k visual entities, 8M videos Abu-El-Haija, Kothari <i>et al.</i> (2016) |
| Kinetics https://deepmind.com/research/open-source/kinetics | Action classes | 700 action classes, 650k videos Carreira and Zisserman (2017) |
| “Something-something” https://20bn.com/datasets/something-something | Actions with objects | 174 actions, 220k videos Goyal, Kahou <i>et al.</i> (2017) |
| AVA https://research.google.com/ava | Actions | 80 actions in 430 15-minute videos Gu, Sun <i>et al.</i> (2018) |
| EPIC-KITCHENS https://epic-kitchens.github.io | Actions and objects | 100 hours of egocentric videos Damen, Doughty <i>et al.</i> (2018) |

Table 6.3 *Datasets for video understanding and action recognition.*

6.6 Vision and language

The ultimate goal of much of computer vision research is not just to solve simpler tasks such as building 3D models of the world or finding relevant images, but to become an essential component of *artificial general intelligence* (AGI). This requires vision to integrate with other components of artificial intelligence such as speech and language understanding and synthesis, logical inference, and commonsense and specialized knowledge representation and reasoning.

Advances in speech and language processing have enabled the widespread deployment of speech-based intelligent virtual assistants such as Siri, Google Assistant, and Alexa. Earlier in this chapter, we’ve seen how computer vision systems can name individual objects in images and find similar images by appearance or keywords. The next natural step of integration with other AI components is to merge vision and language, i.e., *natural language processing* (NLP).

While this area has been studied for a long time (Duygulu, Barnard *et al.* 2002; Farhadi, Hejrati *et al.* 2010), the last decade has seen a rapid increase in performance and capabilities (Mogadala, Kalimuthu, and Klakow 2021; Gan, Yu *et al.* 2020). An example of this is the BabyTalk system developed by Kulkarni, Premraj *et al.* (2013), which first detects objects, their attributes, and their positional relationships, then infers a likely compatible labeling of these objects, and finally generates an image caption, as shown in Figure 6.45a.

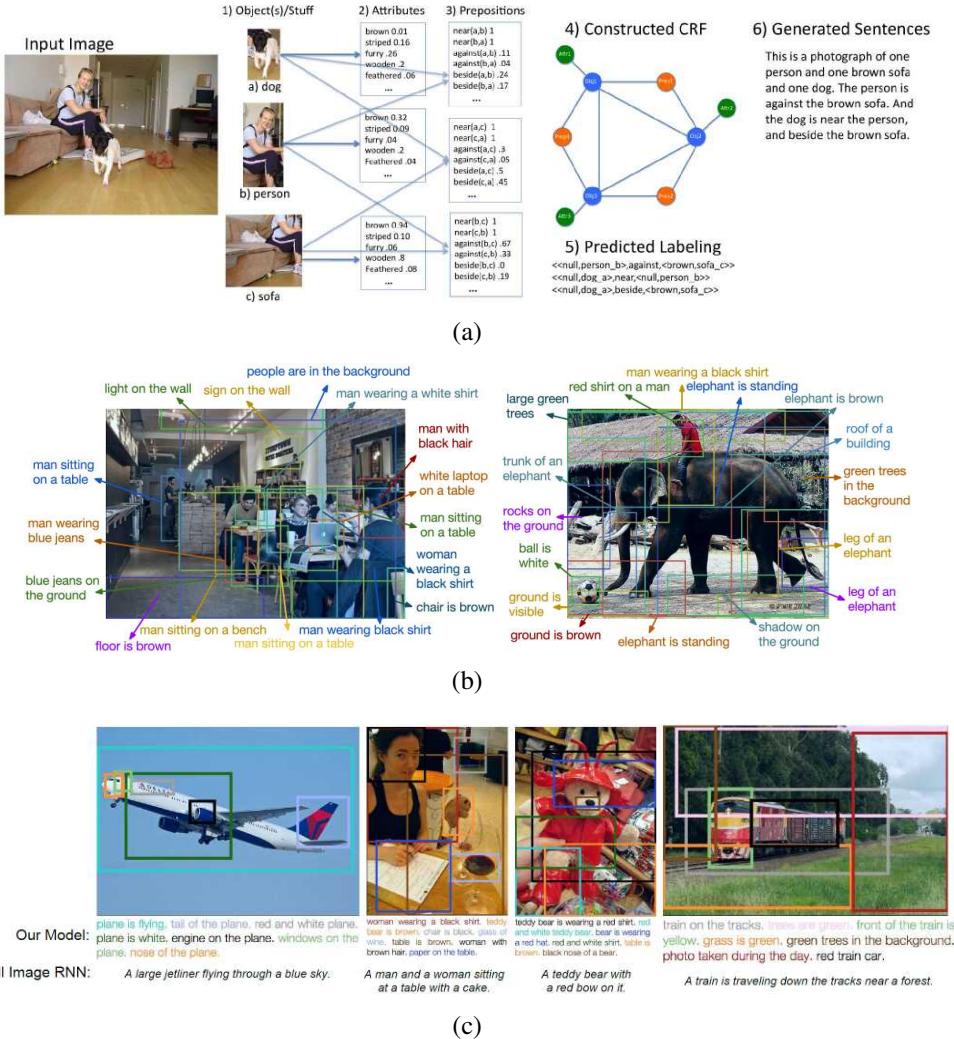
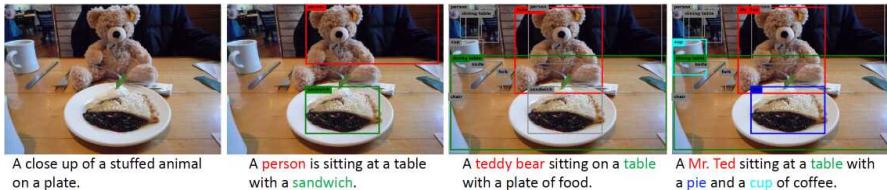


Figure 6.45 Image captioning systems: (a) BabyTalk detects objects, attributes, and positional relationships and composes these into image captions (Kulkarni, Premraj et al. 2013) © 2013 IEEE; (b–c) DenseCap associates word phrases with regions and then uses an RNN to construct plausible sentences (Johnson, Karpathy, and Fei-Fei 2016) © 2016 IEEE.



(a)



(b)

Figure 6.46 *Image captioning with attention: (a) The “Show, Attend, and Tell” system, which uses hard attention to align generated words with image regions © Xu, Ba et al. (2015); (b) Neural Baby Talk captions generated using different detectors, showing the association between words and grounding regions (Lu, Yang et al. 2018) © 2018 IEEE.*

Visual captioning

The next few years brought a veritable explosion of papers on the topic of image captioning and description, including (Chen and Lawrence Zitnick 2015; Donahue, Hendricks *et al.* 2015; Fang, Gupta *et al.* 2015; Karpathy and Fei-Fei 2015; Vinyals, Toshev *et al.* 2015; Xu, Ba *et al.* 2015; Johnson, Karpathy, and Fei-Fei 2016; Yang, He *et al.* 2016; You, Jin *et al.* 2016). Many of these systems combine CNN-based image understanding components (mostly object and human action detectors) with RNNs or LSTMs to generate the description, often in conjunction with other techniques such as multiple instance learning, maximum entropy language models, and visual attention. One somewhat surprising early result was that nearest-neighbor techniques, i.e., finding sets of similar looking images with captions and then creating a consensus caption, work surprisingly well (Devlin, Gupta *et al.* 2015).

Over the last few years, attention-based systems have continued to be essential components of image captioning systems (Lu, Xiong *et al.* 2017; Anderson, He *et al.* 2018; Lu, Yang *et al.* 2018). Figure 6.46 shows examples from two such papers, where each word in the generated caption is *grounded* with a corresponding image region. The CVPR 2020 tutorial by (Zhou 2020) summarizes over two dozen related papers from the last five years, including papers that use transformers (Section 5.5.3) to do the captioning. It also covers video descrip-



Figure 6.47 An adversarial typographic attack used against CLIP (Radford, Kim et al. 2021) discovered by ©Goh, Cammarata et al. (2021). Instead of predicting the object that exists in the scene, CLIP predicts the output based on the adversarial handwritten label.

tion and dense video captioning (Aafaq, Mian et al. 2019; Zhou, Kalantidis et al. 2019) and vision-language pre-training (Sun, Myers et al. 2019; Zhou, Palangi et al. 2020; Li, Yin et al. 2020). The tutorial also has lectures on visual question answering and reasoning (Gan 2020), text-to-image synthesis (Cheng 2020), and vision-language pre-training (Yu, Chen, and Li 2020).

For the task of image classification (Section 6.2), one of the major restrictions is that a model can only predict a label from the discrete pre-defined set of labels it trained on. CLIP (Radford, Kim et al. 2021) proposes an alternative approach that relies on image captions to enable zero-shot transfer to any possible set of labels. Given an image with a set of labels (e.g., {dog, cat, . . . , house}), CLIP predicts the label that maximizes the probability that the image is captioned with a prompt similar to “A photo of a {label}”. Section 5.4.7 discusses the training aspect of CLIP, which collects 400 million text-image pairs and uses contrastive learning to determine how likely it is for an image to be paired with a caption.

Remarkably, without having seen or fine-tuned to many popular image classification benchmarks (e.g., ImageNet, Caltech 101), CLIP can outperform independently fine-tuned ResNet-50 models supervised on each specific dataset. Moreover, compared to state-of-the-art classification models, CLIP’s zero-shot generalization is significantly more robust to dataset distribution shifts, performing well on each of ImageNet Sketch (Wang, Ge et al. 2019), ImageNetV2 (Recht, Roelofs et al. 2019), and ImageNet-R (Hendrycks, Basart et al. 2020), without being specifically trained on any of them. In fact, Goh, Cammarata et al. (2021) found that CLIP units responded similarly with concepts presented in different modalities (e.g., an image of Spiderman, text of the word spider, and a drawing of Spiderman). Figure 6.47 shows the adversarial typographic attack they discovered that could fool CLIP. By simply placing a handwritten class label (e.g., iPod) on a real-world object (e.g., Apple), CLIP often predicted the class written on the label.

As with other areas of visual recognition and learning-based systems, datasets have played an important role in the development of vision and language systems. Some widely used datasets of images with captions include Conceptual Captions (Sharma, Ding et al. 2018),

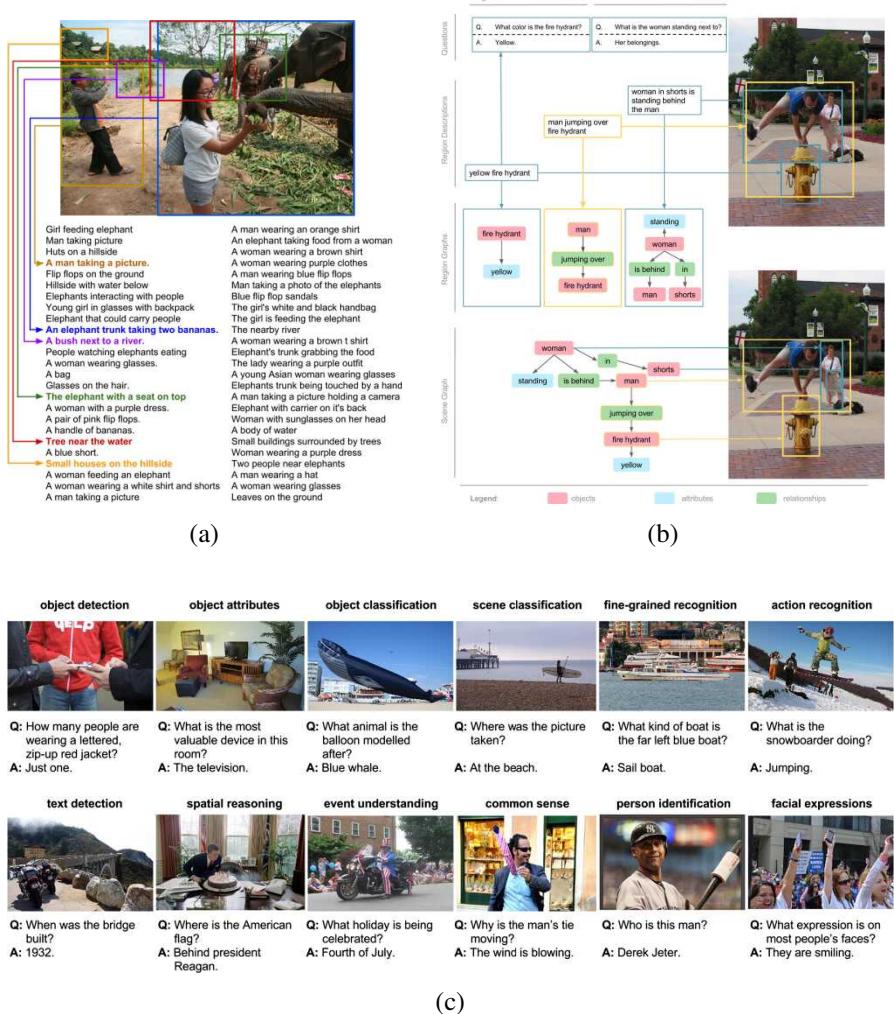


Figure 6.48 Images and data from the Visual Genome dataset (Krishna, Zhu et al. 2017) © 2017 Springer. (a) An example image with its region descriptors. (b) Each region has a graph representation of objects, attributes, and pairwise relationships, which are combined into a scene graph where all the objects are grounded to the image, and also associated questions and answers. (c) Some sample question and answer pairs, which cover a spectrum of visual tasks from recognition to high-level reasoning.

| Name/URL | Metadata | Contents/Reference |
|----------------------|---|--|
| Flickr30k (Entities) | Image captions (grounded) https://shannon.cs.illinois.edu/DenotationGraph http://bryanplummer.com/Flickr30kEntities | 30k images (+ bounding boxes) Young, Lai <i>et al.</i> (2014) Plummer, Wang <i>et al.</i> (2017) |
| COCO Captions | Whole image captions https://cocodataset.org/#captions-2015 | 1.5M captions, 330k images Chen, Fang <i>et al.</i> (2015) |
| Conceptual Captions | Whole image captions https://ai.google.com/research/ConceptualCaptions | 3.3M image caption pairs Sharma, Ding <i>et al.</i> (2018) |
| YFCC100M | Flickr metadata http://projects.dfki.uni-kl.de/yfcc100m | 100M images with metadata Thomee, Shamma <i>et al.</i> (2016) |
| Visual Genome | Dense annotations https://visualgenome.org | 108k images with region graphs Krishna, Zhu <i>et al.</i> (2017) |
| VQA v2.0 | Question/answer pairs https://visualqa.org | 265k images Goyal, Khot <i>et al.</i> (2017) |
| VCR | Multiple choice questions https://visualcommonsense.com | 110k movie clips, 290k QAs Zellers, Bisk <i>et al.</i> (2019) |
| GQA | Compositional QA https://visualreasoning.net | 22M questions on Visual Genome Hudson and Manning (2019) |
| VisDial | Dialogs for chatbot https://visualitydialog.org | 120k COCO images + dialogs Das, Kottur <i>et al.</i> (2017) |

Table 6.4 *Image datasets for vision and language research.*

the UIUC Pascal Sentence Dataset (Farhadi, Hejrati *et al.* 2010), the SBU Captioned Photo Dataset (Ordonez, Kulkarni, and Berg 2011), Flickr30k (Young, Lai *et al.* 2014), COCO Captions (Chen, Fang *et al.* 2015), and their extensions to 50 sentences per image (Vedantam, Lawrence Zitnick, and Parikh 2015) (see Table 6.4). More densely annotated datasets such as Visual Genome (Krishna, Zhu *et al.* 2017) describe different sub-regions of an image with their own phrases, i.e., provide *dense captioning*, as shown in Figure 6.48. YFCC100M (Thomee, Shamma *et al.* 2016) contains around 100M images from Flickr, but it only includes the raw user uploaded metadata for each image, such as the title, time of upload, description, tags, and (optionally) the location of the image.

Metrics for measuring sentence similarity also play an important role in the development of image captioning and other vision and language systems. Some widely used metrics include BLEU: BiLingual Evaluation Understudy (Papineni, Roukos *et al.* 2002), ROUGE: Recall Oriented Understudy of Gisting Evaluation (Lin 2004), METEOR: Metric for Evaluation of Translation with Explicit ORdering (Banerjee and Lavie 2005), CIDEr: Consensus-based

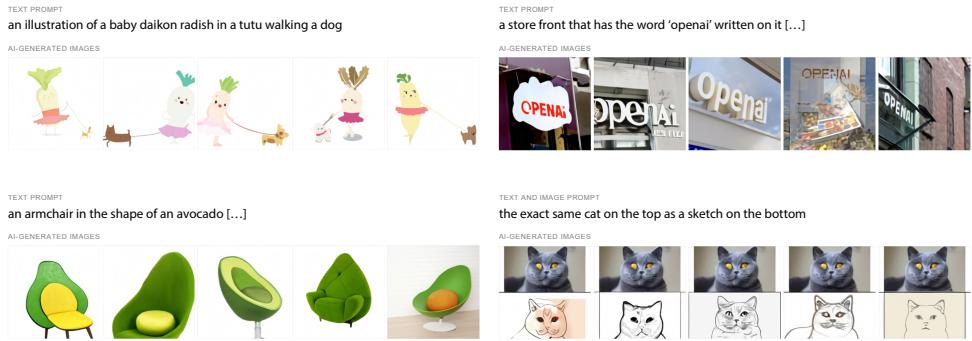


Figure 6.49 Qualitative text-to-image generation results from DALL-E, showing a wide range of generalization abilities ©Ramesh, Pavlov et al. (2021). The bottom right example provides a partially complete image prompt of a cat, along with text, and has the model fill in the rest of the image. The other three examples only start with the text prompt as input, with the model generating the entire image.

Image Description Evaluation (Vedantam, Lawrence Zitnick, and Parikh 2015), and SPICE: Semantic Propositional Image Caption Evaluation (Anderson, Fernando et al. 2016).²⁷

Text-to-image generation

The task of text-to-image generation is the inverse of visual captioning, i.e., given a text prompt, generate the image. Since images are represented in such high dimensionality, generating them to look coherent has historically been difficult. Generating images from a text prompt can be thought of as a generalization of generating images from a small set of class labels (Section 5.5.4). Since there is a near-infinite number of possible text prompts, successful models must be able to generalize from the relatively small fraction seen during training.

Early work on this task from Mansimov, Parisotto et al. (2016) used an RNN to iteratively draw an image from scratch. Their results showed some resemblance to the text prompts, although the generated images were quite blurred. The following year, Reed, Akata et al. (2016) applied a GAN to the problem, where unseen text prompts began to show promising results. Their generated images were relatively small (64×64), which was improved in later papers, which often first generated a small-scale image and then conditioned on that image and the text input to generate a higher-resolution image (Zhang, Xu et al. 2017, 2018; Xu, Zhang et al. 2018; Li, Qi et al. 2019).

DALL-E (Ramesh, Pavlov et al. 2021) uses orders of magnitude of more data (250 million

²⁷See https://www.cs.toronto.edu/~fidler/slides/2017/CSC2539/Kaustav_slides.pdf.

text-image pairs on the internet) and compute to achieve astonishing qualitative results (Figure 6.49).²⁸ Their approach produces promising results for generalizing beyond training data, even compositionally piecing together objects that are not often related (e.g., an armchair and an avocado), producing many styles (e.g., painting, cartoon, charcoal drawings), and working reasonably well with difficult objects (e.g., mirrors or text).

The model for DALL·E consists of two components: a VQ-VAE-2 (Section 5.5.4) and a decoder transformer (Section 5.5.3). The text is tokenized into 256 tokens, each of which is one of 16,384 possible vectors using a BPE-encoding (Sennrich, Haddow, and Birch 2015). The VQ-VAE-2 uses a codebook of size 8,192 (significantly larger than the codebook of size 512 used in the original VQ-VAE-2 paper) to compress images as a 32×32 grid of vector tokens. At inference time, DALL·E uses a transformer decoder, which starts with the 256 text tokens to autoregressively predict the 32×32 grid of image tokens. Given such a grid, the VQ-VAE-2 is able to use its decoder to generate the final RGB image of size 256×256 . To achieve better empirical results, DALL·E generates 512 image candidates and reranks them using CLIP (Radford, Kim *et al.* 2021), which determines how likely a given caption is associated with a given image.

An intriguing extension of DALL·E is to use the VQ-VAE-2 encoder to predict a subset of the compressed image tokens. For instance, suppose we are given a text input and an image. The text input can be tokenized into its 256 tokens, and one can obtain the 32×32 image tokens using the VQ-VAE-2 encoder. If we then discard the bottom half of the image tokens, the transformer decoder can be used to autoregressively predict which tokens might be there. These tokens, along with the non-discarded ones from the original image, can be passed into the VQ-VAE-2 decoder to produce a completed image. Figure 6.49 (bottom right) shows how such a text and partial image prompt can be used for applications such as image-to-image translation (Section 5.5.4).

Visual Question Answering and Reasoning

Image and video captioning are useful tasks that bring us closer to building artificially intelligent systems, as they demonstrate the ability to put together visual cues such as object identities, attributes, and actions. However, it remains unclear if the system has understood the scene at a deeper level and if it can reason about the constituent pieces and how they fit together.

To address these concerns, researchers have been building *visual question answering* (VQA) systems, which require the vision algorithm to answer open-ended questions about

²⁸Play with the results at <https://openai.com/blog/dall-e>.

the image, such as the ones shown in Figure 6.48c. A lot of this work started with the creation of the Visual Question Answering (VQA) dataset (Antol, Agrawal *et al.* 2015), which spurred a large amount of subsequent research. The following year, VQA v2.0 improved this dataset by creating a *balanced* set of image pairs, where each question had different answers in the two images (Goyal, Khot *et al.* 2017).²⁹ This dataset was further extended to reduce the influence of prior assumptions and data distributions and to encourage answers to be grounded in the images (Agrawal, Batra *et al.* 2018).

Since then, many additional VQA datasets have been created. These include the VCR dataset for visual commonsense reasoning (Zellers, Bisk *et al.* 2019) and the GQA dataset and metrics for evaluating visual reasoning and compositional question answering (Hudson and Manning 2019), which is built on top of the information about objects, attributes, and relations provided through the Visual Genome scene graphs (Krishna, Zhu *et al.* 2017). A discussion of these and other datasets for VQA can be found in the CVPR 2020 tutorial by Gan (2020), including datasets that test visual grounding and referring expression comprehension, visual entailment, using external knowledge, reading text, answering sub-questions, and using logic. Some of these datasets are summarized in Table 6.4.

As with image and video captioning, VQA systems use various flavors of attention to associate pixel regions with semantic concepts (Yang, He *et al.* 2016). However, instead of using sequence models such as RNNs, LSTMs, or transformers to generate text, the natural language question is first parsed to produce an encoding that is then fused with the image embedding to generate the desired answer.

The image semantic features can either be computed on a coarse grid, or a “bottom-up” object detector can be combined with a “top-down” attention mechanism to provide feature weightings (Anderson, He *et al.* 2018). In recent years, the pendulum has swung back and forth between techniques that use bottom-up regions and gridded feature descriptors, with two of the recent best-performing algorithms going back to the simpler (and much faster) gridded approach (Jiang, Misra *et al.* 2020; Huang, Zeng *et al.* 2020). The CVPR 2020 tutorial by Gan (2020) discusses these and dozens of other VQA systems as well as their subcomponents, such as multimodal fusion variants (bilinear pooling, alignment, relational reasoning), neural module networks, robust VQA, and multimodal pre-training. The survey by Mogadala, Kalimuthu, and Klakow (2021) and the annual VQA Challeng workshop (Shrivastava, Hudson *et al.* 2020) are also excellent sources of additional information. And if you would like to test out the current state of VQA systems, you can upload your own image to <https://vqa.cloudcv.org> and ask the system your own questions.

²⁹<https://visualqa.org>

Visual Dialog. An even more challenging version of VQA is *visual dialog*, where a chatbot is given an image and asked to answer open-ended questions about the image while also referring to previous elements of the conversation. The VisDial dataset was the earliest to be widely used for this task (Das, Kottur *et al.* 2017).³⁰ You can find pointers to systems that have been developed for this task at the Visual Dialog workshop and challenge (Shrivastava, Hudson *et al.* 2020). There’s also a chatbot at <https://visualchatbot.cloudcv.org> where you can upload your own image and start a conversation, which can sometimes lead to humorous (or weird) outcomes (Shane 2019).

Vision-language pre-training. As with many other recognition tasks, pre-training has had some dramatic success in the last few years, with systems such as ViLBERT (Lu, Batra *et al.* 2019), Oscar (Li, Yin *et al.* 2020), and many other systems described in the CVPR 2020 tutorial on self-supervised learning for vision-and-language (Yu, Chen, and Li 2020).

6.7 Additional reading

Unlike machine learning or deep learning, there are no recent textbooks or surveys devoted specifically to the general topics of image recognition and scene understanding. Some earlier surveys (Pinz 2005; Andreopoulos and Tsotsos 2013) and collections of papers (Ponce, Hebert *et al.* 2006; Dickinson, Leonardis *et al.* 2007) review the “classic” (pre-deep learning) approaches, but given the tremendous changes in the last decade, many of these techniques are no longer used. Currently, some of the best sources for the latest material, in addition to this chapter and university computer vision courses, are tutorials at the major vision conferences such as ICCV (Xie, Girshick *et al.* 2019), CVPR (Girshick, Kirillov *et al.* 2020), and ECCV (Xie, Girshick *et al.* 2020). Image recognition datasets such as those listed in Tables 6.1–6.4 that maintain active leaderboards can also be a good source for recent papers.

Algorithms for instance recognition, i.e., the detection of static manufactured objects that only vary slightly in appearance but may vary in 3D pose, are still often based on detecting 2D points of interest and describing them using viewpoint-invariant descriptors, as discussed in Chapter 7 and (Lowe 2004), Rothganger, Lazebnik *et al.* (2006), and Gordon and Lowe (2006). In more recent years, attention has shifted to the more challenging problem of *instance retrieval* (also known as *content-based image retrieval*), in which the number of images being searched can be very large (Sivic and Zisserman 2009). Section 7.1.4 in the next chapter reviews such techniques, as does the survey in (Zheng, Yang, and Tian 2018). This topic is also related to visual similarity search (Bell and Bala 2015; Arandjelovic, Gronat *et*

³⁰<https://visualdialog.org>

al. 2016; Song, Xiang *et al.* 2016; Gordo, Almazán *et al.* 2017; Rawat and Wang 2017; Bell, Liu *et al.* 2020), which was covered in Section 6.2.3.

A number of surveys, collections of papers, and course notes have been written on the topic of feature-based whole image (single-object) category recognition (Pinz 2005; Ponce, Hebert *et al.* 2006; Dickinson, Leonardis *et al.* 2007; Fei-Fei, Fergus, and Torralba 2009). Some of these papers use a bag of words or keypoints (Csurka, Dance *et al.* 2004; Lazebnik, Schmid, and Ponce 2006; Csurka, Dance *et al.* 2006; Grauman and Darrell 2007b; Zhang, Marszałek *et al.* 2007; Boiman, Shechtman, and Irani 2008; Ferencz, Learned-Miller, and Malik 2008). Other papers recognize objects based on their contours, e.g., using shape contexts (Belongie, Malik, and Puzicha 2002) or other techniques (Shotton, Blake, and Cipolla 2005; Opelt, Pinz, and Zisserman 2006; Ferrari, Tuytelaars, and Van Gool 2006a).

Many object recognition algorithms use part-based decompositions to provide greater invariance to articulation and pose. Early algorithms focused on the relative positions of the parts (Fischler and Elschlager 1973; Kanade 1977; Yuille 1991) while later algorithms used more sophisticated models of appearance (Felzenszwalb and Huttenlocher 2005; Fergus, Perona, and Zisserman 2007; Felzenszwalb, McAllester, and Ramanan 2008). Good overviews on part-based models for recognition can be found in the course notes by Fergus (2009). Carneiro and Lowe (2006) discuss a number of graphical models used for part-based recognition, which include trees and stars, k -fans, and constellations.

Classical recognition algorithms often used scene context as part of their recognition strategy. Representative papers in this area include Torralba (2003), Torralba, Murphy *et al.* (2003), Rabinovich, Vedaldi *et al.* (2007), Russell, Torralba *et al.* (2007), Sudderth, Torralba *et al.* (2008), and Divvala, Hoiem *et al.* (2009). Machine learning also became a key component of classical object detection and recognition algorithms (Felzenszwalb, McAllester, and Ramanan 2008; Sivic, Russell *et al.* 2008), as did exploiting large human-labeled databases (Russell, Torralba *et al.* 2007; Torralba, Freeman, and Fergus 2008).

The breakthrough success of the “AlexNet” SuperVision system of Krizhevsky, Sutskever, and Hinton (2012) shifted the focus in category recognition research from feature-based approaches to deep neural networks. The rapid improvement in recognition accuracy, captured in Figure 5.40 and described in more detail in Section 5.4.3 has been driven to a large degree by deeper networks and better training algorithms, and also in part by larger (unlabeled) training datasets (Section 5.4.7).

More specialized recognition systems such as those for recognizing faces underwent a similar evolution. While some of the earliest approaches to face recognition involved finding the distinctive image features and measuring the distances between them (Fischler and Elschlager 1973; Kanade 1977; Yuille 1991), later approaches relied on comparing gray-

level images, often projected onto lower dimensional subspaces (Turk and Pentland 1991; Belhumeur, Hespanha, and Kriegman 1997; Heisele, Ho *et al.* 2003) or local binary patterns (Ahonen, Hadid, and Pietikäinen 2006). A variety of shape and pose deformation models were also developed (Beymer 1996; Vetter and Poggio 1997), including Active Shape Models (Cootes, Cooper *et al.* 1995), 3D Morphable Models (Blanz and Vetter 1999; Egger, Smith *et al.* 2020), and Active Appearance Models (Cootes, Edwards, and Taylor 2001; Matthews and Baker 2004; Ramnath, Koterba *et al.* 2008). Additional information about classic face recognition algorithms can be found in a number of surveys and books on this topic (Chellappa, Wilson, and Sirohey 1995; Zhao, Chellappa *et al.* 2003; Li and Jain 2005).

The concept of shape models for *frontalization* continued to be used as the community shifted to deep neural network approaches (Taigman, Yang *et al.* 2014). Some more recent deep face recognizers, however, omit the frontalization stage and instead use data augmentation to create synthetic inputs with a larger variety of poses (Schroff, Kalenichenko, and Philbin 2015; Parkhi, Vedaldi, and Zisserman 2015). Masi, Wu *et al.* (2018) provide an excellent tutorial and survey on deep face recognition, including a list of widely used training and testing datasets, a discussion of frontalization and dataset augmentation, and a section on training losses.

As the problem of whole-image (single object) category recognition became more “solved”, attention shifted to multiple object delineation and labeling, i.e., object detection. Object detection was originally studied in the context of specific categories such as faces, pedestrians, cars, etc. Seminal papers in face detection include those by Osuna, Freund, and Girosi (1997); Sung and Poggio (1998); Rowley, Baluja, and Kanade (1998); Viola and Jones (2004); Heisele, Ho *et al.* (2003), with Yang, Kriegman, and Ahuja (2002) providing a comprehensive survey of early work in this field. Early work in pedestrian and car detection was carried out by Gavrila and Philomin (1999); Gavrila (1999); Papageorgiou and Poggio (2000); Schneiderman and Kanade (2004). Subsequent papers include (Mikolajczyk, Schmid, and Zisserman 2004; Dalal and Triggs 2005; Leibe, Seemann, and Schiele 2005; Andriluka, Roth, and Schiele 2009; Dollár, Belongie, and Perona 2010; Felzenszwalb, Girshick *et al.* 2010).

Modern generic object detectors are typically constructed using a region proposal algorithm (Uijlings, Van De Sande *et al.* 2013; Zitnick and Dollár 2014) that then feeds selected regions of the image (either as pixels or precomputed neural features) into a multi-way classifier, resulting in architectures such as R-CNN (Girshick, Donahue *et al.* 2014), Fast R-CNN (Girshick 2015), Faster R-CCNN (Ren, He *et al.* 2015), and FPN (Lin, Dollár *et al.* 2017). An alternative to this two-stage approach is a *single-stage network*, which uses a single network to output detections at a variety of locations. Examples of such architectures include

SSD (Liu, Anguelov *et al.* 2016), RetinaNet (Lin, Goyal *et al.* 2017), and YOLO (Redmon, Divvala *et al.* 2016; Redmon and Farhadi 2017, 2018; Bochkovskiy, Wang, and Liao 2020). These and more recent convolutional object detectors are described in the recent survey by Jiao, Zhang *et al.* (2019).

While object detection can be sufficient in many computer vision applications such as counting cars or pedestrians or even describing images, a detailed pixel-accurate labeling can be potentially even more useful, e.g., for photo editing. This kind of labeling comes in several flavors, including semantic segmentation (what stuff is this?), instance segmentation (which countable object is this?), panoptic segmentation (what stuff or object is it?). One early approach to this problem was to pre-segment the image into pieces and then match these pieces to portions of the model (Mori, Ren *et al.* 2004; Russell, Efros *et al.* 2006; Borenstein and Ullman 2008; Gu, Lim *et al.* 2009). Another popular approach was to use conditional random fields (Kumar and Hebert 2006; He, Zemel, and Carreira-Perpiñán 2004; Winn and Shotton 2006; Rabinovich, Vedaldi *et al.* 2007; Shotton, Winn *et al.* 2009), which at that time produced some of the best results on the PASCAL VOC segmentation challenge. Modern semantic segmentation algorithms use pyramidal fully-convolutional architectures to map input pixels to class labels (Long, Shelhamer, and Darrell 2015; Zhao, Shi *et al.* 2017; Xiao, Liu *et al.* 2018; Wang, Sun *et al.* 2020).

The more challenging task of instance segmentation, where each distinct object gets its own unique label, is usually tackled using a combination of object detectors and per-object segmentation, as exemplified in the seminal Mask R-CNN paper by He, Gkioxari *et al.* (2017). Follow-on work uses more sophisticated backbone architectures (Liu, Qi *et al.* 2018; Chen, Pang *et al.* 2019). Two recent workshops that highlight the latest results in this area are the COCO + LVIS Joint Recognition Challenge (Kirillov, Lin *et al.* 2020) and the Robust Vision Challenge (Zendel *et al.* 2020).

Putting semantic and instance segmentation together has long been a goal of semantic scene understanding (Yao, Fidler, and Urtasun 2012; Tighe and Lazebnik 2013; Tu, Chen *et al.* 2005). Doing this on a per-pixel level results in a *panoptic segmentation* of the scene, where all of the objects are correctly segmented and the remaining stuff is correctly labeled (Kirillov, He *et al.* 2019; Kirillov, Girshick *et al.* 2019). The COCO dataset has now been extended to include a panoptic segmentation task, on which some recent results can be found in the ECCV 2020 workshop on this topic (Kirillov, Lin *et al.* 2020).

Research in video understanding, or more specifically human activity recognition, dates back to the 1990s; some good surveys include (Aggarwal and Cai 1999; Poppe 2010; Aggarwal and Ryoo 2011; Weinland, Ronfard, and Boyer 2011). In the last decade, video understanding techniques shifted to using deep networks (Ji, Xu *et al.* 2013; Karpathy, Toderici *et*

al. 2014; Simonyan and Zisserman 2014a; Donahue, Hendricks *et al.* 2015; Tran, Bourdev *et al.* 2015; Feichtenhofer, Pinz, and Zisserman 2016; Carreira and Zisserman 2017; Tran, Wang *et al.* 2019; Wu, Feichtenhofer *et al.* 2019; Feichtenhofer, Fan *et al.* 2019). Some widely used datasets used for evaluating these algorithms are summarized in Table 6.3.

While associating words with images has been studied for a while (Duygulu, Barnard *et al.* 2002), sustained research into describing images with captions and complete sentences started in the early 2010s (Farhadi, Hejrati *et al.* 2010; Kulkarni, Premraj *et al.* 2013). The last decade has seen a rapid increase in performance and capabilities of such systems (Mogadala, Kalimuthu, and Klakow 2021; Gan, Yu *et al.* 2020). The first sub-problem to be widely studied was image captioning (Donahue, Hendricks *et al.* 2015; Fang, Gupta *et al.* 2015; Karpathy and Fei-Fei 2015; Vinyals, Toshev *et al.* 2015; Xu, Ba *et al.* 2015; Devlin, Gupta *et al.* 2015), with later systems using attention mechanisms (Anderson, He *et al.* 2018; Lu, Yang *et al.* 2018). More recently, researchers have developed systems for visual question answering (Antol, Agrawal *et al.* 2015) and visual commonsense reasoning (Zellers, Bisk *et al.* 2019).

The CVPR 2020 tutorial on recent advances in visual captioning (Zhou 2020) summarizes over two dozen related papers from the last five years, including papers that use Transformers to do the captioning. It also covers video description and dense video captioning (Aafaq, Mian *et al.* 2019; Zhou, Kalantidis *et al.* 2019) and vision-language pre-training (Sun, Myers *et al.* 2019; Zhou, Palangi *et al.* 2020; Li, Yin *et al.* 2020). The tutorial also has lectures on visual question answering and reasoning (Gan 2020), text-to-image synthesis (Cheng 2020), and vision-language pre-training (Yu, Chen, and Li 2020).

6.8 Exercises

Ex 6.1: Pre-trained recognition networks. Find a pre-trained network for image classification, segmentation, or some other task such as face recognition or pedestrian detection.

After running the network, can you characterize the most common kinds of errors the network is making? Create a “confusion matrix” indicating which categories get classified as other categories. Now try the network on your own data, either from a web search or from your personal photo collection. Are there surprising results?

My own favorite code to try is Detectron2,³¹ which I used to generate the panoptic segmentation results shown in Figure 6.39.

³¹Click on the “Colab Notebook” link at <https://github.com/facebookresearch/detectron2> and then edit the input image URL to try your own.

Ex 6.2: Re-training recognition networks. After analyzing the performance of your pre-trained network, try re-training it on the original dataset on which it was trained, but with modified parameters (numbers of layers, channels, training parameters) or with additional examples. Can you get the network to perform more to your liking?

Many of the online tutorials, such as the Detectron2 Collab notebook mentioned above, come with instructions on how to re-train the network from scratch on a different dataset. Can you create your own dataset, e.g., using a web search and figure out how to label the examples? A low effort (but not very accurate) way is to trust the results of the web search. Russakovsky, Deng *et al.* (2015), Kovashka, Russakovsky *et al.* (2016), and other papers on image datasets discuss the challenges in obtaining accurate labels.

Train your network, try to optimize its architecture, and report on the challenges you faced and discoveries you made.

Note: the following exercises were suggested by Matt Deitke.

Ex 6.3: Image perturbations. Download either ImageNet or Imagenette.³² Now, perturb each image by adding a small square to the top left of the image, where the color of the square is unique for each label, as shown in the following figure:



(a) cassette player



(b) golf ball



(c) English Springer

Using any image classification model,³³ e.g., ResNet, EfficientNet, or ViT, train the model from scratch on the perturbed images. Does the model overfit to the color of the square and ignore the rest of the image? When evaluating the model on the training and validation data, try adversarially swapping colors between different labels.

Ex 6.4: Image normalization. Using the same dataset downloaded for the previous exercise, take a ViT model and remove all the intermediate layer normalization operations. Are you able to train the network? Using techniques in Li, Xu *et al.* (2018), how do the plots of the loss landscape appear with and without the intermediate layer normalization operations?

Ex 6.5: Semantic segmentation. Explain the differences between instance segmentation, semantic segmentation, and panoptic segmentation. For each type of segmentation, can it be post-processed to obtain the other kinds of segmentation?

³²Imagenette, <https://github.com/fastai/imagenette>, is a smaller 10-class subset of ImageNet that is easier to use with limited computing resources. .

³³You may find the PyTorch Image Models at <https://github.com/rwightman/pytorch-image-models> useful.

Ex 6.6: Class encoding. Categorical inputs to a neural network, such as a word or object, can be encoded with one-hot encoded vector.³⁴ However, it is common to pass the one-hot encoded vector through an embedding matrix, where the output is then passed into the neural network loss function. What are the advantages of vector embedding over using one-hot encoding?

Ex 6.7: Object detection. For object detection, how do the number of parameters for DETR, Faster-RCNN, and YOLOv4 compare? Try training each of them on MS COCO. Which one tends to train the slowest? How long does it take each model to evaluate a single image at inference time?

Ex 6.8: Image classification vs. description. For image classification, list at least two significant differences between using categorical labels and natural language descriptions.

Ex 6.9: ImageNet Sketch. Try taking several pre-trained models on ImageNet and evaluating them, without any fine-tuning, on ImageNet Sketch (Wang, Ge *et al.* 2019). For each of these models, to what extent does the performance drop due to the shift in distribution?

Ex 6.10: Self-supervised learning. Provide examples of self-supervised learning pretext tasks for each of the following data types: static images, videos, and vision-and-language.

Ex 6.11: Video understanding. For many video understanding tasks, we may be interested in tracking an object through time. Why might this be preferred to making predictions independently for each frame? Assume that inference speed is not a problem.

Ex 6.12: Fine-tuning a new head. Take the backbone of a network trained for object classification and fine-tune it for object detection with a variant of YOLO. Why might it be desirable to freeze the early layers of the network?

Ex 6.13: Movie understanding. Currently, most video understanding networks, such as those discussed in this chapter, tend to only deal with short video clips as input. What modifications might be necessary in order to operate over longer sequences such as an entire movie?

³⁴With a categorical variable, one-hot encoding is used to represent which label is chosen, i.e., when a label is chosen, its entry in the vector is 1 with all other entries being 0.

Chapter 7

Feature detection and matching

| | | |
|-------|--|-----|
| 7.1 | Points and patches | 419 |
| 7.1.1 | Feature detectors | 422 |
| 7.1.2 | Feature descriptors | 434 |
| 7.1.3 | Feature matching | 441 |
| 7.1.4 | Large-scale matching and retrieval | 448 |
| 7.1.5 | Feature tracking | 452 |
| 7.1.6 | <i>Application:</i> Performance-driven animation | 454 |
| 7.2 | Edges and contours | 455 |
| 7.2.1 | Edge detection | 456 |
| 7.2.2 | Contour detection | 461 |
| 7.2.3 | <i>Application:</i> Edge editing and enhancement | 465 |
| 7.3 | Contour tracking | 466 |
| 7.3.1 | Snakes and scissors | 467 |
| 7.3.2 | Level Sets | 474 |
| 7.3.3 | <i>Application:</i> Contour tracking and rotoscoping | 476 |
| 7.4 | Lines and vanishing points | 477 |
| 7.4.1 | Successive approximation | 477 |
| 7.4.2 | Hough transforms | 477 |
| 7.4.3 | Vanishing points | 481 |
| 7.5 | Segmentation | 483 |
| 7.5.1 | Graph-based segmentation | 486 |
| 7.5.2 | Mean shift | 487 |
| 7.5.3 | Normalized cuts | 489 |
| 7.6 | Additional reading | 491 |
| 7.7 | Exercises | 495 |

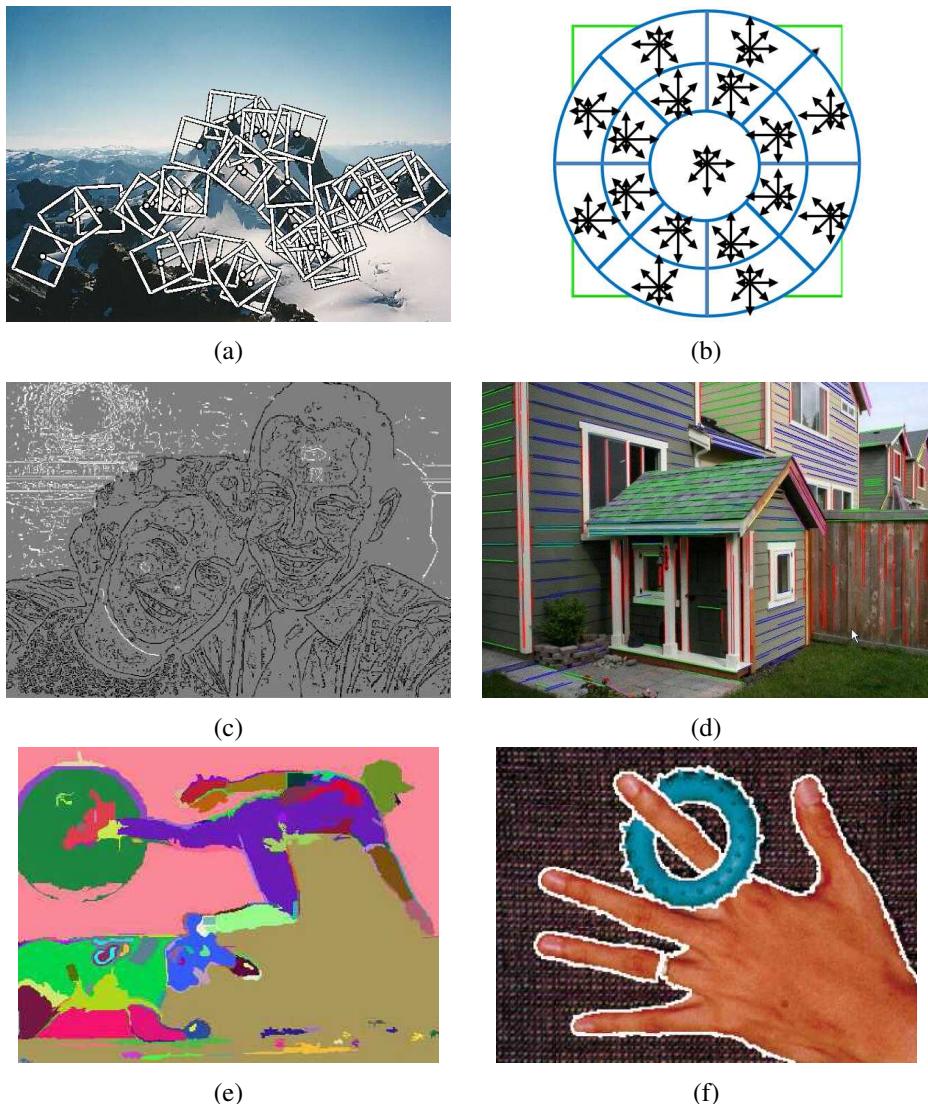


Figure 7.1 Feature detectors and descriptors can be used to analyze, describe and match images: (a) point-like interest operators (Brown, Szeliski, and Winder 2005) © 2005 IEEE; (b) GLOH descriptor (Mikolajczyk and Schmid 2005); (c) edges (Elder and Goldberg 2001) © 2001 IEEE; (d) straight lines (Sinha, Steedly et al. 2008) © 2008 ACM; (e) graph-based merging (Felzenszwalb and Huttenlocher 2004) © 2004 Springer; (f) mean shift (Comaniciu and Meer 2002) © 2002 IEEE.

Feature detection and matching are an essential component of many computer vision applications. Consider the two pairs of images shown in Figure 7.2. For the first pair, we may wish to *align* the two images so that they can be seamlessly stitched into a composite mosaic (Section 8.2). For the second pair, we may wish to establish a dense set of *correspondences* so that a 3D model can be constructed or an in-between view can be generated (Chapter 12). In either case, what kinds of *features* should you detect and then match to establish such an alignment or set of correspondences? Think about this for a few moments before reading on.

The first kind of feature that you may notice are specific locations in the images, such as mountain peaks, building corners, doorways, or interestingly shaped patches of snow. These kinds of localized features are often called *keypoint features* or *interest points* (or even *corners*) and are often described by the appearance of pixel patches surrounding the point location (Section 7.1). Another class of important features are *edges*, e.g., the profile of mountains against the sky (Section 7.2). These kinds of features can be matched based on their orientation and local appearance (edge profiles) and can also be good indicators of object boundaries and *occlusion* events in image sequences. Edges can be grouped into longer *curves* and *contours*, which can then be tracked (Section 7.3). They can also be grouped into *straight line segments*, which can be directly matched or analyzed to find *vanishing points* and hence internal and external camera parameters (Section 7.4).

In this chapter, we describe some practical approaches to detecting such features and also discuss how feature correspondences can be established across different images. Point features are now used in such a wide variety of applications that it is good practice to read and implement some of the algorithms from Section 7.1. Edges and lines provide information that is complementary to both keypoint and region-based descriptors and are well suited to describing the boundaries of manufactured objects. These alternative descriptors, while extremely useful, can be skipped in a short introductory course.

The last part of this chapter (Section 7.5) discusses bottom-up non-semantic segmentation techniques. While these were once widely used as essential components of both recognition and matching algorithms, they have mostly been supplanted by the *semantic segmentation* techniques we studied in Section 6.4. They are still used occasionally to group pixels together for faster or more reliable matching.

7.1 Points and patches

Point features can be used to find a sparse set of corresponding locations in different images, often as a precursor to computing camera pose (Chapter 11), which is a prerequisite for computing a denser set of correspondences using stereo matching (Chapter 12). Such cor-



Figure 7.2 Two pairs of images to be matched. What kinds of features might one use to establish a set of correspondences between these images?

respondences can also be used to align different images, e.g., when stitching image mosaics (Section 8.2) or high dynamic range images (Section 10.2), or performing video stabilization (Section 9.2.1). They are also used extensively to perform object instance recognition (Section 6.1). A key advantage of keypoints is that they permit matching even in the presence of clutter (occlusion) and large scale and orientation changes.

Feature-based correspondence techniques have been used since the early days of stereo matching (Hannah 1974; Moravec 1983; Hannah 1988) and subsequently gained popularity for image-stitching applications (Zoghami, Faugeras, and Deriche 1997; Brown and Lowe 2007) as well as fully automated 3D modeling (Beardsley, Torr, and Zisserman 1996; Schaffalitzky and Zisserman 2002; Brown and Lowe 2005; Snavely, Seitz, and Szeliski 2006).

There are two main approaches to finding feature points and their correspondences. The first is to find features in one image that can be accurately *tracked* using a local search technique, such as correlation or least squares (Section 7.1.5). The second is to independently detect features in all the images under consideration and then *match* features based on their local appearance (Section 7.1.3). The former approach is more suitable when images are taken from nearby viewpoints or in rapid succession (e.g., video sequences), while the latter is more suitable when a large amount of motion or appearance change is expected, e.g., in stitching together panoramas (Brown and Lowe 2007), establishing correspondences in *wide baseline stereo* (Schaffalitzky and Zisserman 2002), or performing object recognition

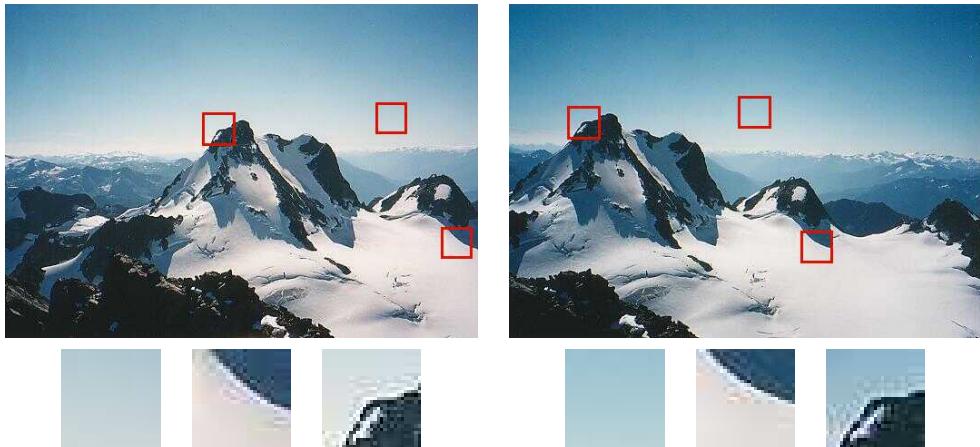


Figure 7.3 Image pairs with extracted patches below. Notice how some patches can be localized or matched with higher accuracy than others.

(Fergus, Perona, and Zisserman 2007).

In this section, we split the keypoint detection and matching pipeline into four separate stages. During the *feature detection* (extraction) stage (Section 7.1.1), each image is searched for locations that are likely to match well in other images. In the *feature description* stage (Section 7.1.2), each region around detected keypoint locations is converted into a more compact and stable (invariant) *descriptor* that can be matched against other descriptors. The *feature matching* stage (Sections 7.1.3 and 7.1.4) efficiently searches for likely matching candidates in other images. The *feature tracking* stage (Section 7.1.5) is an alternative to the third stage that only searches a small neighborhood around each detected feature and is therefore more suitable for video processing.

A wonderful example of all of these stages can be found in David Lowe's (2004) paper, which describes the development and refinement of his *Scale Invariant Feature Transform* (SIFT). Comprehensive descriptions of alternative techniques can be found in a series of survey and evaluation papers covering both feature detection (Schmid, Mohr, and Bauckhage 2000; Mikolajczyk, Tuytelaars *et al.* 2005; Tuytelaars and Mikolajczyk 2008) and feature descriptors (Mikolajczyk and Schmid 2005; Balntas, Lenc *et al.* 2020). Shi and Tomasi (1994) and Triggs (2004) also provide nice reviews of classic (pre-neural network) feature detection techniques.

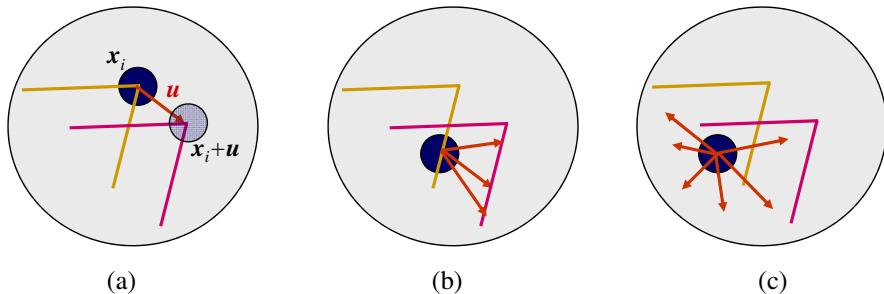


Figure 7.4 Aperture problems for different image patches: (a) stable (“corner-like”) flow; (b) classic aperture problem (barber-pole illusion); (c) textureless region. The two images I_0 (yellow) and I_1 (red) are overlaid. The red vector \mathbf{u} indicates the displacement between the patch centers and the $w(\mathbf{x}_i)$ weighting function (patch window) is shown as a dark circle.

7.1.1 Feature detectors

How can we find image locations where we can reliably find correspondences with other images, i.e., what are good features to track (Shi and Tomasi 1994; Triggs 2004)? Look again at the image pair shown in Figure 7.3 and at the three sample *patches* to see how well they might be matched or tracked. As you may notice, textureless patches are nearly impossible to localize. Patches with large contrast changes (gradients) are easier to localize, although straight line segments at a single orientation suffer from the *aperture problem* (Horn and Schunck 1981; Lucas and Kanade 1981; Anandan 1989), i.e., it is only possible to align the patches along the direction *normal* to the edge direction (Figure 7.4b). Patches with gradients in at least two (significantly) different orientations are the easiest to localize, as shown schematically in Figure 7.4a.

These intuitions can be formalized by looking at the simplest possible matching criterion for comparing two image patches, i.e., their (weighted) summed square difference,

$$E_{\text{WSSD}}(\mathbf{u}) = \sum_i w(\mathbf{x}_i) [I_1(\mathbf{x}_i + \mathbf{u}) - I_0(\mathbf{x}_i)]^2, \quad (7.1)$$

where I_0 and I_1 are the two images being compared, $\mathbf{u} = (u, v)$ is the *displacement* vector, $w(\mathbf{x})$ is a spatially varying weighting (or window) function, and the summation i is over all the pixels in the patch. Note that this is the same formulation we later use to estimate motion between complete images (Section 9.1).

When performing feature detection, we do not know which other image locations the feature will end up being matched against. Therefore, we can only compute how stable this metric is with respect to small variations in position $\Delta\mathbf{u}$ by comparing an image patch against

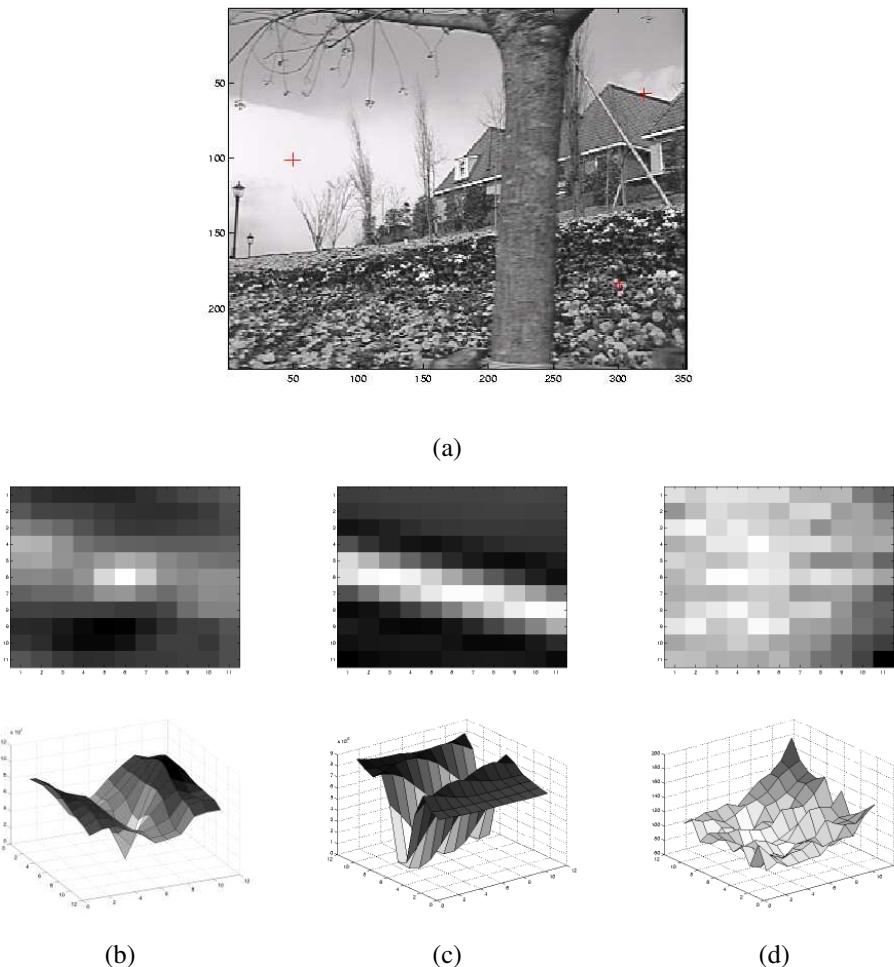


Figure 7.5 Three auto-correlation surfaces $E_{AC}(\Delta u)$ shown as both grayscale images and surface plots: (a) The original image is marked with three red crosses to denote where the auto-correlation surfaces were computed; (b) this patch is from the flower bed (good unique minimum); (c) this patch is from the roof edge (one-dimensional aperture problem); and (d) this patch is from the cloud (no good peak). Each grid point in figures b–d is one value of Δu .

itself, which is known as an *auto-correlation function* or *surface*

$$E_{AC}(\Delta \mathbf{u}) = \sum_i w(\mathbf{x}_i) [I_0(\mathbf{x}_i + \Delta \mathbf{u}) - I_0(\mathbf{x}_i)]^2 \quad (7.2)$$

(Figure 7.5).¹ Note how the auto-correlation surface for the textured flower bed (Figure 7.5b and the red cross in the lower right quadrant of Figure 7.5a) exhibits a strong minimum, indicating that it can be well localized. The correlation surface corresponding to the roof edge (Figure 7.5c) has a strong ambiguity along one direction, while the correlation surface corresponding to the cloud region (Figure 7.5d) has no stable minimum.

Using a Taylor Series expansion of the image function $I_0(\mathbf{x}_i + \Delta \mathbf{u}) \approx I_0(\mathbf{x}_i) + \nabla I_0(\mathbf{x}_i) \cdot \Delta \mathbf{u}$ (Lucas and Kanade 1981; Shi and Tomasi 1994), we can approximate the auto-correlation surface as

$$E_{AC}(\Delta \mathbf{u}) = \sum_i w(\mathbf{x}_i) [I_0(\mathbf{x}_i + \Delta \mathbf{u}) - I_0(\mathbf{x}_i)]^2 \quad (7.3)$$

$$\approx \sum_i w(\mathbf{x}_i) [I_0(\mathbf{x}_i) + \nabla I_0(\mathbf{x}_i) \cdot \Delta \mathbf{u} - I_0(\mathbf{x}_i)]^2 \quad (7.4)$$

$$= \sum_i w(\mathbf{x}_i) [\nabla I_0(\mathbf{x}_i) \cdot \Delta \mathbf{u}]^2 \quad (7.5)$$

$$= \Delta \mathbf{u}^T \mathbf{A} \Delta \mathbf{u}, \quad (7.6)$$

where

$$\nabla I_0(\mathbf{x}_i) = \left(\frac{\partial I_0}{\partial x}, \frac{\partial I_0}{\partial y} \right)(\mathbf{x}_i) \quad (7.7)$$

is the *image gradient* at \mathbf{x}_i . This gradient can be computed using a variety of techniques (Schmid, Mohr, and Bauckhage 2000). The classic “Harris” detector (Harris and Stephens 1988) uses a $[-2 -1 0 1 2]$ filter, but more modern variants (Schmid, Mohr, and Bauckhage 2000; Triggs 2004) convolve the image with horizontal and vertical derivatives of a Gaussian (typically with $\sigma = 1$).

The auto-correlation matrix \mathbf{A} can be written as

$$\mathbf{A} = w * \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix}, \quad (7.8)$$

where we have replaced the weighted summations with discrete convolutions with the weighting kernel w . This matrix can be interpreted as a tensor (multiband) image, where the outer products of the gradients ∇I are convolved with a weighting function w to provide a per-pixel estimate of the local (quadratic) shape of the auto-correlation function.

¹Strictly speaking, a correlation is the *product* of two patches (3.12); I’m using the term here in a more qualitative sense. The weighted sum of squared differences is often called an *SSD surface* (Section 9.1).

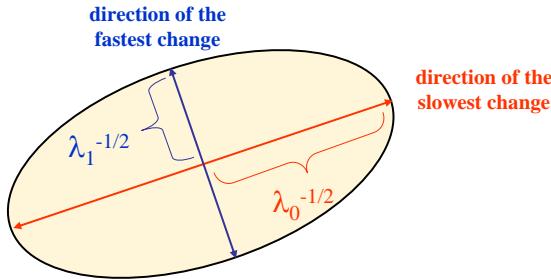


Figure 7.6 Uncertainty ellipse corresponding to an eigenvalue analysis of the auto-correlation matrix \mathbf{A} .

As first shown by Anandan (1984; 1989) and further discussed in Section 9.1.3 and Equation (9.37), the inverse of the matrix \mathbf{A} provides a lower bound on the uncertainty in the location of a matching patch. It is therefore a useful indicator of which patches can be reliably matched. The easiest way to visualize and reason about this uncertainty is to perform an eigenvalue analysis of the auto-correlation matrix \mathbf{A} , which produces two eigenvalues (λ_0, λ_1) and two eigenvector directions (Figure 7.6). Since the larger uncertainty depends on the smaller eigenvalue, i.e., $\lambda_0^{-1/2}$, it makes sense to find maxima in the smaller eigenvalue to locate good features to track (Shi and Tomasi 1994).

Förstner–Harris. While Anandan (1984) and Lucas and Kanade (1981) were the first to analyze the uncertainty structure of the auto-correlation matrix, they did so in the context of associating certainties with optical flow measurements. Förstner (1986) and Harris and Stephens (1988) were the first to propose using local maxima in rotationally invariant scalar measures derived from the auto-correlation matrix to locate keypoints for the purpose of sparse feature matching.² Both of these techniques also proposed using a Gaussian weighting window instead of the previously used square patches, which makes the detector response insensitive to in-plane image rotations.

The minimum eigenvalue λ_0 (Shi and Tomasi 1994) is not the only quantity that can be used to find keypoints. A simpler quantity, proposed by Harris and Stephens (1988), is

$$\det(\mathbf{A}) - \alpha \operatorname{trace}(\mathbf{A})^2 = \lambda_0 \lambda_1 - \alpha(\lambda_0 + \lambda_1)^2 \quad (7.9)$$

with $\alpha = 0.06$. Unlike eigenvalue analysis, this quantity does not require the use of square roots and yet is still rotationally invariant and also downweights edge-like features where

²Schmid, Mohr, and Bauckhage (2000) and Triggs (2004) give more detailed historical reviews of feature detection algorithms.

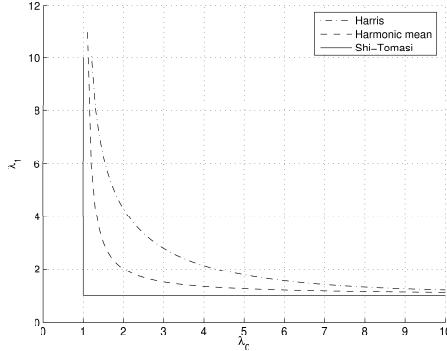


Figure 7.7 Isocontours of popular keypoint detection functions (Brown, Szeliski, and Winder 2004). Each detector looks for points where the eigenvalues λ_0, λ_1 of $\mathbf{A} = w * \nabla I \nabla I^T$ are both large.

$\lambda_1 \gg \lambda_0$. Triggs (2004) suggests using the quantity

$$\lambda_0 - \alpha \lambda_1 \quad (7.10)$$

(say, with $\alpha = 0.05$), which also reduces the response at 1D edges, where aliasing errors sometimes inflate the smaller eigenvalue. He also shows how the basic 2×2 Hessian can be extended to parametric motions to detect points that are also accurately localizable in scale and rotation. Brown, Szeliski, and Winder (2005), on the other hand, use the harmonic mean,

$$\frac{\det \mathbf{A}}{\text{tr } \mathbf{A}} = \frac{\lambda_0 \lambda_1}{\lambda_0 + \lambda_1}, \quad (7.11)$$

which is a smoother function in the region where $\lambda_0 \approx \lambda_1$. Figure 7.7 shows isocontours of the various interest point operators, from which we can see how the two eigenvalues are blended to determine the final interest value. Figure 7.8 shows the resulting interest operator responses for the classic Harris detector as well as the difference of Gaussian (DoG) detector discussed below.

Adaptive non-maximal suppression (ANMS). While most feature detectors simply look for local maxima in the interest function, this can lead to an uneven distribution of feature points across the image, e.g., points will be denser in regions of higher contrast. To mitigate this problem, Brown, Szeliski, and Winder (2005) only detect features that are both local maxima and whose response value is significantly (10%) greater than that of all of its neighbors within a radius r (Figure 7.9c–d). They devise an efficient way to associate suppression radii with all local maxima by first sorting them by their response strength and then creating

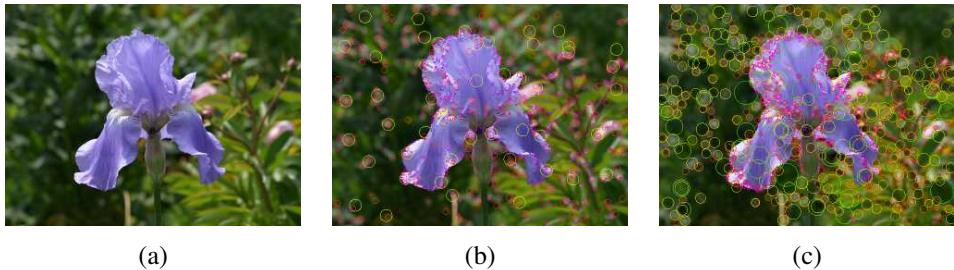


Figure 7.8 Interest operator responses: (a) Sample image, (b) Harris response, and (c) DoG response. The circle sizes and colors indicate the scale at which each interest point was detected. Notice how the two detectors tend to respond at complementary locations.

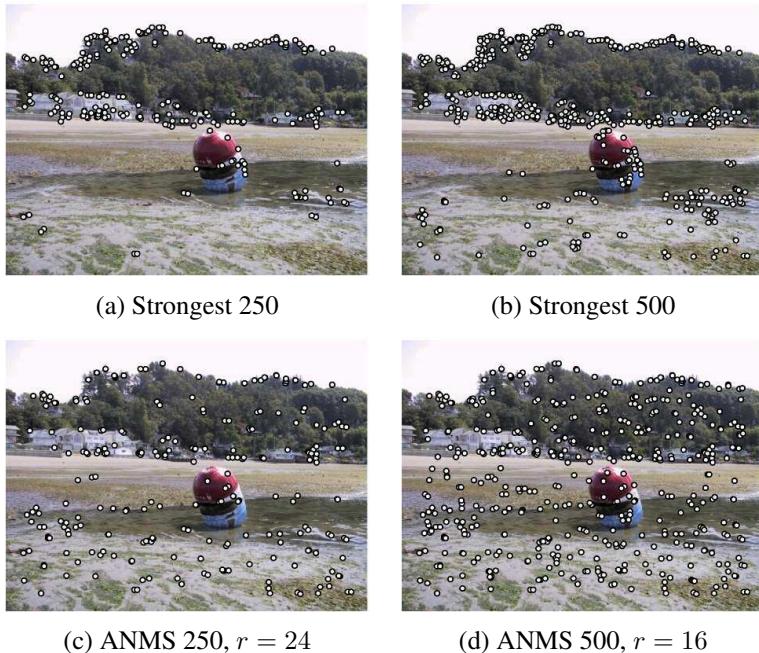


Figure 7.9 Adaptive non-maximal suppression (ANMS) (Brown, Szeliski, and Winder 2005) © 2005 IEEE: The upper two images show the strongest 250 and 500 interest points, while the lower two images show the interest points selected with adaptive non-maximal suppression, along with the corresponding suppression radius r . Note how the latter features have a much more uniform spatial distribution across the image.

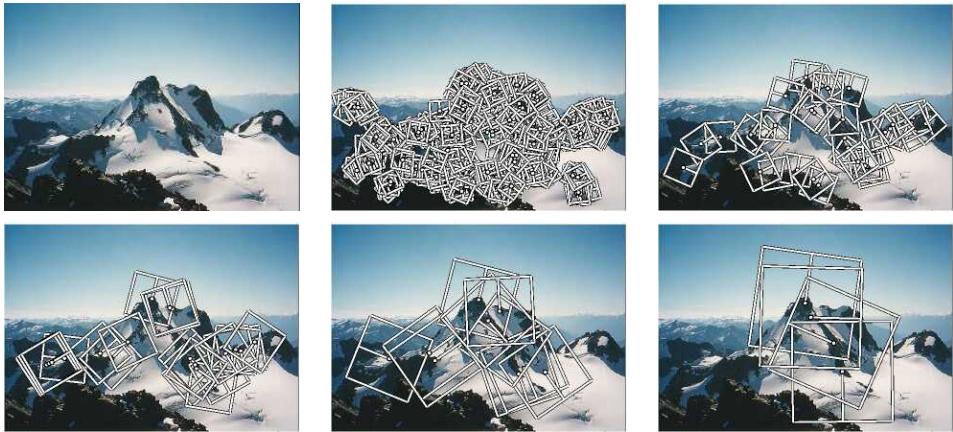


Figure 7.10 Multi-scale oriented patches (MOPS) extracted at five pyramid levels (Brown, Szeliski, and Winder 2005) © 2005 IEEE. The boxes show the feature orientation and the region from which the descriptor vectors are sampled.

a second list sorted by decreasing suppression radius (Brown, Szeliski, and Winder 2005). Figure 7.9 shows a qualitative comparison of selecting the top n features and using ANMS. Note that non-maximal suppression is now also an essential component of DNN-based object detectors, as discussed in Section 6.3.3.

Measuring repeatability. Given the large number of feature detectors that have been developed in computer vision, how can we decide which ones to use? Schmid, Mohr, and Bauckhage (2000) were the first to propose measuring the *repeatability* of feature detectors, which they define as the frequency with which keypoints detected in one image are found within ϵ (say, $\epsilon = 1.5$) pixels of the corresponding location in a transformed image. In their paper, they transform their planar images by applying rotations, scale changes, illumination changes, viewpoint changes, and adding noise. They also measure the *information content* available at each detected feature point, which they define as the entropy of a set of rotationally invariant local grayscale descriptors. Among the techniques they survey, they find that the improved (Gaussian derivative) version of the Harris operator with $\sigma_d = 1$ (scale of the derivative Gaussian) and $\sigma_i = 2$ (scale of the integration Gaussian) works best.

Scale invariance

In many situations, detecting features at the finest stable scale possible may not be appropriate. For example, when matching images with little high-frequency detail (e.g., clouds),

fine-scale features may not exist.

One solution to the problem is to extract features at a variety of scales, e.g., by performing the same operations at multiple resolutions in a pyramid and then matching features at the same level. This kind of approach is suitable when the images being matched do not undergo large scale changes, e.g., when matching successive aerial images taken from an airplane or stitching panoramas taken with a fixed-focal-length camera. Figure 7.10 shows the output of one such approach: the multi-scale oriented patch detector of Brown, Szeliski, and Winder (2005), for which responses at five different scales are shown.

However, for most object recognition applications, the scale of the object in the image is unknown. Instead of extracting features at many different scales and then matching all of them, it is more efficient to extract features that are stable in both location *and* scale (Lowe 2004; Mikolajczyk and Schmid 2004).

Early investigations into scale selection were performed by Lindeberg (1993; 1998b), who first proposed using extrema in the Laplacian of Gaussian (LoG) function as interest point locations. Based on this work, Lowe (2004) proposed computing a set of sub-octave Difference of Gaussian filters (Figure 7.11a), looking for 3D (space+scale) maxima in the resulting structure (Figure 7.11b), and then computing a sub-pixel space+scale location using a quadratic fit (Brown and Lowe 2002). The number of sub-octave levels was determined, after careful empirical investigation, to be three, which corresponds to a quarter-octave pyramid, which is the same as used by Triggs (2004).

As with the Harris operator, pixels where there is strong asymmetry in the local curvature of the indicator function (in this case, the DoG) are rejected. This is implemented by first computing the local Hessian of the difference image D ,

$$\mathbf{H} = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{bmatrix}, \quad (7.12)$$

and then rejecting keypoints for which

$$\frac{\text{Tr}(\mathbf{H})^2}{\text{Det}(\mathbf{H})} > 10. \quad (7.13)$$

While Lowe's Scale Invariant Feature Transform (SIFT) performs well in practice, it is not based on the same theoretical foundation of maximum spatial stability as the auto-correlation-based detectors. (In fact, its detection locations are often complementary to those produced by such techniques and can therefore be used in conjunction with these other approaches.) In order to add a scale selection mechanism to the Harris corner detector, Mikolajczyk and Schmid (2004) evaluate the Laplacian of Gaussian function at each detected Harris point (in a multi-scale pyramid) and keep only those points for which the Laplacian is extremal (larger

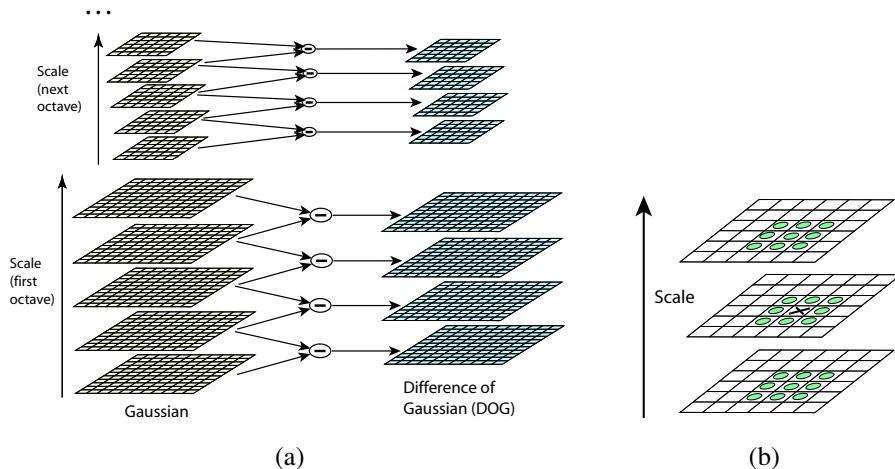


Figure 7.11 Scale-space feature detection using a sub-octave Difference of Gaussian pyramid (Lowe 2004) © 2004 Springer: (a) Adjacent levels of a sub-octave Gaussian pyramid are subtracted to produce Difference of Gaussian images; (b) extrema (maxima and minima) in the resulting 3D volume are detected by comparing a pixel to its 26 neighbors.

or smaller than both its coarser and finer-level values). An optional iterative refinement for both scale and position is also proposed and evaluated. Additional examples of scale-invariant region detectors are discussed by Mikolajczyk, Tuytelaars *et al.* (2005) and Tuytelaars and Mikolajczyk (2008).

Rotational invariance and orientation estimation

In addition to dealing with scale changes, most image matching and object recognition algorithms need to deal with (at least) in-plane image rotation. One way to deal with this problem is to design descriptors that are rotationally invariant (Schmid and Mohr 1997), but such descriptors have poor discriminability, i.e. they map different looking patches to the same descriptor.

A better method is to estimate a *dominant orientation* at each detected keypoint. Once the local orientation and scale of a keypoint have been estimated, a scaled and oriented patch around the detected point can be extracted and used to form a feature descriptor (Figures 7.10 and 7.15).

The simplest possible orientation estimate is the average gradient within a region around the keypoint. If a Gaussian weighting function is used (Brown, Szeliski, and Winder 2005),

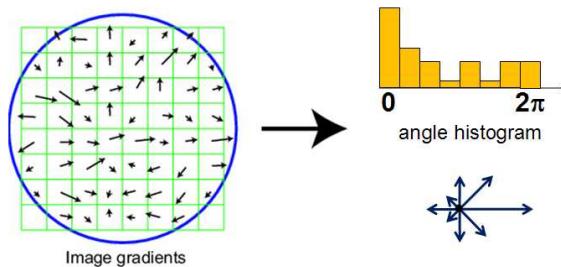


Figure 7.12 A dominant orientation estimate can be computed by creating a histogram of all the gradient orientations (weighted by their magnitudes or after thresholding out small gradients) and then finding the significant peaks in this distribution (Lowe 2004) © 2004 Springer.

this average gradient is equivalent to a first-order steerable filter (Section 3.2.3), i.e., it can be computed using an image convolution with the horizontal and vertical derivatives of Gaussian filter (Freeman and Adelson 1991). To make this estimate more reliable, it is usually preferable to use a larger aggregation window (Gaussian kernel size) than detection window (Brown, Szeliski, and Winder 2005). The orientations of the square boxes shown in Figure 7.10 were computed using this technique.

Sometimes, however, the averaged (signed) gradient in a region can be small and therefore an unreliable indicator of orientation. A more reliable technique is to look at the *histogram* of orientations computed around the keypoint. Lowe (2004) computes a 36-bin histogram of edge orientations weighted by both gradient magnitude and Gaussian distance to the center, finds all peaks within 80% of the global maximum, and then computes a more accurate orientation estimate using a three-bin parabolic fit (Figure 7.12).

Affine invariance

While scale and rotation invariance are highly desirable, for many applications such as *wide baseline stereo matching* (Pritchett and Zisserman 1998; Schaffalitzky and Zisserman 2002) or location recognition (Chum, Philbin *et al.* 2007), full affine invariance is preferred. Affine-invariant detectors not only respond at consistent locations after scale and orientation changes, they also respond consistently across affine deformations such as (local) perspective foreshortening (Figure 7.13). In fact, for a small enough patch, any continuous image warping can be well approximated by an affine deformation.

To introduce affine invariance, several authors have proposed fitting an ellipse to the auto-

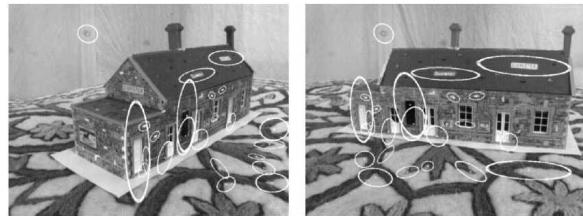


Figure 7.13 Affine region detectors used to match two images taken from dramatically different viewpoints (Mikolajczyk and Schmid 2004) © 2004 Springer.



Figure 7.14 Maximally stable extremal regions (MSERs) extracted and matched from a number of images (Matas, Chum et al. 2004) © 2004 Elsevier.

correlation or Hessian matrix (using eigenvalue analysis) and then using the principal axes and ratios of this fit as the affine coordinate frame (Lindeberg and Gårding 1997; Baumberg 2000; Mikolajczyk and Schmid 2004; Mikolajczyk, Tuytelaars *et al.* 2005; Tuytelaars and Mikolajczyk 2008).

Another important affine invariant region detector is the maximally stable extremal region (MSER) detector developed by Matas, Chum *et al.* (2004). To detect MSERs, binary regions are computed by thresholding the image at all possible gray levels (the technique therefore only works for grayscale images). This operation can be performed efficiently by first sorting all pixels by gray value and then incrementally adding pixels to each connected component as the threshold is changed (Nistér and Stewénius 2008). As the threshold is changed, the area of each component (region) is monitored; regions whose rate of change of area with respect to the threshold is minimal are defined as *maximally stable*. Such regions are therefore invariant to both affine geometric and photometric (linear bias-gain or smooth monotonic) transformations (Figure 7.14). If desired, an affine coordinate frame can be fit to each detected region using its moment matrix.

The area of feature point detection continues to be very active, with papers appearing every year at major computer vision conferences. Mikolajczyk, Tuytelaars *et al.* (2005) and Tuytelaars and Mikolajczyk (2008) survey a number of popular (pre-DNN) affine region detectors and provide experimental comparisons of their invariance to common image transfor-

mations such as scaling, rotations, noise, and blur.

More recent papers published in the last decade include:

- SURF (Bay, Ess *et al.* 2008), which uses integral images for faster convolutions;
- FAST and FASTER (Rosten, Porter, and Drummond 2010), one of the first learned detectors;
- BRISK (Leutenegger, Chli, and Siegwart 2011), which uses a scale-space FAST detector together with a bit-string descriptor;
- ORB (Rublee, Rabaud *et al.* 2011), which adds orientation to FAST; and
- KAZE (Alcantarilla, Bartoli, and Davison 2012) and Accelerated-KAZE (Alcantarilla, Nuevo, and Bartoli 2013), which use non-linear diffusion to select the scale for feature detection.

While FAST introduced the idea of machine learning for feature detectors, more recent papers use convolutional neural networks to perform the detection. These include:

- Learning covariant feature detectors (Lenc and Vedaldi 2016);
- Learning to assign orientations to feature points (Yi, Verdie *et al.* 2016);
- LIFT, learned invariant feature transforms (Yi, Trulls *et al.* 2016), SuperPoint, self-supervised interest point detection and description (DeTone, Malisiewicz, and Rabivich 2018), and LF-Net, learning local features from images (Ono, Trulls *et al.* 2018), all three of which jointly optimize the detectors and descriptors in a single (multi-head) pipeline;
- AffNet (Mishkin, Radenovic, and Matas 2018), which detects matchable affine-covariant regions;
- Key.Net (Barroso-Laguna, Riba *et al.* 2019), which uses a combination of handcrafted and learned CNN features; and
- D2-Net (Dusmanu, Rocco *et al.* 2019), R2D2 (Revaud, Weinzaepfel *et al.* 2019), and D2D (Tian, Balntas *et al.* 2020), which all extract dense local feature descriptors and then keeps the ones that have high saliency or repeatability.

These last two papers also contains a nice review of other recent feature detectors, as does the paper by Balntas, Lenc *et al.* (2020).

Of course, keypoints are not the only features that can be used for registering images. Zoghami, Faugeras, and Deriche (1997) use line segments as well as point-like features to estimate homographies between pairs of images, whereas Bartoli, Coquerelle, and Sturm (2004) use line segments with local correspondences along the edges to extract 3D structure and motion. Tuytelaars and Van Gool (2004) use affine invariant regions to detect correspondences for wide baseline stereo matching, whereas Kadir, Zisserman, and Brady (2004) detect salient regions where patch entropy and its rate of change with scale are locally maximal. Corso and Hager (2005) use a related technique to fit 2D oriented Gaussian kernels to homogeneous regions. More details on techniques for finding and matching curves, lines, and regions can be found later in this chapter.

7.1.2 Feature descriptors

After detecting keypoint features, we must *match* them, i.e., we must determine which features come from corresponding locations in different images. In some situations, e.g., for video sequences (Shi and Tomasi 1994) or for stereo pairs that have been *rectified* (Zhang, Deriche *et al.* 1995; Loop and Zhang 1999; Scharstein and Szeliski 2002), the local motion around each feature point may be mostly translational. In this case, simple error metrics, such as the *sum of squared differences* or *normalized cross-correlation*, described in Section 9.1, can be used to directly compare the intensities in small patches around each feature point. (The comparative study by Mikolajczyk and Schmid (2005), discussed below, uses cross-correlation.) Because feature points may not be exactly located, a more accurate matching score can be computed by performing incremental motion refinement as described in Section 9.1.3, but this can be time-consuming and can sometimes even decrease performance (Brown, Szeliski, and Winder 2005).

In most cases, however, the local appearance of features will change in orientation and scale, and sometimes even undergo affine deformations. Extracting a local scale, orientation, or affine frame estimate and then using this to resample the patch before forming the feature descriptor is thus usually preferable (Figure 7.15).

Even after compensating for these changes, the local appearance of image patches will usually still vary from image to image. How can we make image descriptors more invariant to such changes, while still preserving discriminability between different (non-corresponding) patches? Mikolajczyk and Schmid (2005) review a number of view-invariant local image descriptors and experimentally compare their performance. More recently, Balntas, Lenc *et al.* (2020) and Jin, Mishkin *et al.* (2021) compare the large number of learned feature descriptors developed in the prior decade.³ Below, we describe a few of these descriptors in

³Many recent publications such as Tian, Yu *et al.* (2019) use their HPatches dataset to compare their performance



Figure 7.15 Once a local scale and orientation estimate has been determined, MOPS descriptors are formed using an 8×8 sampling of bias and gain normalized intensity values, with a sample spacing of five pixels relative to the detection scale (Brown, Szeliski, and Winder 2005) © 2005 IEEE. This low frequency sampling gives the features some robustness to interest point location error and is achieved by sampling at a higher pyramid level than the detection scale.

more detail.

Bias and gain normalization (MOPS). For tasks that do not exhibit large amounts of foreshortening, such as image stitching, simple normalized intensity patches perform reasonably well and are simple to implement (Brown, Szeliski, and Winder 2005) (Figure 7.15). To compensate for slight inaccuracies in the feature point detector (location, orientation, and scale), multi-scale oriented patches (MOPS) are sampled at a spacing of five pixels relative to the detection scale, using a coarser level of the image pyramid to avoid aliasing. To compensate for affine photometric variations (linear exposure changes or bias and gain, (3.3)), patch intensities are re-scaled so that their mean is zero and their variance is one.

Scale invariant feature transform (SIFT). SIFT features (Lowe 2004) are formed by computing the gradient at each pixel in a 16×16 window around the detected keypoint, using the appropriate level of the Gaussian pyramid at which the keypoint was detected. The gradient magnitudes are downweighted by a Gaussian fall-off function (shown as a blue circle in Figure 7.16a) to reduce the influence of gradients far from the center, as these are more affected by small misregistrations.

In each 4×4 quadrant, a gradient orientation histogram is formed by (conceptually) adding the gradient values weighted by the Gaussian fall-off function to one of eight orientation histogram bins. To reduce the effects of location and dominant orientation miscalibration, each of the original 256 weighted gradient magnitudes is softly added to $2 \times 2 \times 2$ adjacent

against previous approaches.

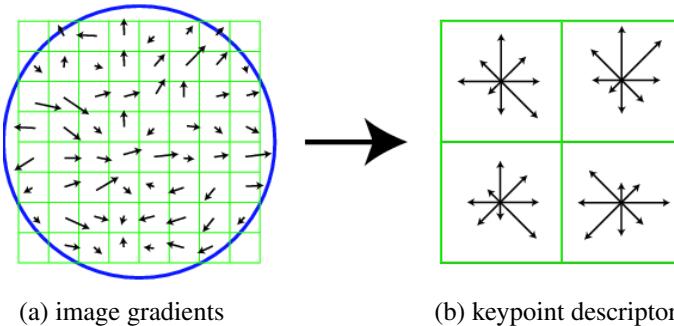


Figure 7.16 A schematic representation of Lowe’s (2004) scale invariant feature transform (SIFT): (a) Gradient orientations and magnitudes are computed at each pixel and weighted by a Gaussian fall-off function (blue circle). (b) A weighted gradient orientation histogram is then computed in each subregion, using trilinear interpolation. While this figure shows an 8×8 pixel patch and a 2×2 descriptor array, Lowe’s actual implementation uses 16×16 patches and a 4×4 array of eight-bin histograms.

histogram bins in the (x, y, θ) space using trilinear interpolation. Softly distributing values to adjacent histogram bins is generally a good idea in any application where histograms are being computed, e.g., for Hough transforms (Section 7.4.2) or local histogram equalization (Section 3.1.4).

The 4×4 array of eight-bin histogram yields 128 non-negative values form a raw version of the SIFT descriptor vector. To reduce the effects of contrast or gain (additive variations are already removed by the gradient), the 128-D vector is normalized to unit length. To further make the descriptor robust to other photometric variations, values are clipped to 0.2 and the resulting vector is once again renormalized to unit length.

PCA-SIFT. Ke and Sukthankar (2004) propose a simpler way to compute descriptors inspired by SIFT; it computes the x and y (gradient) derivatives over a 39×39 patch and then reduces the resulting 3042-dimensional vector to 36 using principal component analysis (PCA) (Section 5.2.3 and Appendix A.1.2). Another popular variant of SIFT is SURF (Bay, Ess *et al.* 2008), which uses box filters to approximate the derivatives and integrals used in SIFT.

RootSIFT. Arandjelović and Zisserman (2012) observe that by simply re-normalizing SIFT descriptors using an L_1 measure and then taking the square root of each component, a dramatic increase in performance (discriminability) can be obtained.

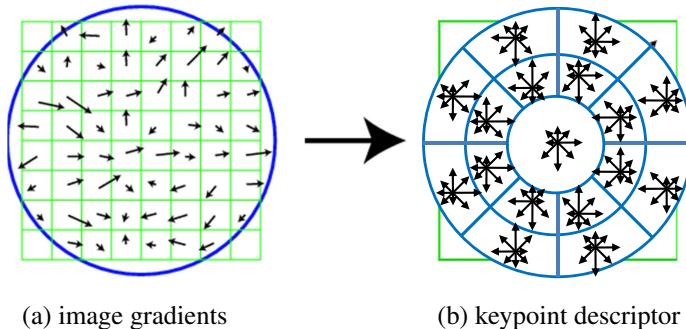


Figure 7.17 *The gradient location-orientation histogram (GLOH) descriptor uses log-polar bins instead of square bins to compute orientation histograms (Mikolajczyk and Schmid 2005). GLOH uses 16 gradient orientations inside each bin, although this figure only shows 8 to appear less cluttered.*

Gradient location-orientation histogram (GLOH). This descriptor, developed by Mikolajczyk and Schmid (2005), is a variant of SIFT that uses a log-polar binning structure instead of the four quadrants used by Lowe (2004) (Figure 7.17). The spatial bins extend over the radii $0\dots 6$, $6\dots 11$, and $11\dots 15$, with eight angular bins (except for the single central region), for a total of 17 spatial bins and GLOH uses 16 orientation bins instead of the 8 used in SIFT. The 272-dimensional histogram is then projected onto a 128-dimensional descriptor using PCA trained on a large database. In their evaluation, Mikolajczyk and Schmid (2005) found that GLOH, which has the best performance overall, outperforms SIFT by a small margin.

Steerable filters. Steerable filters (Section 3.2.3) are combinations of derivative of Gaussian filters that permit the rapid computation of even and odd (symmetric and anti-symmetric) edge-like and corner-like features at all possible orientations (Freeman and Adelson 1991). Because they use reasonably broad Gaussians, they too are somewhat insensitive to localization and orientation errors.

Performance of local descriptors. Among the local descriptors that Mikolajczyk and Schmid (2005) compared, they found that GLOH performed best, followed closely by SIFT. They also present results for many other descriptors not covered in this book.

The field of feature descriptors continued to advance rapidly, with some techniques looking at local color information (van de Weijer and Schmid 2006; Abdel-Hakim and Farag 2006). Winder and Brown (2007) develop a multi-stage framework for feature descriptor computation that subsumes both SIFT and GLOH (Figure 7.18a) and also allows them to

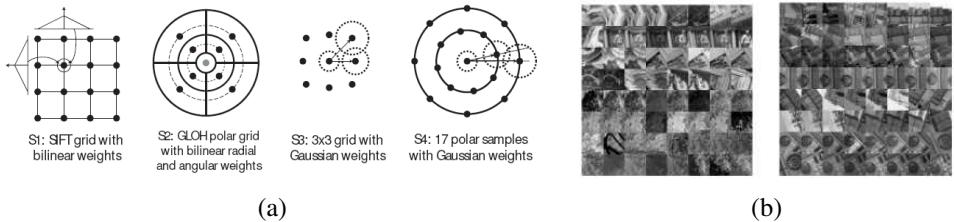


Figure 7.18 *Spatial summation blocks for SIFT, GLOH, and some related feature descriptors (Winder and Brown 2007) © 2007 IEEE: (a) The parameters for the features, e.g., their Gaussian weights, are learned from a training database of (b) matched real-world image patches obtained from robust structure from motion applied to internet photo collections (Hua, Brown, and Winder 2007).*

learn optimal parameters for newer descriptors that outperform previous hand-tuned descriptors. Hua, Brown, and Winder (2007) extend this work by learning lower-dimensional projections of higher-dimensional descriptors that have the best discriminative power, and Brown, Hua, and Winder (2011) further extend it by learning the optimal placement of the pooling regions. All of these papers use a database of real-world image patches (Figure 7.18b) obtained by sampling images at locations that were reliably matched using a robust structure-from-motion algorithm applied to internet photo collections (Snavely, Seitz, and Szeliski 2006; Goesele, Snavely *et al.* 2007). In concurrent work, Tola, Lepetit, and Fua (2010) developed a similar DAISY descriptor for dense stereo matching and optimized its parameters based on ground truth stereo data.

While these techniques construct feature detectors that optimize for repeatability across *all* object classes, it is also possible to develop class- or instance-specific feature detectors that maximize *discriminability* from other classes (Ferencz, Learned-Miller, and Malik 2008). If planar surface orientations can be determined in the images being matched, it is also possible to extract viewpoint-invariant patches (Wu, Clipp *et al.* 2008).

A more recent trend has been the development of binary bit-string feature descriptors, which can take advantage of fast Hamming distance operators in modern computer architectures. The BRIEF descriptor (Calonder, Lepetit *et al.* 2010) compares 128 different pairs of pixel values (denoted as line segments in Figure 7.19a) scattered around the keypoint location to obtain a 128-bit vector. ORB (Rublee, Rabaud *et al.* 2011) adds an orientation component to the FAST detector before computing oriented BRIEF descriptors. BRISK (Leutenegger, Chli, and Siegwart 2011) adds scale-space analysis to the FAST detector and a radially symmetric sampling pattern (Figure 7.19b) to produce the binary descriptor. FREAK (Alahi,

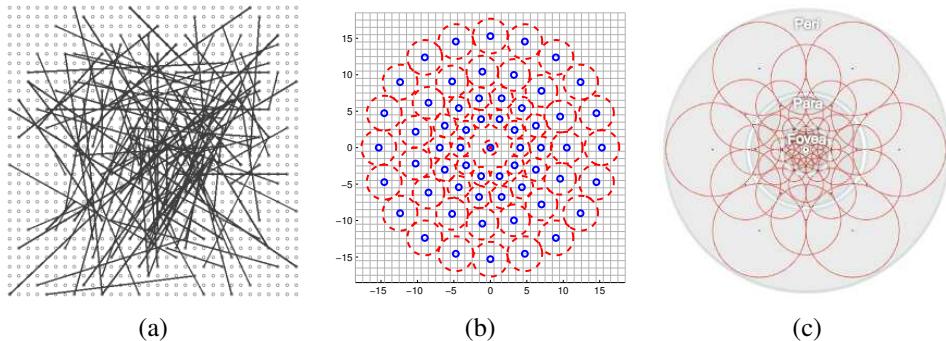


Figure 7.19 *Binary bit-string feature descriptors: (a) the BRIEF descriptor compares 128 pairs of pixel values (denoted by line segments) and stores the comparison results in a 128-bit vector (Calonder, Lepetit et al. 2010) © 2010 Springer; (b) BRISK sampling pattern and Gaussian blur radii; (Leutenegger, Chli, and Siegwart 2011) © 2011 IEEE; (c) FREAK retinal sampling pattern (Alahi, Ortiz, and Vanderghenst 2012) © 2012 IEEE.*

Ortiz, and Vanderghenst 2012) uses a more pronounced “retinal” (log-polar) sampling pattern paired with a cascade of bit comparisons for even greater speed and efficiency. The survey and evaluation by Mukherjee, Wu, and Wang (2015) compares all of these “classic” feature detectors and descriptors.

Since 2015 or so, most of the new feature descriptors are constructed using deep learning techniques, as surveyed in Balntas, Lenc *et al.* (2020) and Jin, Mishkin *et al.* (2021). Some of these descriptors, such as LIFT (Yi, Trulls *et al.* 2016), TFeat (Balntas, Riba *et al.* 2016), HPatches (Balntas, Lenc *et al.* 2020), L2-Net (Tian, Fan, and Wu 2017), HardNet (Mishchuk, Mishkin *et al.* 2017), Geodesc (Luo, Shen *et al.* 2018), LF-Net (Ono, Trulls *et al.* 2018), SOSNet (Tian, Yu *et al.* 2019), and Key.Net (Barroso-Laguna, Riba *et al.* 2019) operate on patches, much like the classical SIFT approach. They hence require an initial local feature detector to determine the center of the patch and use a predetermined patch size when constructing the input to the network.

In contrast, approaches such as DELF (Noh, Araujo *et al.* 2017), SuperPoint (DeTone, Malisiewicz, and Rabinovich 2018), D2-Net (Dusmanu, Rocco *et al.* 2019), ContextDesc (Luo, Shen *et al.* 2019), R2D2 (Revaud, Weinzaepfel *et al.* 2019), ASLFeat (Luo, Zhou *et al.* 2020), and CAPS (Wang, Zhou *et al.* 2020) use the entire image as the input to the descriptor computation. This has the added benefit that the receptive field used to compute the descriptor can be learned from the data and does not require specifying a patch size. Theoretically, these CNN models can learn receptive fields that use all of the pixels in the image, although in

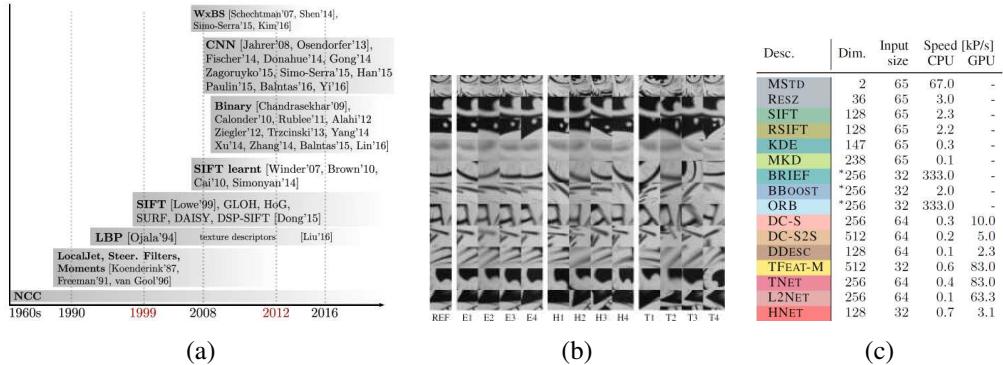


Figure 7.20 HPatches local descriptors benchmark (Balntas, Lenc et al. 2020) © 2019 IEEE: (a) chronology of feature descriptors; (b) typical patches in the dataset (grouped by Easy, Hard, and Tough); (c) size and speed of different descriptors.

practice they tend to use Gaussian-like receptive fields (Zhou, Khosla et al. 2015; Luo, Li et al. 2016; Selvaraju, Cogswell et al. 2017).

In the HPatches benchmark (Figure 7.20) for evaluating patch matching by Balntas, Lenc et al. (2020), HardNet and L2-net performed the best on average. Another paper (Wang, Zhou et al. 2020) shows CAPS and R2D2 as the best performers, while S2DNet (Germain, Bourmaud, and Lepetit 2020) and LISRD (Pautrat, Larsson et al. 2020) also claim state-of-the-art performance, while the WISW benchmark (Bellavia and Colombo 2020) shows that traditional descriptors such as SIFT enhanced with more recent ideas do the best. On the wide baseline image matching benchmark by Jin, Mishkin et al. (2021),⁴ HardNet, KeyNet, and D2-Net were top performers (e.g., D2-Net had the highest number of landmarks), although the results were quite task-dependent and the Difference of Gaussian detector was still the best. The performance of these descriptors on matching features across large illumination differences (day-night) has also been studied (Radenović, Schönberger et al. 2016; Zhou, Sattler, and Jacobs 2016; Mishkin 2021).

The most recent trend in wide-baseline matching has been to densely extract features without a detector stage and to then match and refine the set of correspondences (Jiang, Trulls et al. 2021; Sarlin, Unagar et al. 2021; Sun, Shen et al. 2021; Truong, Danelljan et al. 2021; Zhou, Sattler, and Leal-Taixé 2021). Some of these more recent techniques have been evaluated by Mishkin (2021).

⁴The benchmark is associated with the CVPR Workshop on Image Matching: Local Features & Beyond: <https://image-matching-workshop.github.io>.

7.1.3 Feature matching

Once we have extracted features and their descriptors from two or more images, the next step is to establish some preliminary feature matches between these images. The approach we take depends partially on the application, e.g., different strategies may be preferable for matching images that are known to overlap (e.g., in image stitching) vs. images that may have no correspondence whatsoever (e.g., when trying to recognize objects from a database).

In this section, we divide this problem into two separate components. The first is to select a *matching strategy*, which determines which correspondences are passed on to the next stage for further processing. The second is to devise efficient *data structures* and *algorithms* to perform this matching as quickly as possible, which we expand on in Section 7.1.4.

Matching strategy and error rates

Determining which feature matches are reasonable to process further depends on the context in which the matching is being performed. Say we are given two images that overlap to a fair amount (e.g., for image stitching or for tracking objects in a video). We know that most features in one image are likely to match the other image, although some may not match because they are occluded or their appearance has changed too much.

On the other hand, if we are trying to recognize how many known objects appear in a cluttered scene (Figure 6.2), most of the features may not match. Furthermore, a large number of potentially matching objects must be searched, which requires more efficient strategies, as described below.

To begin with, we assume that the feature descriptors have been designed so that Euclidean (vector magnitude) distances in feature space can be directly used for ranking potential matches. If it turns out that certain parameters (axes) in a descriptor are more reliable than others, it is usually preferable to re-scale these axes ahead of time, e.g., by determining how much they vary when compared against other known good matches (Hua, Brown, and Winder 2007). A more general process, which involves transforming feature vectors into a new scaled basis, is called *whitening* and is discussed in more detail in the context of eigenface-based face recognition (Section 5.2.3).

Given a Euclidean distance metric, the simplest matching strategy is to set a threshold (maximum distance) and to return all matches from other images within this threshold. Setting the threshold too high results in too many *false positives*, i.e., incorrect matches being returned. Setting the threshold too low results in too many *false negatives*, i.e., too many correct matches being missed (Figure 7.21).

We can quantify the performance of a matching algorithm at a particular threshold by

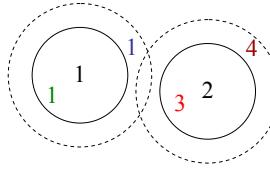


Figure 7.21 False positives and negatives: The black digits 1 and 2 are features being matched against a database of features in other images. At the current threshold setting (the solid circles), the green 1 is a true positive (good match), the blue 1 is a false negative (failure to match), and the red 3 is a false positive (incorrect match). If we set the threshold higher (the dashed circles), the blue 1 becomes a true positive but the brown 4 becomes an additional false positive.

| | True matches | True non-matches | |
|-----------------------|--------------|------------------|-------------|
| Predicted matches | TP = 18 | FP = 4 | P' = 22 |
| Predicted non-matches | FN = 2 | TN = 76 | N' = 78 |
| | P = 20 | N = 80 | Total = 100 |
| | TPR = 0.90 | FPR = 0.05 | PPV = 0.82 |
| | | | ACC = 0.94 |

Table 7.1 The number of matches correctly and incorrectly estimated by a feature matching algorithm, showing the number of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). The columns sum up to the actual number of positives (P) and negatives (N), while the rows sum up to the predicted number of positives (P') and negatives (N'). The formulas for the true positive rate (TPR), the false positive rate (FPR), the positive predictive value (PPV), and the accuracy (ACC) are given in the text.

first counting the number of true and false matches and match failures, using the following definitions (Fawcett 2006), which we already discussed in Section 6.3.3:

- TP: true positives, i.e., number of correct matches;
- FN: false negatives, matches that were not correctly detected;
- FP: false positives, proposed matches that are incorrect;
- TN: true negatives, non-matches that were correctly rejected.

Table 7.1 shows a sample *confusion matrix* (contingency table) containing such numbers.

We can convert these numbers into *unit rates* by defining the following quantities (Fawcett 2006):

- true positive rate (**TPR**),

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{\text{TP}}{\text{P}}; \quad (7.14)$$

- false positive rate (**FPR**),

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} = \frac{\text{FP}}{\text{N}}; \quad (7.15)$$

- positive predictive value (**PPV**),

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{\text{TP}}{\text{P}'}; \quad (7.16)$$

- accuracy (**ACC**),

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{P} + \text{N}}. \quad (7.17)$$

In the *information retrieval* (or document retrieval) literature (Baeza-Yates and Ribeiro-Neto 1999; Manning, Raghavan, and Schütze 2008), the term *precision* (how many returned documents are relevant) is used instead of PPV and *recall* (what fraction of relevant documents was found) is used instead of TPR (see also Section 6.3.3). The precision and recall can be combined into a single measure called the *F-score*, which is their harmonic mean. This single measure is often used to rank vision algorithms (Knapsch, Park *et al.* 2017).

Any particular matching strategy (at a particular threshold or parameter setting) can be rated by the TPR and FPR numbers; ideally, the true positive rate will be close to 1 and the false positive rate close to 0. As we vary the matching threshold, we obtain a family of such points, which are collectively known as the *receiver operating characteristic (ROC) curve* (Fawcett 2006) (Figure 7.22a). The closer this curve lies to the upper left corner, i.e., the larger the area under the curve (AUC), the better its performance. Figure 7.22b shows how we can plot the number of matches and non-matches as a function of inter-feature distance d . These curves can then be used to plot an ROC curve (Exercise 7.3). The ROC curve can also be used to calculate the *mean average precision*, which is the average precision (PPV) as you vary the threshold to select the best results, then the two top results, etc. (see Section 6.3.3 and Figure 6.27).

The problem with using a fixed threshold is that it is difficult to set; the useful range of thresholds can vary a lot as we move to different parts of the feature space (Lowe 2004; Mikolajczyk and Schmid 2005). A better strategy in such cases is to simply match the *nearest neighbor* in feature space. Since some features may have no matches (e.g., they may be part of background clutter in object recognition or they may be occluded in the other image), a threshold is still used to reduce the number of false positives.

Ideally, this threshold itself will adapt to different regions of the feature space. If sufficient training data is available (Hua, Brown, and Winder 2007), it is sometimes possible to learn

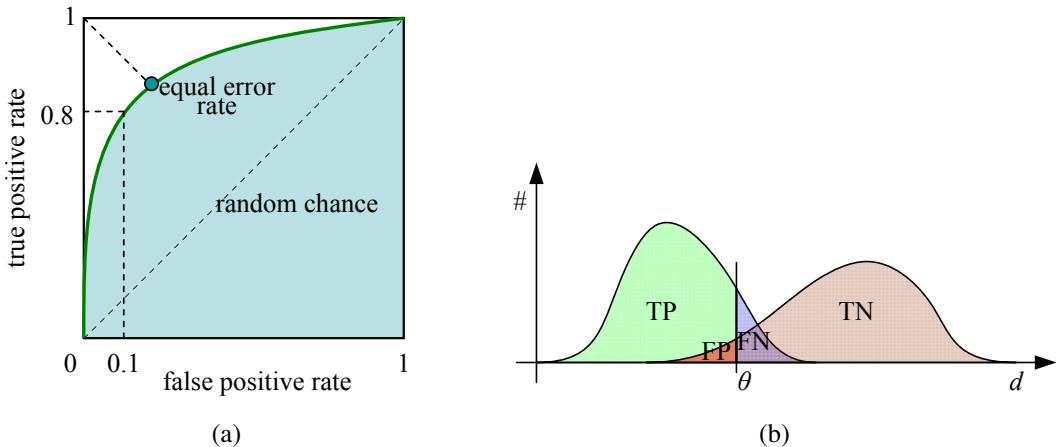


Figure 7.22 ROC curve and its related rates: (a) The ROC curve plots the true positive rate against the false positive rate for a particular combination of feature extraction and matching algorithms. Ideally, the true positive rate should be close to 1, while the false positive rate is close to 0. The area under the ROC curve (AUC) is often used as a single (scalar) measure of algorithm performance. Alternatively, the equal error rate is sometimes used. (b) The distribution of positives (matches) and negatives (non-matches) as a function of inter-feature distance d . As the threshold θ is increased, the number of true positives (TP) and false positives (FP) increases.

different thresholds for different features. Often, however, we are simply given a collection of images to match, e.g., when stitching images or constructing 3D models from unordered photo collections (Brown and Lowe 2007, 2005; Snavely, Seitz, and Szeliski 2006). In this case, a useful heuristic can be to compare the nearest neighbor distance to that of the second nearest neighbor, preferably taken from an image that is known not to match the target (e.g., a different object in the database) (Brown and Lowe 2002; Lowe 2004; Mishkin, Matas, and Perdoch 2015). We can define this *nearest neighbor distance ratio* (Mikolajczyk and Schmid 2005) as

$$\text{NNDR} = \frac{d_1}{d_2} = \frac{\|D_A - D_B\|}{\|D_A - D_C\|}, \quad (7.18)$$

where d_1 and d_2 are the nearest and second nearest neighbor distances, D_A is the target descriptor, and D_B and D_C are its closest two neighbors (Figure 7.23). Recent work has shown that mutual NNDR (or, at least NNDR with cross-consistency check) work noticeably better than one-way NNDR (Bellavia and Colombo 2020; Jin, Mishkin *et al.* 2021).

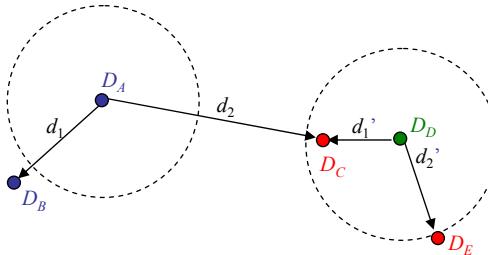


Figure 7.23 Fixed threshold, nearest neighbor, and nearest neighbor distance ratio matching. At a fixed distance threshold (dashed circles), descriptor D_A fails to match D_B and D_D incorrectly matches D_C and D_E . If we pick the nearest neighbor, D_A correctly matches D_B but D_D incorrectly matches D_C . Using nearest neighbor distance ratio (NNDR) matching, the small NNDR d_1/d_2 correctly matches D_A with D_B , and the large NNDR d'_1/d'_2 correctly rejects matches for D_D .

Efficient matching

Once we have decided on a matching strategy, we still need to efficiently search for potential candidates. The simplest way to find all corresponding feature points is to compare all features against all other features in each pair of potentially matching images. While traditionally this has been too computationally expensive, modern GPUs have enabled such comparisons.

A more efficient approach is to devise an *indexing structure*, such as a multi-dimensional search tree or a hash table, to rapidly search for features near a given feature. Such indexing structures can either be built for each image independently (which is useful if we want to only consider certain potential matches, e.g., searching for a particular object) or globally for all the images in a given database, which can potentially be faster, since it removes the need to iterate over each image. For extremely large databases (millions of images or more), even more efficient structures based on ideas from document retrieval, e.g., *vocabulary trees* (Nistér and Stewénius 2006), *product quantization* (Jégou, Douze, and Schmid 2010; Johnson, Douze, and Jégou 2021), or an *inverted multi-index* (Babenko and Lempitsky 2015b) can be used, as discussed in Section 7.1.4.

One of the simpler techniques to implement is multi-dimensional hashing, which maps descriptors into fixed size buckets based on some function applied to each descriptor vector. At matching time, each new feature is hashed into a bucket, and a search of nearby buckets is used to return potential candidates, which can then be sorted or graded to determine which are valid matches.

A simple example of hashing is the Haar wavelets used by Brown, Szeliski, and Winder

(2005) in their MOPS paper. During the matching structure construction, each 8×8 scaled, oriented, and normalized MOPS patch is converted into a three-element index by performing sums over different quadrants of the patch. The resulting three values are normalized by their expected standard deviations and then mapped to the two (of $b = 10$) nearest 1D bins. The three-dimensional indices formed by concatenating the three quantized values are used to index the $2^3 = 8$ bins where the feature is stored (added). At query time, only the primary (closest) indices are used, so only a single three-dimensional bin needs to be examined. The coefficients in the bin can then be used to select k approximate nearest neighbors for further processing (such as computing the NNDR).

A more complex, but more widely applicable, version of hashing is called *locality sensitive hashing*, which uses unions of independently computed hashing functions to index the features (Gionis, Indyk, and Motwani 1999; Shakhnarovich, Darrell, and Indyk 2006). Shakhnarovich, Viola, and Darrell (2003) extend this technique to be more sensitive to the distribution of points in parameter space, which they call *parameter-sensitive hashing*. More recent work converts high-dimensional descriptor vectors into binary codes that can be compared using Hamming distances (Torralba, Weiss, and Fergus 2008; Weiss, Torralba, and Fergus 2008) or that can accommodate arbitrary kernel functions (Kulis and Grauman 2009; Raginsky and Lazebnik 2009).

Another widely used class of indexing structures are multi-dimensional search trees. The best known of these are *k-d trees*, also often written as *kd-trees*, which divide the multi-dimensional feature space along alternating axis-aligned hyperplanes, choosing the threshold along each axis so as to maximize some criterion, such as the search tree balance (Samet 1989). Figure 7.24 shows an example of a two-dimensional k-d tree. Here, eight different data points A–H are shown as small diamonds arranged on a two-dimensional plane. The k-d tree recursively splits this plane along axis-aligned (horizontal or vertical) cutting planes. Each split can be denoted using the dimension number and split value (Figure 7.24b). The splits are arranged so as to try to balance the tree, i.e., to keep its maximum depth as small as possible. At query time, a classic k-d tree search first locates the query point (+) in its appropriate bin (D), and then searches nearby leaves in the tree (C, B, ...) until it can guarantee that the nearest neighbor has been found. The best bin first (BBF) search (Beis and Lowe 1999) searches bins in order of their spatial proximity to the query point and is therefore usually more efficient.

Many additional data structures have been developed for solving exact and approximate nearest neighbor problems (Arya, Mount *et al.* 1998; Liang, Liu *et al.* 2001; Hjaltason and Samet 2003). For example, Nene and Nayar (1997) developed a technique they call *slicing* that uses a series of 1D binary searches on the point list sorted along different dimensions

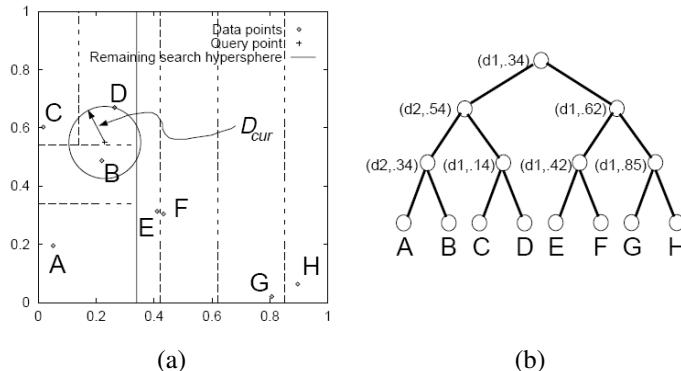


Figure 7.24 *K-d tree and best bin first (BBF) search (Beis and Lowe 1999) © 1999 IEEE:* (a) The spatial arrangement of the axis-aligned cutting planes is shown using dashed lines. Individual data points are shown as small diamonds. (b) The same subdivision can be represented as a tree, where each interior node represents an axis-aligned cutting plane (e.g., the top node cuts along dimension d_1 at value .34) and each leaf node is a data point. During a BBF search, a query point (denoted by “+”) first looks in its containing bin (D) and then in its nearest adjacent bin (B), rather than its closest neighbor in the tree (C).

to efficiently cull down a list of candidate points that lie within a hypercube of the query point. Grauman and Darrell (2005) reweight the matches at different levels of an indexing tree, which allows their technique to be less sensitive to discretization errors in the tree construction. Nistér and Stewénius (2006) use a *metric tree*, which compares feature descriptors to a small number of prototypes at each level in a hierarchy. The resulting quantized *visual words* can then be used with classical information retrieval (document relevance) techniques to quickly winnow down a set of potential candidates from a database of millions of images (Section 7.1.4). Muja and Lowe (2009) compare a number of these approaches, introduce a new one of their own (priority search on hierarchical k-means trees), and conclude that multiple randomized k-d trees often provide the best performance. Modern libraries for computing approximate nearest neighbors include FLANN (Muja and Lowe 2014) and Faiss (Johnson, Douze, and Jégou 2021), which are discussed in Section 5.1.1 and Appendix C.2.

Feature match verification and densification

Once we have some candidate matches, we can use geometric alignment (Section 8.1) to verify which matches are *inliers* and which ones are *outliers*. For example, if we expect the whole image to be translated or rotated in the matching view, we can fit a global geometric

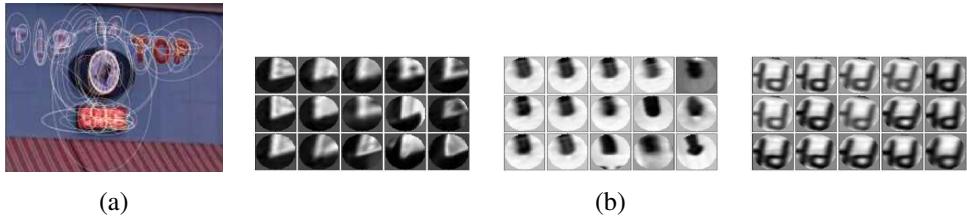


Figure 7.25 Visual words obtained from elliptical normalized affine regions (Sivic and Zisserman 2009) © 2009 IEEE. (a) Affine covariant regions are extracted from each frame and clustered into visual words using k-means clustering on SIFT descriptors with a learned Mahalanobis distance. (b) The central patch in each grid shows the query and the surrounding patches show the nearest neighbors.

transform and keep only those feature matches that are sufficiently close to this estimated transformation. The process of selecting a small set of seed matches and then verifying a larger set is often called *random sampling* or RANSAC (Section 8.1.4). Once an initial set of correspondences has been established, some systems look for additional matches, e.g., by looking for additional correspondences along epipolar lines (Section 12.1) or in the vicinity of estimated locations based on the global transform. It is also possible to use deep neural networks to perform feature matching and filtering, as in the SuperGlue system of Sarlin, DeTone *et al.* (2020). These topics are discussed further in Sections 8.1 and 12.2.

7.1.4 Large-scale matching and retrieval

As the number of objects in the database starts to grow (say, billions of objects or video frames), the time it takes to match a new image against each database image can become prohibitive. Instead of comparing the images one at a time, techniques are needed to quickly narrow down the search to a few likely images, which can then be compared using a more conservative verification stage.

The problem of quickly finding partial matches between documents is one of the central problems in *information retrieval* (IR) (Baeza-Yates and Ribeiro-Neto 1999; Manning, Raghavan, and Schütze 2008). In computer vision, the problem of finding a particular object in a large collection is called *content-based image retrieval (CBIR)* (Smeulders, Worring *et al.* 2000; Lew, Sebe *et al.* 2006; Vasconcelos 2007; Datta, Joshi *et al.* 2008) or *instance retrieval* (Zheng, Yang, and Tian 2018). The basic approach in fast document retrieval algorithms is to precompute an *inverted index* between individual words and the documents (or web pages or news stories) where they occur. More precisely, the *frequency* of occurrence of particular

words in a document is used to quickly find documents that match a particular query.

Sivic and Zisserman (2009) were the first to adapt IR techniques to visual search. In their Video Google system, affine invariant features are first detected in all the video frames they are indexing using both *shape adapted* regions around Harris feature points (Schaffalitzky and Zisserman 2002; Mikolajczyk and Schmid 2004) and maximally stable extremal regions (Matas, Chum *et al.* 2004; Section 7.1.1), as shown in Figure 7.25a. Next, 128-dimensional SIFT descriptors are computed from each normalized region (i.e., the patches shown in Figure 7.25b). Then, an average covariance matrix for these descriptors is estimated by accumulating statistics for features tracked from frame to frame. The feature descriptor covariance Σ is then used to define a Mahalanobis distance (5.32) between feature descriptors. In practice, feature descriptors are *whitened* by pre-multiplying them by $\Sigma^{-1/2}$ so that Euclidean distances can be used.⁵

To apply fast information retrieval techniques to images, the high-dimensional feature descriptors that occur in each image must first be mapped into discrete *visual words*. Sivic and Zisserman (2003) perform this mapping using k-means clustering, while some of the later methods (Nistér and Stewénius 2006; Philbin, Chum *et al.* 2007) use alternative techniques, such as vocabulary trees or randomized forests. To keep the clustering time manageable, only a few hundred video frames are used to learn the cluster centers, which still involves estimating several thousand clusters from about 300,000 descriptors, although subsequent work has greatly extended this capacity (Nistér and Stewénius 2006; Philbin, Chum *et al.* 2007; Mikulik, Perdoch *et al.* 2013). At visual query time, each feature in a new query region (e.g., Figure 7.25a, which is a cropped region from a larger video frame) is mapped to its corresponding visual word. To keep very common patterns from contaminating the results, a *stop list* of the most common visual words is created and such words are dropped from further consideration.

Once a query image or region has been mapped into its constituent visual words, likely matching images must then be retrieved from the database. The exact details of how this is done can be found in Sivic and Zisserman (2009), Nistér and Stewénius (2006), Philbin, Chum *et al.* (2007), Chum, Philbin *et al.* (2007), Philbin, Chum *et al.* (2008), and also in the first edition of this book (Szeliski 2010, Section 14.3.2). Because of the high efficiency in both quantizing and scoring features, the vocabulary-tree-based recognition system built by Nistér and Stewénius (2006) was able to process incoming images in real time against a database of 40,000 CD covers and at 1Hz when matching a database of one million frames

⁵Note that the computation of feature covariances from matched feature points is much more sensible than simply performing a PCA on the descriptor space (Winder and Brown 2007). This corresponds roughly to the *within-class* scatter matrix we studied in Section 5.2.3.

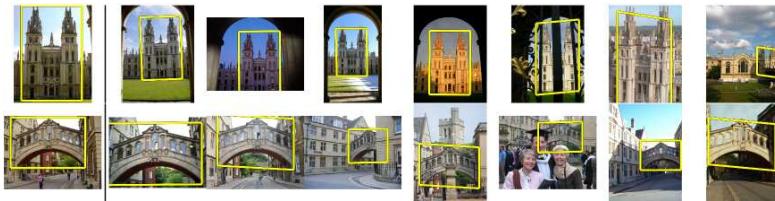


Figure 7.26 Location or building recognition using randomized trees (Philbin, Chum et al. 2007) © 2007 IEEE. The left image is the query, the other images are the highest-ranked results.

taken from six feature-length movies.

Instance recognition systems continued to improve rapidly in the 2000s. Philbin, Chum et al. (2007) showed that randomized forest of k-d trees perform better than vocabulary trees on a large location recognition task (Figure 7.26). They also compared the effects of using different 2D motion models (Section 2.1.1) in the verification stage. In follow-on work, Chum, Philbin et al. (2007) applied another idea from information retrieval, namely *query expansion*, which involves re-submitting top-ranked images from the initial query as additional queries to generate additional candidate results.⁶ Philbin, Chum et al. (2008) showed how to mitigate quantization problems in visual words selection using *soft assignment*, where each feature descriptor is mapped to a number of nearby visual words, which is similar to the multiple assignment idea proposed earlier by Jégou, Harzallah, and Schmid (2007). However, such techniques tend to reduce the sparsity of visual word vectors and increase the memory and computation costs. Jégou, Douze, and Schmid (2008) incorporated partial geometrical information and an explicit matching scheme between local descriptors in the initial large-scale image ranking stage. Taken together, these algorithms helped instance recognition algorithms perform Web-scale retrieval, matching, 3D reconstruction tasks (Agarwal, Furukawa et al. 2010, 2011; Frahm, Fite-Georgel et al. 2010; Snavely, Simon et al. 2010).

Since the “deep learning revolution” in 2012, researchers have started developing neural feature detectors and descriptors (Sections 7.1.1 and 7.1.2) and sometimes combining them into end-to-end matching systems.⁷ Figure 7.27 shows some of the major milestones in instance retrieval, while Figure 7.28 shows the variety of different classic and CNN-based retrieval architectures that have been considered. The survey paper by Zheng, Yang, and Tian (2018) describes and contrasts these various algorithms in more detail and also provides an

⁶An alternative to query expansion is *database-side augmentation* (Arandjelović and Zisserman 2012).

⁷But note that some popular open-source large-scale reconstruction systems such as COLMAP still use traditional features and indexing schemes (Schönberger and Frahm 2016).

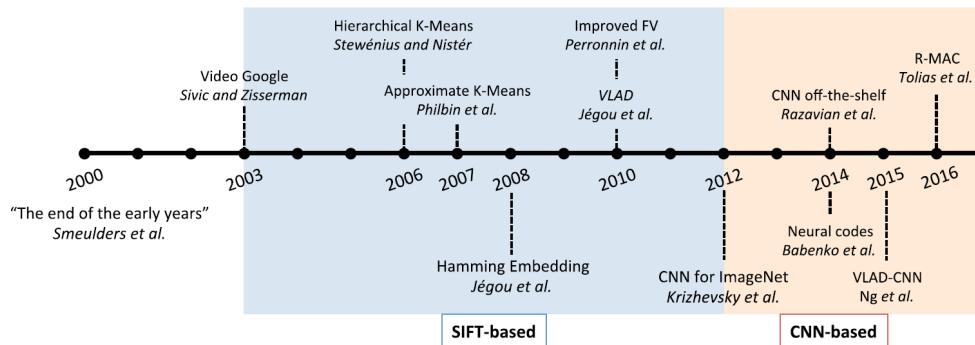


Figure 7.27 Milestones in instance retrieval (Zheng, Yang, and Tian 2018) © 2018 IEEE, showing the shift from hand-crafted feature-based retrieval to CNN-based approaches.

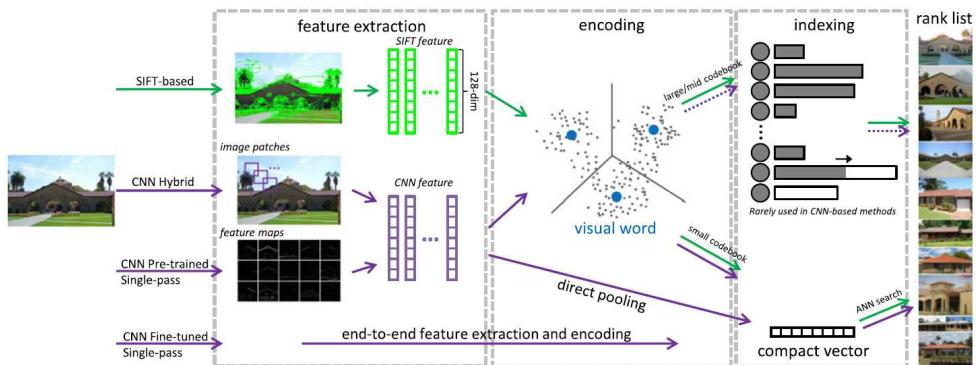


Figure 7.28 Typical pipeline for feature-based instance retrieval (Zheng, Yang, and Tian 2018) © 2018 IEEE, showing the feature extraction, encoding, and indexing portions, which are often collapsed when using a deep learning framework.

experimental comparison of some of these algorithms on image retrieval datasets. You can also find more details on related techniques and systems in Section 6.2.3 on visual similarity search, which discusses global descriptors that represent an image with a single vector (Arandjelovic, Gronat *et al.* 2016; Radenović, Tolias, and Chum 2019; Yang, Kien Nguyen *et al.* 2019; Cao, Araujo, and Sim 2020; Ng, Balntas *et al.* 2020; Tolias, Jenicek, and Chum 2020) as alternatives to bags of local features, Section 11.2.3 on location recognition, and Section 11.4.6 on large-scale 3D reconstruction from community (internet) photos.

7.1.5 Feature tracking

An alternative to independently finding features in all candidate images and then matching them is to find a set of likely feature locations in a first image and to then *search* for their corresponding locations in subsequent images. This kind of *detect then track* approach is more widely used for video tracking applications, where the expected amount of motion and appearance deformation between adjacent frames is expected to be small.

The process of selecting good features to track is closely related to selecting good features for more general recognition applications. In practice, regions containing high gradients in both directions, i.e., which have high eigenvalues in the auto-correlation matrix (7.8), provide stable locations at which to find correspondences (Shi and Tomasi 1994).

In subsequent frames, searching for locations where the corresponding patch has low squared difference (7.1) often works well enough. However, if the images are undergoing brightness change, explicitly compensating for such variations (9.9) or using *normalized cross-correlation* (9.11) may be preferable. If the search range is large, it is also often more efficient to use a *hierarchical* search strategy, which uses matches in lower-resolution images to provide better initial guesses and hence speed up the search (Section 9.1.1). Alternatives to this strategy involve learning what the appearance of the patch being tracked should be and then searching for it in the vicinity of its predicted position (Avidan 2001; Jurie and Dhome 2002; Williams, Blake, and Cipolla 2003). These topics are all covered in more detail in Section 9.1.3.

If features are being tracked over longer image sequences, their appearance can undergo larger changes. You then have to decide whether to continue matching against the originally detected patch (feature) or to re-sample each subsequent frame at the matching location. The former strategy is prone to failure, as the original patch can undergo appearance changes such as foreshortening. The latter runs the risk of the feature drifting from its original location to some other location in the image (Shi and Tomasi 1994). (Mathematically, small misregistration errors compound to create a *Markov random walk*, which leads to larger drift over time.)



Figure 7.29 Feature tracking using an affine motion model (Shi and Tomasi 1994) © 1994 IEEE, Top row: image patch around the tracked feature location. Bottom row: image patch after warping back toward the first frame using an affine deformation. Even though the speed sign gets larger from frame to frame, the affine transformation maintains a good resemblance between the original and subsequent tracked frames.

A preferable solution is to compare the original patch to later image locations using an *affine* motion model (Section 9.2). Shi and Tomasi (1994) first compare patches in neighboring frames using a translational model and then use the location estimates produced by this step to initialize an affine registration between the patch in the current frame and the base frame where a feature was first detected (Figure 7.29). In their system, features are only detected infrequently, i.e., only in regions where tracking has failed. In the usual case, an area around the current *predicted* location of the feature is searched with an incremental registration algorithm (Section 9.1.3). The resulting tracker is often called the Kanade–Lucas–Tomasi (KLT) tracker.

Since their original work on feature tracking, Shi and Tomasi’s approach has generated a plethora of follow-on papers and applications. Beardsley, Torr, and Zisserman (1996) use extended feature tracking combined with structure from motion (Chapter 11) to incrementally build up sparse 3D models from video sequences. Kang, Szeliski, and Shum (1997) tie together the corners of adjacent (regularly gridded) patches to provide some additional stability to the tracking, at the cost of poorer handling of occlusions. Tommasini, Fusiello *et al.* (1998) provide a better spurious match rejection criterion for the basic Shi and Tomasi algorithm, Collins and Liu (2003) provide improved mechanisms for feature selection and dealing with larger appearance changes over time, and Shafique and Shah (2005) develop algorithms for feature matching (data association) for videos with large numbers of moving objects or points. Lepetit and Fua (2005) and Yilmaz, Javed, and Shah (2006) survey the larger field of object tracking, which includes not only feature-based techniques but also alternative techniques based on contour and region (Section 7.3).



Figure 7.30 Real-time head tracking using fast trained classifiers (Lepetit, Pilet, and Fua 2004) © 2004 IEEE.

A more recent development in feature tracking is the use of learning algorithms to build special-purpose recognizers to rapidly search for matching features anywhere in an image (Lepetit, Pilet, and Fua 2006; Hinterstoisser, Benhimane *et al.* 2008; Rogez, Rihan *et al.* 2008; Özyusal, Calonder *et al.* 2010). By taking the time to train classifiers on sample patches and their affine deformations, extremely fast and reliable feature detectors can be constructed, which enables much faster motions to be supported (Figure 7.30). Coupling such features to deformable models (Pilet, Lepetit, and Fua 2008) or structure-from-motion algorithms (Klein and Murray 2008) can result in even higher stability.

While feature-based tracking is still widely used in real-time applications such as SLAM, autonomous navigation, and augmented reality (Section 11.5), a lot of current work on tracking is focused on whole *object tracking* (Chellappa, Sankaranarayanan *et al.* 2010; Smeulders, Chu *et al.* 2014), which we study in more detail in Section 9.4.4.

7.1.6 Application: Performance-driven animation

One of the most compelling applications of fast feature tracking is *performance-driven animation*, i.e., the interactive deformation of a 3D graphics model based on tracking a user's motions (Williams 1990; Litwinowicz and Williams 1994; Lepetit, Pilet, and Fua 2004).

Buck, Finkelstein *et al.* (2000) present a system that tracks a user's facial expressions and head motions and then uses them to morph among a series of hand-drawn sketches. An animator first extracts the eye and mouth regions of each sketch and draws control lines over each image (Figure 7.31a). At run time, a face-tracking system (Toyama 1998) determines the current location of these features (Figure 7.31b). The animation system decides which



Figure 7.31 *Performance-driven, hand-drawn animation (Buck, Finkelstein et al. 2000)*
 © 2000 ACM: (a) eye and mouth portions of hand-drawn sketch with their overlaid control lines; (b) an input video frame with the tracked features overlaid; (c) a different input video frame along with its (d) corresponding hand-drawn animation.

input images to morph based on nearest neighbor feature appearance matching and triangular barycentric interpolation. It also computes the global location and orientation of the head from the tracked features. The resulting morphed eye and mouth regions are then composited back into the overall head model to yield a frame of hand-drawn animation (Figure 7.31d).

In more recent work, Barnes, Jacobs *et al.* (2008) watch users animate paper cutouts on a desk and then turn the resulting motions and drawings into seamless 2D animations. Feature-based facial trackers continue to be widely used (Zollhöfer, Thies *et al.* 2018), both in the visual effects industry, as well as for real-time smartphone augmented reality effects such as Facebook’s Spark AR Face Masks.

7.2 Edges and contours

While interest points are useful for finding image locations that can be accurately matched in 2D, edge points are far more plentiful and often carry important semantic associations. For example, the boundaries of objects, which also correspond to occlusion events in 3D, are usually delineated by visible contours. Other kinds of edges correspond to shadow boundaries or crease edges, where surface orientation changes rapidly. Isolated edge points can also be grouped into longer *curves* or *contours*, as well as *straight line segments* (Section 7.4). It is interesting that even young children have no difficulty in recognizing familiar objects or animals from such simple line drawings.

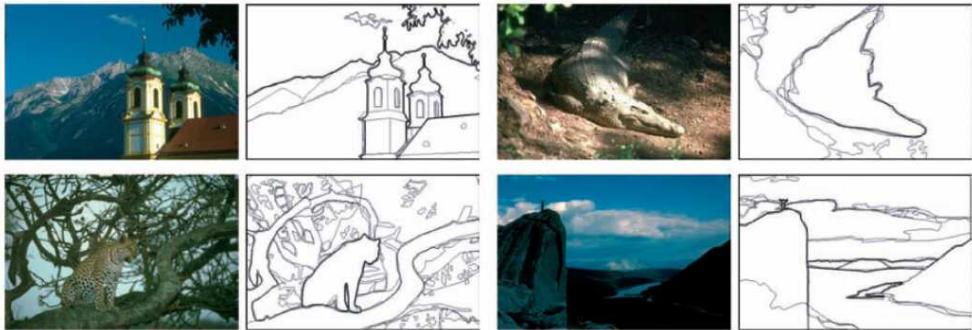


Figure 7.32 Human boundary detection (Martin, Fowlkes, and Malik 2004) © 2004 IEEE. The darkness of the edges corresponds to how many human subjects marked an object boundary at that location.

7.2.1 Edge detection

Given an image, how can we find the salient edges? Consider the color images in Figure 7.32. If someone asked you to point out the most “salient” or “strongest” edges or the object boundaries, which ones would you trace? How closely do your perceptions match the edge images shown in Figure 7.32?

Qualitatively, edges occur at boundaries between regions of different color, intensity, or texture (Martin, Fowlkes, and Malik 2004; Arbeláez, Maire *et al.* 2011; Pont-Tuset, Arbeláez *et al.* 2017). Unfortunately, segmenting an image into coherent regions is a difficult task, which we address in Section 7.5. Often, it is preferable to detect edges using only purely local information.

Under such conditions, a reasonable approach is to define an edge as a location of *rapid intensity or color variation*. Think of an image as a height field. On such a surface, edges occur at locations of *steep slopes*, or equivalently, in regions of closely packed contour lines (on a topographic map).

A mathematical way to define the slope and direction of a surface is through its gradient,

$$\mathbf{J}(\mathbf{x}) = \nabla I(\mathbf{x}) = \left(\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y} \right)(\mathbf{x}). \quad (7.19)$$

The local gradient vector \mathbf{J} points in the direction of *steepest ascent* in the intensity function. Its magnitude is an indication of the slope or strength of the variation, while its orientation points in a direction *perpendicular* to the local contour.

Unfortunately, taking image derivatives accentuates high frequencies and hence amplifies noise, as the proportion of noise to signal is larger at high frequencies. It is therefore prudent

to smooth the image with a low-pass filter prior to computing the gradient. Because we would like the response of our edge detector to be independent of orientation, a circularly symmetric smoothing filter is desirable. As we saw in Section 3.2, the Gaussian is the only separable circularly symmetric filter, so it is used in most edge detection algorithms. Canny (1986) discusses alternative filters and a number of researchers review alternative edge detection algorithms and compare their performance (Davis 1975; Nalwa and Binford 1986; Nalwa 1987; Deriche 1987; Freeman and Adelson 1991; Nalwa 1993; Heath, Sarkar *et al.* 1998; Crane 1997; Ritter and Wilson 2000; Bowyer, Kranenburg, and Dougherty 2001; Arbeláez, Maire *et al.* 2011; Pont-Tuset, Arbeláez *et al.* 2017).

Because differentiation is a linear operation, it commutes with other linear filtering operations. The gradient of the smoothed image can therefore be written as

$$\mathbf{J}_\sigma(\mathbf{x}) = \nabla[G_\sigma(\mathbf{x}) * I(\mathbf{x})] = [\nabla G_\sigma](\mathbf{x}) * I(\mathbf{x}), \quad (7.20)$$

i.e., we can convolve the image with the horizontal and vertical derivatives of the Gaussian kernel function,

$$\nabla G_\sigma(\mathbf{x}) = \left(\frac{\partial G_\sigma}{\partial x}, \frac{\partial G_\sigma}{\partial y} \right)(\mathbf{x}) = [-x \ -y] \frac{1}{\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right), \quad (7.21)$$

where the parameter σ indicates the width of the Gaussian. This is the same computation that is performed by Freeman and Adelson's (1991) first-order steerable filter, which we have already covered in Section 3.2.3.

For many applications, however, we wish to thin such a continuous gradient image to return isolated edges only, i.e., as single pixels at discrete locations along the edge contours. This can be achieved by looking for *maxima* in the edge strength (gradient magnitude) in a direction *perpendicular* to the edge orientation, i.e., along the gradient direction.

Finding this maximum corresponds to taking a directional derivative of the strength field in the direction of the gradient and then looking for zero crossings. The desired directional derivative is equivalent to the dot product between a second gradient operator and the results of the first,

$$S_\sigma(\mathbf{x}) = \nabla \cdot \mathbf{J}_\sigma(\mathbf{x}) = [\nabla^2 G_\sigma](\mathbf{x}) * I(\mathbf{x}). \quad (7.22)$$

The gradient operator dot product with the gradient is called the *Laplacian*. The convolution kernel

$$\nabla^2 G_\sigma(\mathbf{x}) = \left(\frac{x^2 + y^2}{\sigma^4} - \frac{2}{\sigma^2} \right) G_\sigma(\mathbf{x}), \quad (7.23)$$

is therefore called the *Laplacian of Gaussian* (LoG) kernel (Marr and Hildreth 1980). This kernel can be split into two separable parts,

$$\nabla^2 G_\sigma(\mathbf{x}) = \left(\frac{x^2}{2\sigma^4} - \frac{1}{\sigma^2} \right) G_\sigma(x) G_\sigma(y) + \left(\frac{y^2}{2\sigma^4} - \frac{1}{\sigma^2} \right) G_\sigma(y) G_\sigma(x) \quad (7.24)$$

(Wiejak, Buxton, and Buxton 1985), which allows for a much more efficient implementation using separable filtering (Section 3.2.1).

In practice, it is quite common to replace the Laplacian of Gaussian convolution with difference of Gaussian (DoG) computation, since the kernel shapes are qualitatively similar (Figure 3.34). This is especially convenient if a “Laplacian pyramid” (Section 3.5) has already been computed.⁸

In fact, it is not strictly necessary to take differences between adjacent levels when computing the edge field. Think about what a zero crossing in a “generalized” difference of Gaussians image represents. The finer (smaller kernel) Gaussian is a noise-reduced version of the original image. The coarser (larger kernel) Gaussian is an estimate of the average intensity over a larger region. Thus, whenever the DoG image changes sign, this corresponds to the (slightly blurred) image going from relatively darker to relatively lighter, as compared to the average intensity in that neighborhood.

Once we have computed the sign function $S(\mathbf{x})$, we must find its *zero crossings* and convert these into edge elements (*edgels*). An easy way to detect and represent zero crossings is to look for adjacent pixel locations \mathbf{x}_i and \mathbf{x}_j where the sign changes value, i.e., $[S(\mathbf{x}_i) > 0] \neq [S(\mathbf{x}_j) > 0]$.

The sub-pixel location of this crossing can be obtained by computing the “ x -intercept” of the “line” connecting $S(\mathbf{x}_i)$ and $S(\mathbf{x}_j)$,

$$\mathbf{x}_z = \frac{\mathbf{x}_i S(\mathbf{x}_j) - \mathbf{x}_j S(\mathbf{x}_i)}{S(\mathbf{x}_j) - S(\mathbf{x}_i)}. \quad (7.25)$$

The orientation and strength of such edgels can be obtained by linearly interpolating the gradient values computed on the original pixel grid.

An alternative edgel representation can be obtained by linking adjacent edgels on the dual grid to form edgels that live *inside* each square formed by four adjacent pixels in the original pixel grid.⁹ The advantage of this representation is that the edgels now live on a grid offset by half a pixel from the original pixel grid and are thus easier to store and access. As before, the orientations and strengths of the edges can be computed by interpolating the gradient field or estimating these values from the difference of Gaussian image (see Exercise 7.7).

In applications where the accuracy of the edge orientation is more important, higher-order steerable filters can be used (Freeman and Adelson 1991) (see Section 3.2.3). Such filters are more selective for more elongated edges and also have the possibility of better modeling curve

⁸Recall that Burt and Adelson’s (1983a) “Laplacian pyramid” actually computes differences of Gaussian-filtered levels.

⁹This algorithm is a 2D version of the 3D *marching cubes* isosurface extraction algorithm (Lorensen and Cline 1987).

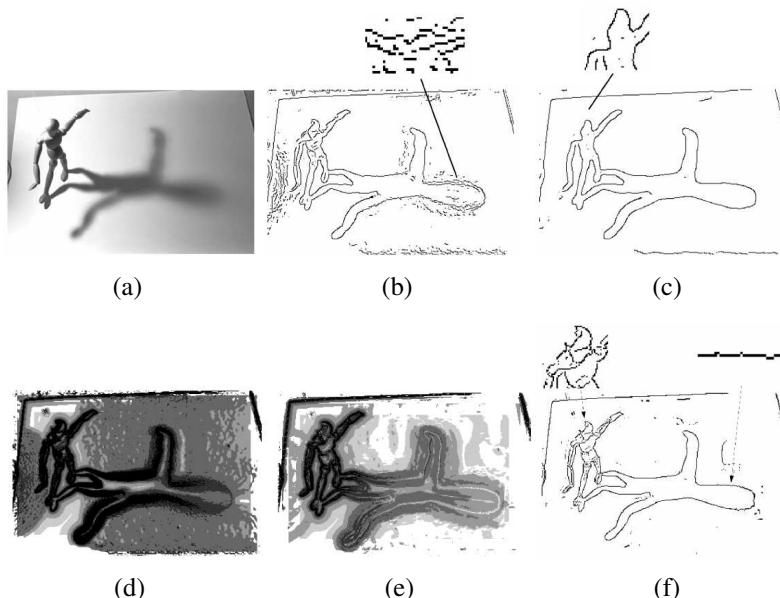


Figure 7.33 Scale selection for edge detection (Elder and Zucker 1998) © 1998 IEEE: (a) original image; (b–c) Canny/Deriche edge detector tuned to the finer (mannequin) and coarser (shadow) scales; (d) minimum reliable scale for gradient estimation; (e) minimum reliable scale for second derivative estimation; (f) final detected edges.

intersections because they can represent multiple orientations at the same pixel (Figure 3.16). Their disadvantage is that they are more expensive to compute and the directional derivative of the edge strength does not have a simple closed form solution.¹⁰

Scale selection and blur estimation

As we mentioned before, the derivative, Laplacian, and Difference of Gaussian filters (7.20–7.23) all require the selection of a spatial scale parameter σ . If we are only interested in detecting sharp edges, the width of the filter can be determined from image noise characteristics (Canny 1986; Elder and Zucker 1998). However, if we want to detect edges that occur at different resolutions (Figures 7.33b–c), a *scale-space* approach that detects and then selects edges at different scales may be necessary (Witkin 1983; Lindeberg 1994, 1998a; Nielsen, Florack, and Deriche 1997).

¹⁰In fact, the edge orientation can have a 180° ambiguity for “bar edges”, which makes the computation of zero crossings in the derivative more tricky.

Elder and Zucker (1998) present a principled approach to solving this problem. Given a known image noise level, their technique computes, for every pixel, the minimum scale at which an edge can be reliably detected (Figure 7.33d). Their approach first computes gradients densely over an image by selecting among gradient estimates computed at different scales, based on their gradient magnitudes. It then performs a similar estimate of minimum scale for directed second derivatives and uses zero crossings of this latter quantity to robustly select edges (Figures 7.33e–f). As an optional final step, the blur width of each edge can be computed from the distance between extrema in the second derivative response minus the width of the Gaussian filter.

Color edge detection

While most edge detection techniques have been developed for grayscale images, color images can provide additional information. For example, noticeable edges between *iso-luminant* colors (colors that have the same luminance) are useful cues but fail to be detected by grayscale edge operators.

One simple approach is to combine the outputs of grayscale detectors run on each color band separately.¹¹ However, some care must be taken. For example, if we simply sum up the gradients in each of the color bands, the signed gradients may actually cancel each other! (Consider, for example a pure red-to-green edge.) We could also detect edges independently in each band and then take the union of these, but this might lead to thickened or doubled edges that are hard to link.

A better approach is to compute the *oriented energy* in each band (Morrone and Burr 1988; Perona and Malik 1990a), e.g., using a second-order steerable filter (Section 3.2.3) (Freeman and Adelson 1991), and then sum up the orientation-weighted energies and find their joint best orientation. Unfortunately, the directional derivative of this energy may not have a closed form solution (as in the case of signed first-order steerable filters), so a simple zero crossing-based strategy cannot be used. However, the technique described by Elder and Zucker (1998) can be used to compute these zero crossings numerically instead.

An alternative approach is to estimate local color statistics in regions around each pixel (Ruzon and Tomasi 2001; Martin, Fowlkes, and Malik 2004). This has the advantage that more sophisticated techniques (e.g., 3D color histograms) can be used to compare regional statistics and that additional measures, such as texture, can also be considered. Figure 7.34 shows the output of such detectors.

¹¹Instead of using the raw RGB space, a more perceptually uniform color space such as L*a*b* (see Section 2.3.2) can be used instead. When trying to match human performance (Martin, Fowlkes, and Malik 2004), this makes sense. However, in terms of the physics of the underlying image formation and sensing, it may be a questionable strategy.

Over the years, many other approaches have been developed for detecting color edges, dating back to early work by Nevatia (1977). Ruzon and Tomasi (2001) and Gevers, van de Weijer, and Stokman (2006) provide good reviews of these approaches, which include ideas such as fusing outputs from multiple channels, using multidimensional gradients, and vector-based methods.

Combining edge feature cues

If the goal of edge detection is to match human *boundary detection* performance (Bowyer, Kranenburg, and Dougherty 2001; Martin, Fowlkes, and Malik 2004; Arbeláez, Maire *et al.* 2011; Pont-Tuset, Arbeláez *et al.* 2017), as opposed to simply finding stable features for matching, even better detectors can be constructed by combining multiple low-level cues such as brightness, color, and texture.

Martin, Fowlkes, and Malik (2004) describe a system that combines brightness, color, and texture edges to produce state-of-the-art performance on a database of hand-segmented natural color images (Martin, Fowlkes *et al.* 2001). First, they construct and train separate oriented half-disc detectors for measuring significant differences in brightness (luminance), color (a^* and b^* channels, summed responses), and texture (un-normalized filter bank responses from the work of Malik, Belongie *et al.* (2001)). Some of the responses are then sharpened using a soft non-maximal suppression technique. Finally, the outputs of the three detectors are combined using a variety of machine-learning techniques, from which logistic regression is found to have the best tradeoff between speed, space, and accuracy . The resulting system (see Figure 7.34 for some examples) is shown to outperform previously developed techniques. Maire, Arbelaez *et al.* (2008) improve on these results by combining the detector based on local appearance with a *spectral* (segmentation-based) detector (Belongie and Malik 1998). In follow-on work, Arbeláez, Maire *et al.* (2011) build a hierarchical segmentation on top of this edge detector using a variant of the watershed algorithm.

7.2.2 Contour detection

While isolated edges can be useful for a variety of applications, such as line detection (Section 7.4) and sparse stereo matching (Section 12.2), they become even more useful when linked into continuous contours.

If the edges have been detected using zero crossings of some function, linking them up is straightforward, since adjacent edgels share common endpoints. Linking the edgels into chains involves picking up an unlinked edgel and following its neighbors in both directions. Either a sorted list of edgels (sorted first by x coordinates and then by y coordinates, for



Figure 7.34 Combined brightness, color, texture boundary detector (Martin, Fowlkes, and Malik 2004) © 2004 IEEE. Successive rows show the outputs of the brightness gradient (BG), color gradient (CG), texture gradient (TG), and combined (BG+CG+TG) detectors. The final row shows human-labeled boundaries derived from a database of hand-segmented images (Martin, Fowlkes et al. 2001).

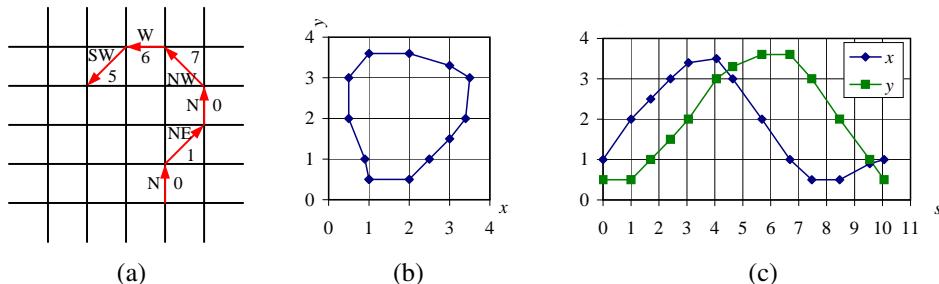


Figure 7.35 Some coding alternatives for linked contours. (a) A chain code representation of a grid-aligned linked edge chain. The code is represented as a series of direction codes, e.g., 0 1 0 7 6 5, which can further be compressed using predictive and run-length coding. (b–c) Arc-length parameterization of a contour. Discrete points along the contour (b) are first transcribed as (c) (x, y) pairs along the arc length s . This curve can then be regularly re-sampled or converted into alternative (e.g., Fourier) representations.

example) or a 2D array can be used to accelerate the neighbor finding. If edges were not detected using zero crossings, finding the continuation of an edgel can be tricky. In this case, comparing the orientation (and, optionally, phase) of adjacent edgelets can be used for disambiguation. Ideas from connected component computation can also sometimes be used to make the edge linking process even faster (see Exercise 7.8).

Once the edgelets have been linked into chains, we can apply an optional thresholding with hysteresis to remove low-strength contour segments (Canny 1986). The basic idea of hysteresis is to set two different thresholds and allow a curve being tracked above the higher threshold to dip in strength down to the lower threshold.

Linked edgelet lists can be encoded more compactly using a variety of alternative representations. A *chain code* encodes a list of connected points lying on an \mathcal{N}_8 grid using a three-bit code corresponding to the eight cardinal directions (N, NE, E, SE, S, SW, W, NW) between a point and its successor (Figure 7.35a). While this representation is more compact than the original edgelet list (especially if predictive variable-length coding is used), it is not very suitable for further processing.

A more useful representation is the *arc length parameterization* of a contour, $\mathbf{x}(s)$, where s denotes the arc length along a curve. Consider the linked set of edgelets shown in Figure 7.35b. We start at one point (the dot at $(1.0, 0.5)$ in Figure 7.35c) and plot it at coordinate $s = 0$ (Figure 7.35c). The next point at $(2.0, 0.5)$ gets plotted at $s = 1$, and the next point at $(2.5, 1.0)$ gets plotted at $s = 1.7071$, i.e., we increment s by the length of each edge segment. The resulting plot can be resampled on a regular (say, integral) s grid before further

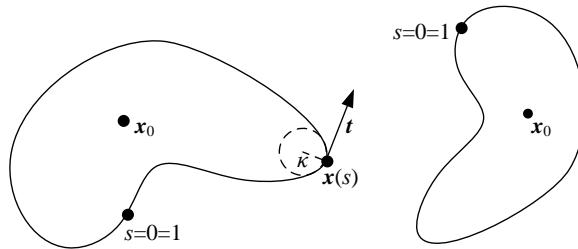


Figure 7.36 Matching two contours using their arc-length parameterization. If both curves are normalized to unit length, $s \in [0, 1]$ and centered around their centroid \mathbf{x}_0 , they will have the same descriptor up to an overall “temporal” shift (due to different starting points for $s = 0$) and a phase (x - y) shift (due to rotation).

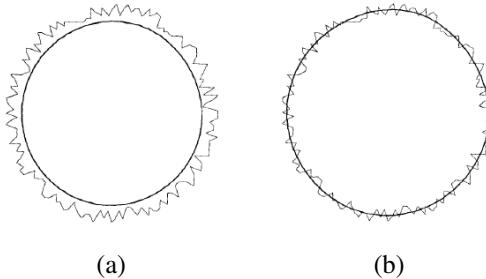


Figure 7.37 Curve smoothing with a Gaussian kernel (Lowe 1988) © 1998 IEEE: (a) without a shrinkage correction term; (b) with a shrinkage correction term.

processing.

The advantage of the arc-length parameterization is that it makes matching and processing (e.g., smoothing) operations much easier. Consider the two curves describing similar shapes shown in Figure 7.36. To compare the curves, we first subtract the average values $\mathbf{x}_0 = \int_s \mathbf{x}(s)$ from each descriptor. Next, we rescale each descriptor so that s goes from 0 to 1 instead of 0 to S , i.e., we divide $\mathbf{x}(s)$ by S . Finally, we take the Fourier transform of each normalized descriptor, treating each $\mathbf{x} = (x, y)$ value as a complex number. If the original curves are the same (up to an unknown scale and rotation), the resulting Fourier transforms should differ only by a scale change in magnitude plus a constant complex phase shift, due to rotation, and a linear phase shift in the domain, due to different starting points for s (see Exercise 7.9).

Arc-length parameterization can also be used to smooth curves to remove digitization

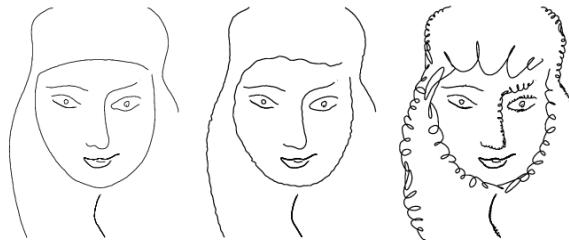


Figure 7.38 *Changing the character of a curve without affecting its sweep (Finkelstein and Salesin 1994) © 1994 ACM: higher frequency wavelets can be replaced with exemplars from a style library to effect different local appearances.*

noise. However, if we just apply a regular smoothing filter, the curve tends to shrink on itself (Figure 7.37a). Lowe (1989) and Taubin (1995) describe techniques that compensate for this shrinkage by adding an offset term based on second derivative estimates or a larger smoothing kernel (Figure 7.37b). An alternative approach, based on selectively modifying different frequencies in a wavelet decomposition, is presented by Finkelstein and Salesin (1994). In addition to controlling shrinkage without affecting its “sweep”, wavelets allow the “character” of a curve to be interactively modified, as shown in Figure 7.38.

The evolution of curves as they are smoothed and simplified is related to “grassfire” (distance) transforms and region skeletons (Section 3.3.3) (Tek and Kimia 2003), and can be used to recognize objects based on their contour shape (Sebastian and Kimia 2005). More local descriptors of curve shape such as *shape contexts* (Belongie, Malik, and Puzicha 2002) can also be used for recognition and are potentially more robust to missing parts due to occlusions.

The field of contour detection and linking continues to evolve rapidly and now includes techniques for global contour grouping, boundary completion, and junction detection (Maire, Arbelaez *et al.* 2008), as well as grouping contours into likely regions (Arbeláez, Maire *et al.* 2011) and wide-baseline correspondence (Meltzer and Soatto 2008). Some additional papers that address contour detection include Xiaofeng and Bo (2012), Lim, Zitnick, and Dollár (2013), Dollár and Zitnick (2015), Xie and Tu (2015), and Pont-Tuset, Arbeláez *et al.* (2017).

7.2.3 Application: Edge editing and enhancement

While edges can serve as components for object recognition or features for matching, they can also be used directly for image editing.

In fact, if the edge magnitude and blur estimate are kept along with each edge, a visually

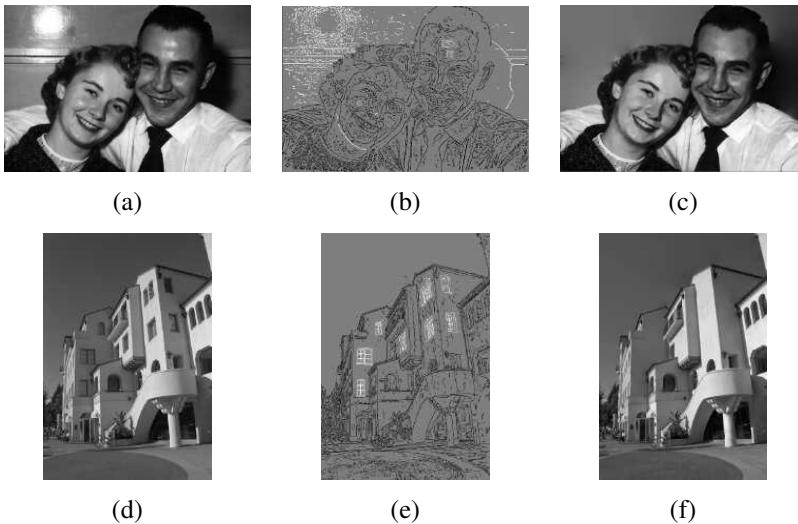


Figure 7.39 Image editing in the contour domain (Elder and Goldberg 2001) © 2001 IEEE: (a) and (d) original images; (b) and (e) extracted edges (edges to be deleted are marked in white); (c) and (f) reconstructed edited images.

similar image can be reconstructed from this information (Elder 1999). Based on this principle, Elder and Goldberg (2001) propose a system for “image editing in the contour domain”. Their system allows users to selectively remove edges corresponding to unwanted features such as specularities, shadows, or distracting visual elements. After reconstructing the image from the remaining edges, the undesirable visual features have been removed (Figure 7.39).

Another potential application is to enhance perceptually salient edges while simplifying the underlying image to produce a cartoon-like or “pen-and-ink” stylized image (DeCarlo and Santella 2002). This application is discussed in more detail in Section 10.5.2.

7.3 Contour tracking

While lines, vanishing points, and rectangles are commonplace in the human-made world, curves corresponding to object boundaries are even more common, especially in the natural environment. In this section, we describe some approaches to locating such boundary curves in images.

The first, originally called *snakes* by its inventors (Kass, Witkin, and Terzopoulos 1988) (Section 7.3.1), is an energy-minimizing, two-dimensional spline curve that evolves (moves)

towards image features such as strong edges. The second, *intelligent scissors* (Mortensen and Barrett 1995) (Section 7.3.1), allows the user to sketch in real time a curve that clings to object boundaries. Finally, *level set* techniques (Section 7.3.2) evolve the curve as the zero-set of a *characteristic function*, which allows them to easily change topology and incorporate region-based statistics.

All three of these are examples of *active contours* (Blake and Isard 1998; Mortensen 1999), since these boundary detectors iteratively move towards their final solution under the combination of image and optional user-guidance forces. The presentation below is heavily shortened from that presented in the first edition of this book (Szeliski 2010, Section 5.1), where interested readers can find more details.

7.3.1 Snakes and scissors

Snakes are a two-dimensional generalization of the 1D energy-minimizing splines first introduced in Section 4.2,

$$\mathcal{E}_{\text{int}} = \int \alpha(s) \|\mathbf{f}_s(s)\|^2 + \beta(s) \|\mathbf{f}_{ss}(s)\|^2 ds, \quad (7.26)$$

where s is the arc-length along the curve $\mathbf{f}(s) = (x(s), y(s))$ and $\alpha(s)$ and $\beta(s)$ are first- and second-order continuity weighting functions analogous to the $s(x, y)$ and $c(x, y)$ terms introduced in (4.24–4.25). We can discretize this energy by sampling the initial curve position evenly along its length (Figure 7.35c) to obtain

$$\begin{aligned} E_{\text{int}} = & \sum_i \alpha(i) \|f(i+1) - f(i)\|^2 / h^2 \\ & + \beta(i) \|f(i+1) - 2f(i) + f(i-1)\|^2 / h^4, \end{aligned} \quad (7.27)$$

where h is the step size, which can be neglected if we resample the curve along its arc-length after each iteration.

In addition to this *internal* spline energy, a snake simultaneously minimizes external image-based and constraint-based potentials. The image-based potentials are the sum of several terms

$$\mathcal{E}_{\text{image}} = w_{\text{line}} \mathcal{E}_{\text{line}} + w_{\text{edge}} \mathcal{E}_{\text{edge}} + w_{\text{term}} \mathcal{E}_{\text{term}}, \quad (7.28)$$

where the *line* term attracts the snake to dark ridges, the *edge* term attracts it to strong gradients (edges), and the *term* term attracts it to line terminations. As the snakes evolve by minimizing their energy, they often “wiggle” and “slither”, which accounts for their popular name.

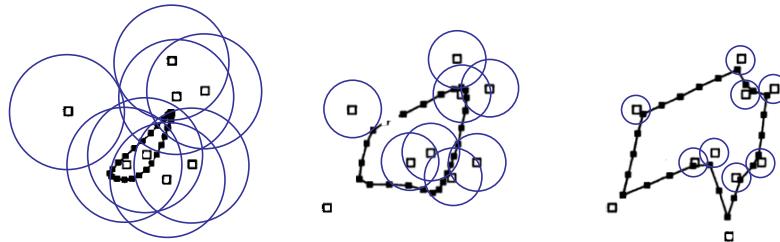


Figure 7.40 *Elastic net: The open squares indicate the cities and the closed squares linked by straight line segments are the tour points. The blue circles indicate the approximate extent of the attraction force of each city, which is reduced over time. Under the Bayesian interpretation of the elastic net, the blue circles correspond to one standard deviation of the circular Gaussian that generates each city from some unknown tour point.*

Because regular snakes have a tendency to shrink, it is usually better to initialize them by drawing the snake outside the object of interest to be tracked. Alternatively, an expansion *ballooning* force can be added to the dynamics (Cohen and Cohen 1993), essentially moving each point outwards along its normal. It is also possible to replace the energy-minimizing variational evolution equations with a deep neural network to significantly improve performance (Peng, Jiang *et al.* 2020).

Elastic nets and slippery springs

An interesting variant on snakes, first proposed by Durbin and Willshaw (1987) and later re-formulated in an energy-minimizing framework by Durbin, Szeliski, and Yuille (1989), is the *elastic net* formulation of the Traveling Salesman Problem (TSP). Recall that in a TSP, the salesman must visit each city once while minimizing the total distance traversed. A snake that is constrained to pass through each city could solve this problem (without any optimality guarantees) but it is impossible to tell ahead of time which snake control point should be associated with each city.

Instead of having a fixed constraint between snake nodes and cities, a city is assumed to pass near *some* point along the tour (Figure 7.40). In a probabilistic interpretation, each city is generated as a *mixture* of Gaussians centered at each tour point,

$$p(\mathbf{d}(j)) = \sum_i p_{ij} \quad \text{with} \quad p_{ij} = e^{-d_{ij}^2/(2\sigma^2)}, \quad (7.29)$$

where σ is the standard deviation of the Gaussian and

$$d_{ij} = \|\mathbf{f}(i) - \mathbf{d}(j)\| \quad (7.30)$$

is the Euclidean distance between a tour point $\mathbf{f}(i)$ and a city location $\mathbf{d}(j)$. The corresponding data fitting energy (negative log likelihood) is

$$E_{\text{slippery}} = - \sum_j \log p(\mathbf{d}(j)) = - \sum_j \log \left[\sum e^{-\|\mathbf{f}(i) - \mathbf{d}(j)\|^2 / 2\sigma^2} \right]. \quad (7.31)$$

This energy derives its name from the fact that, unlike a regular spring, which couples a given snake point to a given constraint, this alternative energy defines a *slippery spring* that allows the association between constraints (cities) and curve (tour) points to evolve over time (Szeliski 1989). Note that this is a soft variant of the popular *iterative closest point* data constraint that is often used in fitting or aligning surfaces to data points or to each other (Section 13.2.1) (Besl and McKay 1992; Chen and Medioni 1992; Zhang 1994).

To compute a good solution to the TSP, the slippery spring data association energy is combined with a regular first-order internal smoothness energy (7.27) to define the cost of a tour. The tour $\mathbf{f}(s)$ is initialized as a small circle around the mean of the city points and σ is progressively lowered (Figure 7.40). For large σ values, the tour tries to stay near the centroid of the points but as σ decreases each city pulls more and more strongly on its closest tour points (Durbin, Szeliski, and Yuille 1989). In the limit as $\sigma \rightarrow 0$, each city is guaranteed to capture at least one tour point and the tours between subsequent cities become straight lines.

Splines and shape priors

While snakes can be very good at capturing the fine and irregular detail in many real-world contours, they sometimes exhibit too many degrees of freedom, making it more likely that they can get trapped in local minima during their evolution.

One solution to this problem is to control the snake with fewer degrees of freedom through the use of B-spline approximations (Menet, Saint-Marc, and Medioni 1990b,a; Cipolla and Blake 1990). The resulting *B-snake* can be written as

$$\mathbf{f}(s) = \sum_k B_k(s) \mathbf{x}_k. \quad (7.32)$$

If the object being tracked or recognized has large variations in location, scale, or orientation, these can be modeled as an additional transformation on the control points, e.g., $\mathbf{x}'_k = s\mathbf{R}\mathbf{x}_k + \mathbf{t}$ (2.18), which can be estimated at the same time as the values of the control points. Alternatively, separate *detection* and *alignment* stages can be run to first localize and orient the objects of interest (Cootes, Cooper *et al.* 1995).

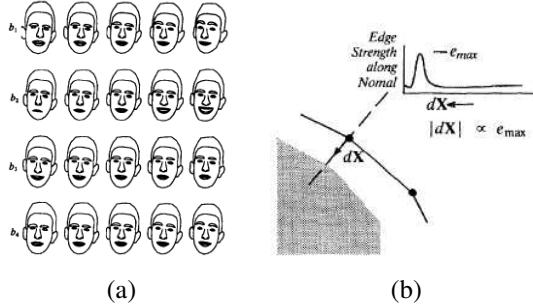


Figure 7.41 Active Shape Model (ASM): (a) the effect of varying the first four shape parameters for a set of faces (Cootes, Taylor et al. 1993) © 1993 IEEE; (b) searching for the strongest gradient along the normal to each control point (Cootes, Cooper et al. 1995) © 1995 Elsevier.

In a B-snake, because the snake is controlled by fewer degrees of freedom, there is less need for the internal smoothness forces used with the original snakes, although these can still be derived and implemented using finite element analysis, i.e., taking derivatives and integrals of the B-spline basis functions (Terzopoulos 1983; Bathe 2007).

In practice, it is more common to estimate a set of *shape priors* on the typical distribution of the control points $\{\mathbf{x}_k\}$ (Cootes, Cooper et al. 1995). One potential way of describing this distribution would be by the location $\bar{\mathbf{x}}_k$ and 2D covariance \mathbf{C}_k of each individual point \mathbf{x}_k . These could then be turned into a quadratic penalty (prior energy) on the point location. In practice, however, the variation in point locations is usually highly correlated.

A preferable approach is to estimate the joint covariance of all the points simultaneously. First, concatenate all of the point locations $\{\mathbf{x}_k\}$ into a single vector \mathbf{x} , e.g., by interleaving the x and y locations of each point. The distribution of these vectors across all training examples can be described with a mean $\bar{\mathbf{x}}$ and a covariance

$$\mathbf{C} = \frac{1}{P} \sum_p (\mathbf{x}_p - \bar{\mathbf{x}})(\mathbf{x}_p - \bar{\mathbf{x}})^T, \quad (7.33)$$

where \mathbf{x}_p are the P training examples. Using *eigenvalue analysis* (Appendix A.1.2), which is also known as *principal component analysis* (PCA) (Section 5.2.3 and Appendix B.1), the covariance matrix can be written as,

$$\mathbf{C} = \boldsymbol{\Phi} \operatorname{diag}(\lambda_0 \dots \lambda_{K-1}) \boldsymbol{\Phi}^T. \quad (7.34)$$

In most cases, the likely appearance of the points can be modeled using only a few eigenvectors with the largest eigenvalues. The resulting *point distribution model* (Cootes, Taylor *et al.* 1993; Cootes, Cooper *et al.* 1995) can be written as

$$\mathbf{x} = \bar{\mathbf{x}} + \hat{\Phi} \mathbf{b}, \quad (7.35)$$

where \mathbf{b} is an $M \ll K$ element *shape parameter* vector and $\hat{\Phi}$ are the first m columns of Φ . To constrain the shape parameters to reasonable values, we can use a quadratic penalty of the form

$$E_{\text{shape}} = \frac{1}{2} \mathbf{b}^T \text{diag}(\lambda_0 \dots \lambda_{M-1}) \mathbf{b} = \sum_m b_m^2 / 2\lambda_m. \quad (7.36)$$

Alternatively, the range of allowable b_m values can be limited to some range, e.g., $|b_m| \leq 3\sqrt{\lambda_m}$ (Cootes, Cooper *et al.* 1995). Alternative approaches for deriving a set of shape vectors are reviewed by Isard and Blake (1998). Varying the individual shape parameters b_m over the range $-2\sqrt{\lambda_m} \leq 2\sqrt{\lambda_m}$ can give a good indication of the expected variation in appearance, as shown in Figure 7.41a.

To align a point distribution model with an image, each control point searches in a direction normal to the contour to find the most likely corresponding image edge point (Figure 7.41b). These individual measurements can be combined with priors on the shape parameters (and, if desired, position, scale, and orientation parameters) to estimate a new set of parameters. The resulting *active shape model* (ASM) can be iteratively minimized to fit images to non-rigidly deforming objects, such as medical images, or body parts, such as hands (Cootes, Cooper *et al.* 1995). The ASM can also be combined with a PCA analysis of the underlying gray-level distribution to create an *active appearance model* (AAM) (Cootes, Edwards, and Taylor 2001), which we discussed in more detail in Section 6.2.4.

Dynamic snakes and CONDENSATION

In many applications of active contours, the object of interest is being tracked from frame to frame as it deforms and evolves. In this case, it makes sense to use estimates from the previous frame to predict and constrain the new estimates.

One way to do this is to use Kalman filtering, which results in a formulation called *Kalman snakes* (Terzopoulos and Szeliski 1992; Blake, Curwen, and Zisserman 1993). The Kalman filter is based on a linear dynamic model of shape parameter evolution,

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{w}_t, \quad (7.37)$$

where \mathbf{x}_t and \mathbf{x}_{t-1} are the current and previous state variables, \mathbf{A} is the linear *transition matrix*, and \mathbf{w} is a noise (perturbation) vector, which is often modeled as a Gaussian (Gelb

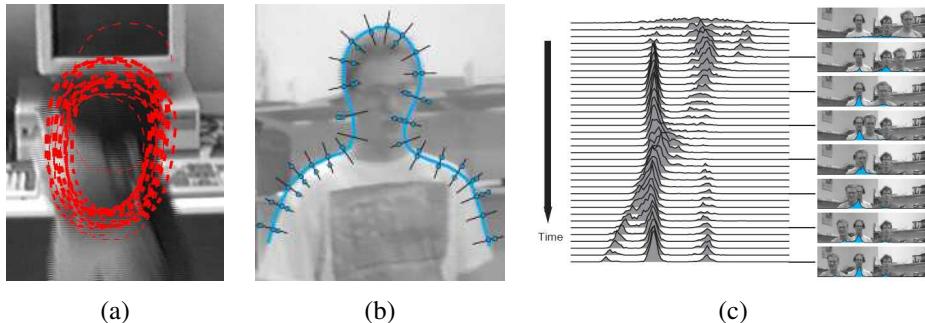


Figure 7.42 Head tracking using CONDENSATION (Isard and Blake 1998) © 1998 Springer: (a) sample set representation of head estimate distribution; (b) multiple measurements at each control vertex location; (c) multi-hypothesis tracking over time.

1974). The matrices \mathbf{A} and the noise covariance can be learned ahead of time by observing typical sequences of the object being tracked (Blake and Isard 1998).

In many situations, however, such as when tracking in clutter, a better estimate for the contour can be obtained if we remove the assumptions that the distributions are Gaussian, which is what the Kalman filter requires. In this case, a general multi-modal distribution is propagated. To model such multi-modal distributions, Isard and Blake (1998) introduced the use of *particle filtering* to the computer vision community.¹² Particle filtering techniques represent a probability distribution using a collection of weighted point samples (Andrieu, de Freitas *et al.* 2003; Bishop 2006; Koller and Friedman 2009).

To update the locations of the samples according to the linear dynamics (deterministic drift), the centers of the samples are updated and multiple samples are generated for each point. These are then perturbed to account for the stochastic diffusion, i.e., their locations are moved by random vectors taken from the distribution of w .¹³ Finally, the weights of these samples are multiplied by the measurement probability density, i.e., we take each sample and measure its likelihood given the current (new) measurements. Because the point samples represent and propagate conditional estimates of the multi-modal density, Isard and Blake (1998) dubbed their algorithm CONditional DENSity propagATION or CONDENSATION.

Figure 7.42a shows what a factored sample of a head tracker might look like, drawing a red B-spline contour for each of (a subset of) the particles being tracked. Figure 7.42b shows why the measurement density itself is often multi-modal: the locations of the edges

¹²Alternatives to modeling multi-modal distributions include *mixtures of Gaussians* (Section 5.2.2) and *multiple hypothesis tracking* (Bar-Shalom and Fortmann 1988; Cham and Rehg 1999).

¹³Note that because of the structure of these steps, non-linear dynamics and non-Gaussian noise can be used.

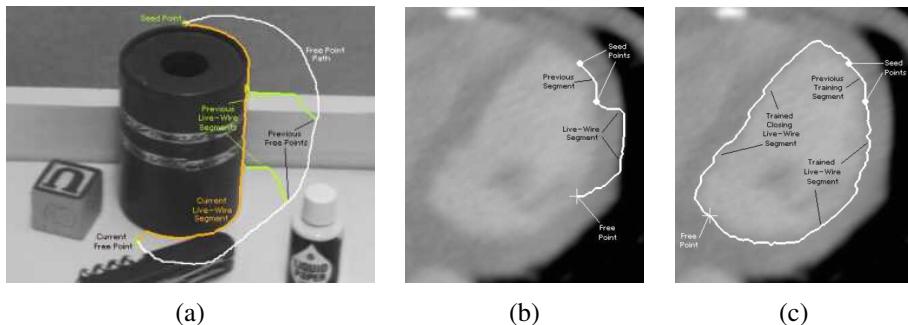


Figure 7.43 Intelligent scissors: (a) as the mouse traces the white path, the scissors follow the orange path along the object boundary (the green curves show intermediate positions) (Mortensen and Barrett 1995) © 1995 ACM; (b) regular scissors can sometimes jump to a stronger (incorrect) boundary; (c) after training to the previous segment, similar edge profiles are preferred (Mortensen and Barrett 1998) © 1995 Elsevier.

perpendicular to the spline curve can have multiple local maxima due to background clutter. Finally, Figure 7.42c shows the temporal evolution of the conditional density (x coordinate of the head and shoulder tracker centroid) as it tracks several people over time.

Scissors

Active contours allow a user to roughly specify a boundary of interest and have the system evolve the contour towards a more accurate location as well as track it over time. The results of this curve evolution, however, may be unpredictable and may require additional user-based hints to achieve the desired result.

An alternative approach is to have the system optimize the contour in real time as the user is drawing (Mortensen 1999). The *intelligent scissors* system developed by Mortensen and Barrett (1995) does just that. As the user draws a rough outline (the white curve in Figure 7.43a), the system computes and draws a better curve that clings to high-contrast edges (the orange curve).

To compute the optimal curve path (*live-wire*), the image is first pre-processed to associate low costs with edges (links between neighboring horizontal, vertical, and diagonal, i.e., \mathcal{N}_8 neighbors) that are likely to be boundary elements. Their system uses a combination of zero-crossing, gradient magnitudes, and gradient orientations to compute these costs.

Next, as the user traces a rough curve, the system continuously recomputes the lowest-cost path between the starting *seed point* and the current mouse location using Dijkstra's al-

gorithm, a breadth-first dynamic programming algorithm that terminates at the current target location.

In order to keep the system from jumping around unpredictably, the system will “freeze” the curve to date (reset the seed point) after a period of inactivity. To prevent the live wire from jumping onto adjacent higher-contrast contours, the system also “learns” the intensity profile under the current optimized curve, and uses this to preferentially keep the wire moving along the same (or a similar looking) boundary (Figure 7.43b–c).

Several extensions have been proposed to the basic algorithm, which works remarkably well even in its original form. Mortensen and Barrett (1999) use *tobogganing*, which is a simple form of watershed region segmentation, to pre-segment the image into regions whose boundaries become candidates for optimized curve paths. The resulting region boundaries are turned into a much smaller graph, where nodes are located wherever three or four regions meet. The Dijkstra algorithm is then run on this reduced graph, resulting in much faster (and often more stable) performance. Another extension to intelligent scissors is to use a probabilistic framework that takes into account the current trajectory of the boundary, resulting in a system called JetStream (Pérez, Blake, and Gangnet 2001).

Instead of re-computing an optimal curve at each time instant, a simpler system can be developed by simply “snapping” the current mouse position to the nearest likely boundary point (Gleicher 1995). Applications of these boundary extraction techniques to image cutting and pasting are presented in Section 10.4.

7.3.2 Level Sets

A limitation of active contours based on parametric curves of the form $\mathbf{f}(s)$, e.g., snakes, B-snakes, and CONDENSATION, is that it is challenging to change the topology of the curve as it evolves (McInerney and Terzopoulos 1999, 2000). Furthermore, if the shape changes dramatically, curve reparameterization may also be required.

An alternative representation for such closed contours is to use a *level set*, where the *zero-crossing(s)* of a *characteristic* (or signed distance (Section 3.3.3)) function define the curve. Level sets evolve to fit and track objects of interest by modifying the underlying *embedding function* (another name for this 2D function) $\phi(x, y)$ instead of the curve $\mathbf{f}(s)$ (Malladi, Sethian, and Vemuri 1995; Sethian 1999; Sapiro 2001; Osher and Paragios 2003). To reduce the amount of computation required, only a small strip (frontier) around the locations of the current zero-crossing needs to be updated at each step, which results in what are called *fast marching methods* (Sethian 1999).

An example of an evolution equation is the *geodesic active contour* proposed by Caselles,

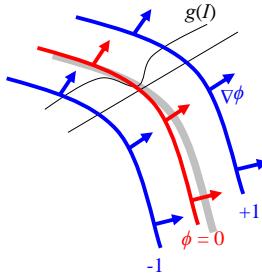


Figure 7.44 Level set evolution for a geodesic active contour. The embedding function ϕ is updated based on the curvature of the underlying surface modulated by the edge/speed function $g(I)$, as well as the gradient of $g(I)$, thereby attracting it to strong edges.

Kimmel, and Sapiro (1997) and Yezzi, Kichenassamy *et al.* (1997),

$$\begin{aligned} \frac{d\phi}{dt} &= |\nabla\phi| \operatorname{div} \left(g(I) \frac{\nabla\phi}{|\nabla\phi|} \right) \\ &= g(I) |\nabla\phi| \operatorname{div} \left(\frac{\nabla\phi}{|\nabla\phi|} \right) + \nabla g(I) \cdot \nabla\phi, \end{aligned} \quad (7.38)$$

where $g(I)$ is a generalized version of the snake edge potential. To get an intuitive sense of the curve's behavior, assume that the embedding function ϕ is a signed distance function away from the curve (Figure 7.44), in which case $|\phi| = 1$. The first term in Equation (7.38) moves the curve in the direction of its curvature, i.e., it acts to straighten the curve, under the influence of the modulation function $g(I)$. The second term moves the curve down the gradient of $g(I)$, encouraging the curve to migrate towards minima of $g(I)$.

While this level-set formulation can readily change topology, it is still susceptible to local minima, since it is based on local measurements such as image gradients. An alternative approach is to re-cast the problem in a segmentation framework, where the energy measures the consistency of the image statistics (e.g., color, texture, motion) inside and outside the segmented regions (Cremers, Rousson, and Deriche 2007; Rousson and Paragios 2008; Houhou, Thiran, and Bresson 2008). These approaches build on earlier energy-based segmentation frameworks introduced by Leclerc (1989), Mumford and Shah (1989), and Chan and Vese (2001), which are discussed in more detail in Section 4.3.2.

For more information on level sets and their applications, please see the collection of papers edited by Osher and Paragios (2003) as well as the series of Workshops on Variational and Level Set Methods in Computer Vision (Paragios, Faugeras *et al.* 2005) and Special Issues on Scale Space and Variational Methods in Computer Vision (Paragios and Sgallari

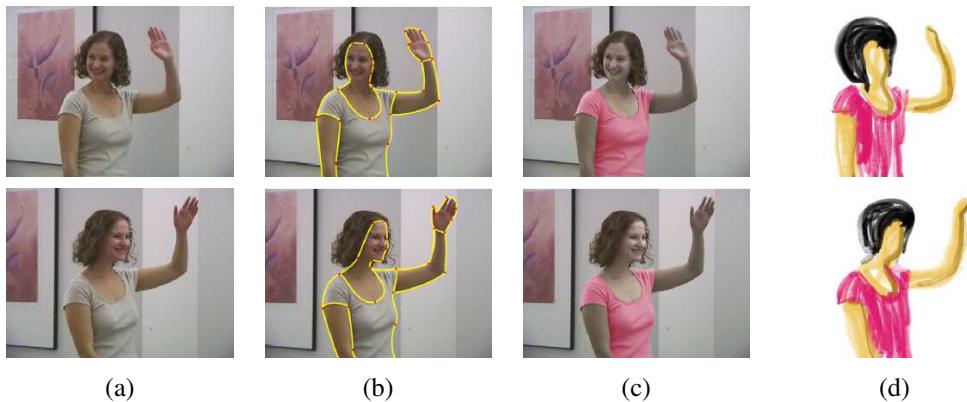


Figure 7.45 Keyframe-based rotoscoping (Agarwala, Hertzmann et al. 2004) © 2004 ACM: (a) original frames; (b) rotoscoped contours; (c) re-colored blouse; (d) rotoscoped hand-drawn animation.

2009).

7.3.3 Application: Contour tracking and rotoscoping

Active contours can be used in a wide variety of object-tracking applications (Blake and Isard 1998; Yilmaz, Javed, and Shah 2006). For example, they can be used to track facial features for performance-driven animation (Terzopoulos and Waters 1990; Lee, Terzopoulos, and Waters 1995; Parke and Waters 1996; Bregler, Covell, and Slaney 1997). They can also be used to track heads and people, as shown in Figure 7.42, as well as moving vehicles (Paragios and Deriche 2000). Additional applications include medical image segmentation, where contours can be tracked from slice to slice in computed tomography (Cootes and Taylor 2001), or over time, as in ultrasound scans.

An interesting application that is closer to computer animation and visual effects is *rotoscoping*, which uses the tracked contours to deform a set of hand-drawn animations (or to modify or replace the original video frames).¹⁴ Agarwala, Hertzmann et al. (2004) present a system based on tracking hand-drawn B-spline contours drawn at selected keyframes, using a combination of geometric and appearance-based criteria (Figure 7.45). They also provide an excellent review of previous rotoscoping and image-based, contour-tracking systems.

Additional applications of rotoscoping (object contour detection and segmentation), such

¹⁴The term comes from a device (a rotoscope) that projected frames of a live-action film underneath an acetate so that artists could draw animations directly over the actors' shapes.

as cutting and pasting objects from one photograph into another, are presented in Section 10.4.

7.4 Lines and vanishing points

While edges and general curves are suitable for describing the contours of natural objects, the human-made world is full of straight lines. Detecting and matching these lines can be useful in a variety of applications, including architectural modeling, pose estimation in urban environments, and the analysis of printed document layouts.

In this section, we present some techniques for extracting *piecewise linear* descriptions from the curves computed in the previous section. We begin with some algorithms for approximating a curve as a piecewise-linear polyline. We then describe the *Hough transform*, which can be used to group edgels into line segments even across gaps and occlusions. Finally, we describe how 3D lines with common *vanishing points* can be grouped together. These vanishing points can be used to calibrate a camera and to determine its orientation relative to a rectahedral scene, as described in Section 11.1.1.

7.4.1 Successive approximation

As we saw in Section 7.2.2, describing a curve as a series of 2D locations $\mathbf{x}_i = \mathbf{x}(s_i)$ provides a general representation suitable for matching and further processing. In many applications, however, it is preferable to approximate such a curve with a simpler representation, e.g., as a piecewise-linear polyline or as a B-spline curve (Farin 2002).

Many techniques have been developed over the years to perform this approximation, which is also known as *line simplification*. One of the oldest, and simplest, is the one proposed by Ramer (1972) and Douglas and Peucker (1973), who recursively subdivide the curve at the point furthest away from the line joining the two endpoints (or the current coarse polyline approximation). Hershberger and Snoeyink (1992) provide a more efficient implementation and also cite some of the other related work in this area.

Once the line simplification has been computed, it can be used to approximate the original curve. If a smoother representation or visualization is desired, either approximating or interpolating splines or curves can be used (Sections 3.5.1 and 7.3.1) (Szeliski and Ito 1986; Bartels, Beatty, and Barsky 1987; Farin 2002).

7.4.2 Hough transforms

While curve approximation with polylines can often lead to successful line extraction, lines in the real world are sometimes broken up into disconnected components or made up of many

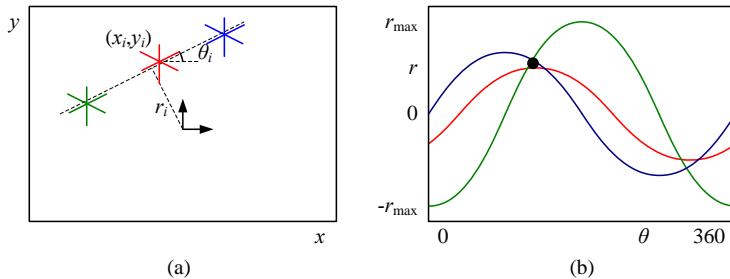


Figure 7.46 Original Hough transform: (a) each point votes for a complete family of potential lines $r_i(\theta) = x_i \cos \theta + y_i \sin \theta$; (b) each pencil of lines sweeps out a sinusoid in (r, θ) ; their intersection provides the desired line equation.

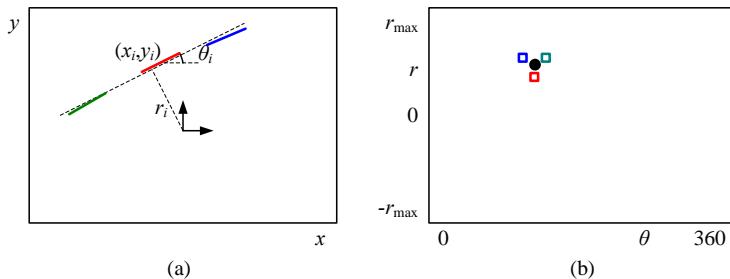


Figure 7.47 Oriented Hough transform: (a) an edgel re-parameterized in polar (r, θ) coordinates, with $\hat{\mathbf{n}}_i = (\cos \theta_i, \sin \theta_i)$ and $r_i = \hat{\mathbf{n}}_i \cdot \mathbf{x}_i$; (b) (r, θ) accumulator array, showing the votes for the three edgels marked in red, green, and blue.

collinear line segments. In many cases, it is desirable to group such collinear segments into extended lines. At a further processing stage (described in Section 7.4.3), we can then group such lines into collections with common vanishing points.

The Hough transform, named after its original inventor (Hough 1962), is a well-known technique for having edges “vote” for plausible line locations (Duda and Hart 1972; Ballard 1981; Illingworth and Kittler 1988). In its original formulation (Figure 7.46), each edge point votes for *all* possible lines passing through it, and lines corresponding to high *accumulator* or *bin* values are examined for potential line fits.¹⁵ Unless the points on a line are truly punctate, a better approach is to use the local orientation information at each edgel to vote for a *single* accumulator cell (Figure 7.47), as described below. A hybrid strategy, where each edgel votes

¹⁵The Hough transform can also be *generalized* to look for other geometric features, such as circles (Ballard 1981).

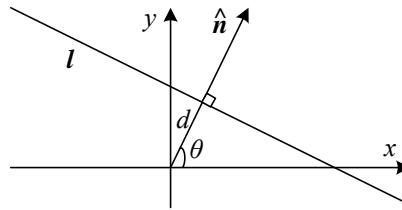


Figure 7.48 2D line equation expressed in terms of the normal $\hat{\mathbf{n}}$ and distance to the origin d .

for a number of possible orientation or location pairs centered around the estimate orientation, may be desirable in some cases.

Before we can vote for line hypotheses, we must first choose a suitable representation. Figure 7.48 (copied from Figure 2.2a) shows the normal-distance $(\hat{\mathbf{n}}, d)$ parameterization for a line. Since lines are made up of edge segments, we adopt the convention that the line normal $\hat{\mathbf{n}}$ points in the same direction (i.e., has the same sign) as the image gradient $\mathbf{J}(\mathbf{x}) = \nabla I(\mathbf{x})$ (7.19). To obtain a minimal two-parameter representation for lines, we convert the normal vector into an angle

$$\theta = \tan^{-1} n_y / n_x, \quad (7.39)$$

as shown in Figure 7.48. The range of possible (θ, d) values is $[-180^\circ, 180^\circ] \times [-\sqrt{2}, \sqrt{2}]$, assuming that we are using normalized pixel coordinates (2.61) that lie in $[-1, 1]$. The number of bins to use along each axis depends on the accuracy of the position and orientation estimate available at each edgel and the expected line density, and is best set experimentally with some test runs on sample imagery.

There are a lot of details in getting the Hough transform to work well, including using edge segment lengths or strengths during the voting process, keeping a list of constituent edgels in the accumulator array for easier post-processing, and optionally combining edges of different “polarity” into the same line segments. These are best worked out by writing an implementation and testing it out on sample data.

An alternative to the 2D polar (θ, d) representation for lines is to use the full 3D $\mathbf{m} = (\hat{\mathbf{n}}, d)$ line equation, projected onto the unit sphere. While the sphere can be parameterized using spherical coordinates (2.8),

$$\hat{\mathbf{m}} = (\cos \theta \cos \phi, \sin \theta \cos \phi, \sin \phi), \quad (7.40)$$

this does not uniformly sample the sphere and still requires the use of trigonometry.

An alternative representation can be obtained by using a *cube map*, i.e., projecting \mathbf{m} onto the face of a unit cube (Figure 7.49a). To compute the cube map coordinate of a 3D vector

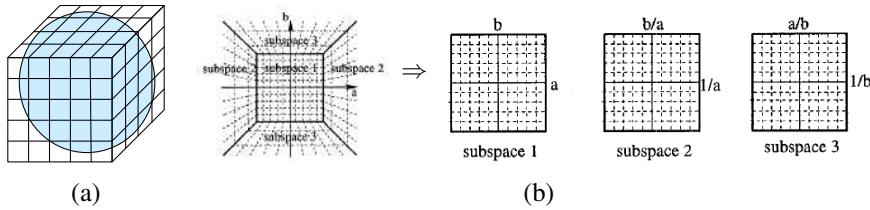


Figure 7.49 Cube map representation for line equations and vanishing points: (a) a cube map surrounding the unit sphere; (b) projecting the half-cube onto three subspaces (Tuytelaars, Van Gool, and Proesmans 1997) © 1997 IEEE.

\mathbf{m} , first find the largest (absolute value) component of \mathbf{m} , i.e., $m = \pm \max(|n_x|, |n_y|, |d|)$, and use this to select one of the six cube faces. Divide the remaining two coordinates by m and use these as indices into the cube face. While this avoids the use of trigonometry, it does require some decision logic.

One advantage of using the cube map, first pointed out by Tuytelaars, Van Gool, and Proesmans (1997), is that all of the lines passing through a point correspond to line segments on the cube faces, which is useful if the original (full voting) variant of the Hough transform is being used. In their work, they represent the line equation as $ax + b + y = 0$, which does not treat the x and y axes symmetrically. Note that if we restrict $d \geq 0$ by ignoring the polarity of the edge orientation (gradient sign), we can use a half-cube instead, which can be represented using only three cube faces, as shown in Figure 7.49b (Tuytelaars, Van Gool, and Proesmans 1997).

RANSAC-based line detection. Another alternative to the Hough transform is the RANdom SAmple Consensus (RANSAC) algorithm described in more detail in Section 8.1.4. In brief, RANSAC randomly chooses pairs of edgels to form a line hypothesis and then tests how many other edgels fall onto this line. (If the edge orientations are accurate enough, a single edgel can produce this hypothesis.) Lines with sufficiently large numbers of *inliers* (matching edgels) are then selected as the desired line segments.

An advantage of RANSAC is that no accumulator array is needed, so the algorithm can be more space efficient and potentially less prone to the choice of bin size. The disadvantage is that many more hypotheses may need to be generated and tested than those obtained by finding peaks in the accumulator array.

Bottom-up grouping. Yet another approach to line segment detection is to iteratively group edgels with similar orientations into oriented rectangular *line-support regions* (Burns, Han-

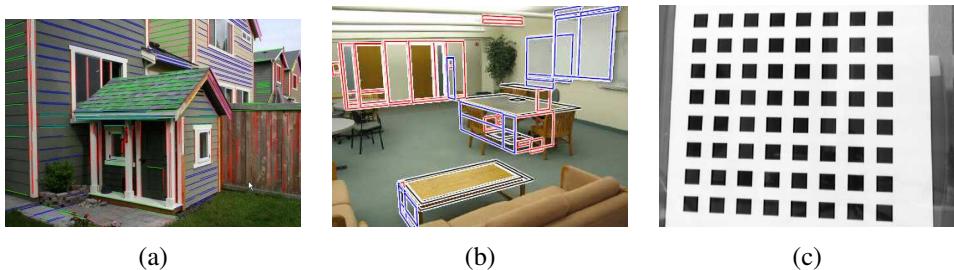


Figure 7.50 Real-world vanishing points: (a) architecture (Sinha, Steedly et al. 2008), (b) furniture (Mičušík, Wildenauer, and Košecká 2008) © 2008 IEEE, and (c) calibration patterns (Zhang 2000).

son, and Riseman 1986). The validity of such regions can then be determined using a statistical analysis, as described in the LSD paper by Grompone von Gioi, Jakubowicz *et al.* (2008). The resulting algorithm is quite fast, does a good job of distinguishing line segments from texture, and is widely used in practice because of its performance and open source availability. Recently, deep neural network algorithms have been developed to simultaneously extract line segments and their junctions (Huang, Wang *et al.* 2018; Zhang, Li *et al.* 2019; Huang, Qin *et al.* 2020; Lin, Pintea, and van Gemert 2020).

In general, there is no clear consensus on which line estimation technique performs best. It is therefore a good idea to think carefully about the problem at hand and to implement several approaches (successive approximation, Hough, and RANSAC) to determine the one that works best for your application.

7.4.3 Vanishing points

In many scenes, structurally important lines have the same vanishing point because they are parallel in 3D. Examples of such lines are horizontal and vertical building edges, zebra crossings, railway tracks, the edges of furniture such as tables and dressers, and of course, the ubiquitous calibration pattern (Figure 7.50). Finding the vanishing points common to such line sets can help refine their position in the image and, in certain cases, help determine the intrinsic and extrinsic orientation of the camera (Section 11.1.1).

Over the years, a large number of techniques have been developed for finding vanishing points (Quan and Mohr 1989; Collins and Weiss 1990; Brillaut-O’Mahoney 1991; McLean and Kotturi 1995; Becker and Bove 1995; Shufelt 1999; Tuytelaars, Van Gool, and Proesmans 1997; Schaffalitzky and Zisserman 2000; Antone and Teller 2002; Rother 2002; Košecká and Zhang 2005; Denis, Elder, and Estrada 2008; Pflugfelder 2008; Tardif 2009; Bazin, Seo *et al.*

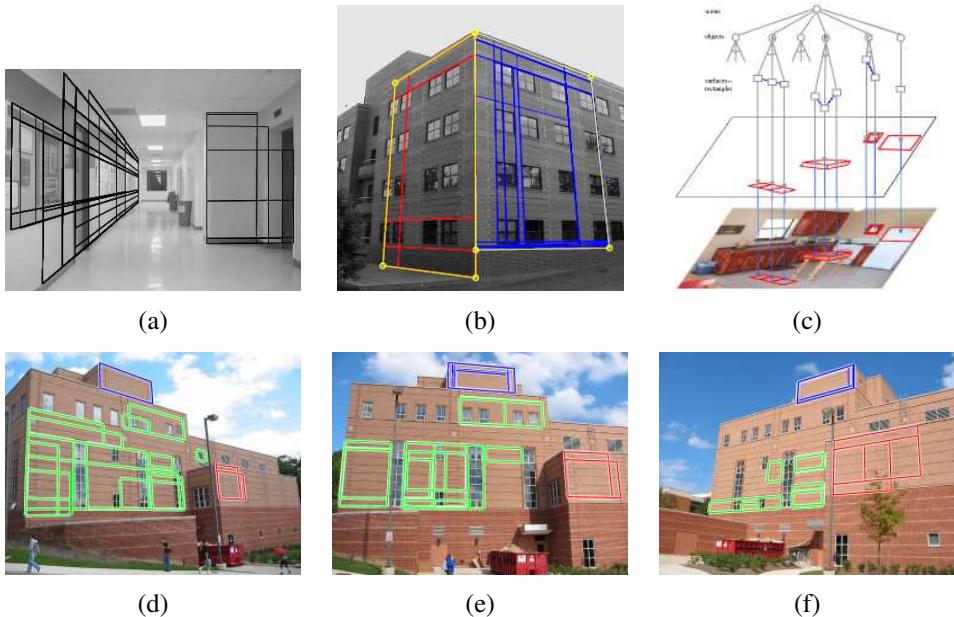


Figure 7.51 *Rectangle detection:* (a) indoor corridor and (b) building exterior with grouped facades (Košecká and Zhang 2005) © 2005 Elsevier; (c) grammar-based recognition (Han and Zhu 2005) © 2005 IEEE; (d–f) rectangle matching using a plane sweep algorithm (Mičušík, Wildenauer, and Košecká 2008) © 2008 IEEE.

2012; Antunes and Barreto 2013; Kluger, Ackermann *et al.* 2017; Zhou, Qi *et al.* 2019a)—see some of the more recent papers for additional references and alternative approaches.

In the first edition of this book (Szeliski 2010, Section 4.3.3), I presented a simple Hough technique based on having line pairs vote for potential vanishing point locations, followed by a robust least squares fitting stage. While my technique proceeds in two discrete stages, better results may be obtained by alternating between assigning lines to vanishing points and refitting the vanishing point locations (Antone and Teller 2002; Košecká and Zhang 2005; Pflugfelder 2008). The results of detecting individual vanishing points can also be made more robust by simultaneously searching for pairs or triplets of mutually orthogonal vanishing points (Shufelt 1999; Antone and Teller 2002; Rother 2002; Sinha, Steedly *et al.* 2008; Li, Kim *et al.* 2020). Some results of such vanishing point detection algorithms can be seen in Figure 7.50. It is also possible to simultaneously detect line segments and their junctions using a neural network (Zhang, Li *et al.* 2019) and to then use these to construct complete 3D wireframe models (Zhou, Qi, and Ma 2019; Zhou, Qi *et al.* 2019b).

Rectangle detection

Once sets of mutually orthogonal vanishing points have been detected, it now becomes possible to search for 3D rectangular structures in the image (Figure 7.51). A variety of techniques have been developed to find such rectangles, primarily focused on architectural scenes (Košecká and Zhang 2005; Han and Zhu 2005; Shaw and Barnes 2006; Mičušík, Wildenauer, and Košecká 2008; Schindler, Krishnamurthy *et al.* 2008).

After detecting orthogonal vanishing directions, Košecká and Zhang (2005) refine the fitted line equations, search for corners near line intersections, and then verify rectangle hypotheses by rectifying the corresponding patches and looking for a preponderance of horizontal and vertical edges (Figures 7.51a–b). In follow-on work, Mičušík, Wildenauer, and Košecká (2008) use a Markov random field (MRF) to disambiguate between potentially overlapping rectangle hypotheses. They also use a plane sweep algorithm to match rectangles between different views (Figures 7.51d–f).

A different approach is proposed by Han and Zhu (2005), who use a grammar of potential rectangle shapes and nesting structures (between rectangles and vanishing points) to infer the most likely assignment of line segments to rectangles (Figure 7.51c). The idea of using regular, repetitive structures as part of the modeling process is now being called *holistic 3D reconstruction* (Zhou, Furukawa, and Ma 2019; Zhou, Furukawa *et al.* 2020; Pintore, Mura *et al.* 2020) and will be discussed in more detail in Section 13.6.1 on modeling 3D architecture.

7.5 Segmentation

Image segmentation is the task of finding groups of pixels that “go together”. In statistics and machine learning, this problem is known as *cluster analysis* or more simply *clustering* and is a widely studied area with hundreds of different algorithms (Jain and Dubes 1988; Kaufman and Rousseeuw 1990; Jain, Duin, and Mao 2000; Jain, Topchy *et al.* 2004; Xu and Wunsch 2005). We’ve already discussed general vector-space clustering algorithms in Section 5.2.1. The main difference between clustering and segmentation is that the former usually ignores pixel layout and neighborhoods, while the latter relies heavily on spatial cues and constraints.

In computer vision, image segmentation is one of the oldest and most widely studied problems (Brice and Fennema 1970; Pavlidis 1977; Riseman and Arbib 1977; Ohlander, Price, and Reddy 1978; Rosenfeld and Davis 1979; Haralick and Shapiro 1985). Early techniques often used region splitting or merging (Brice and Fennema 1970; Horowitz and Pavlidis 1976; Ohlander, Price, and Reddy 1978; Pavlidis and Liow 1990), which correspond to *divisive* and *agglomerative* algorithms (Jain, Topchy *et al.* 2004; Xu and Wunsch 2005), which we intro-

duced in Section 5.2.1. More recent algorithms typically optimize some global criterion, such as intra-region consistency and inter-region boundary lengths or dissimilarity (Leclerc 1989; Mumford and Shah 1989; Shi and Malik 2000; Comaniciu and Meer 2002; Felzenszwalb and Huttenlocher 2004; Cremers, Rousson, and Deriche 2007; Pont-Tuset, Arbeláez *et al.* 2017).

We have already seen examples of image segmentation using image morphology (Section 3.3.3), Markov random fields (Section 4.3), active contours (Section 7.3), and level sets (Section 7.3.2). In the recognition chapter (Section 6.4), we studied *semantic segmentation*, whose goal is to break the image up into semantically labeled regions such as sky, grass, and individual people and animals. In this section, we review some additional techniques for bottom-up general (non-semantic) image segmentation. These include algorithms based on region splitting and merging, graph-based segmentation, and probabilistic aggregation (Section 7.5.1), *mean shift* mode finding (Section 7.5.2), and *normalized cuts* splitting based on pixel similarity metrics (Section 7.5.3). Since many of these algorithms are no longer widely used, a lot of the descriptions have been considerably shortened from those found in the first edition of this book (Szeliski 2010, Chapter 5), where you can find longer descriptions.

Since the literature on image segmentation is so vast, a good way to get a handle on some of the better performing algorithms is to look at experimental comparisons on human-labeled databases (Arbeláez, Maire *et al.* 2011; Pont-Tuset, Arbeláez *et al.* 2017). The best known of these is the Berkeley Segmentation Dataset and Benchmark (Martin, Fowlkes *et al.* 2001), which consists of 1,000 images from a Corel image dataset that were hand-labeled by 30 human subjects, for which Unnikrishnan, Pantofaru, and Hebert (2007) propose new metrics for comparing segmentation algorithms, while Estrada and Jepson (2009) compare four well-known segmentation algorithms. A newer database of foreground and background segmentations, used by Alpert, Galun *et al.* (2007), is also available.

As mentioned in Section 3.3.3, the simplest possible technique for segmenting a grayscale image is to select a threshold and then compute connected components. Unfortunately, a single threshold is rarely sufficient for the whole image because of lighting and intra-object statistical variations.

Region splitting (divisive clustering). Splitting the image into successively finer regions is one of the oldest techniques in computer vision. Ohlander, Price, and Reddy (1978) present such a technique, which first computes a histogram for the whole image and then finds a threshold that best separates the large peaks in the histogram. This process is repeated until regions are either fairly uniform or below a certain size. More recent splitting algorithms often optimize some metric of intra-region similarity and inter-region dissimilarity. These are covered in Sections 4.3.2 and Sections 7.5.3.

Region merging (agglomerative clustering). Region merging techniques also date back to the beginnings of computer vision. Brice and Fennema (1970) use a dual grid for representing boundaries between pixels and merge regions based on their relative boundary lengths and the strength of the visible edges at these boundaries.

A very simple version of pixel-based merging combines adjacent regions whose average color difference is below a threshold or whose regions are too small. Segmenting the image into such *superpixels* (Mori, Ren *et al.* 2004), which are not semantically meaningful, can be a useful pre-processing stage to make higher-level algorithms such as stereo matching (Zitnick, Kang *et al.* 2004; Taguchi, Wilburn, and Zitnick 2008), optical flow (Zitnick, Jojic, and Kang 2005; Brox, Bregler, and Malik 2009), and recognition (Mori, Ren *et al.* 2004; Mori 2005; Gu, Lim *et al.* 2009; Lim, Arbeláez *et al.* 2009) both faster and more robust. It is also possible to combine both splitting and merging by starting with a medium-grain segmentation (in a quadtree representation) and then allowing both merging and splitting operations (Horowitz and Pavlidis 1976; Pavlidis and Liow 1990).

Watershed. A technique related to thresholding, since it operates on a grayscale image, is *watershed* computation (Vincent and Soille 1991). This technique segments an image into several *catchment basins*, which are the regions of an image (interpreted as a height field or landscape) where rain would flow into the same lake. An efficient way to compute such regions is to start flooding the landscape at all of the local minima and to label ridges wherever differently evolving components meet. The whole algorithm can be implemented using a priority queue of pixels and breadth-first search (Vincent and Soille 1991).¹⁶

Since images rarely have dark regions separated by lighter ridges, watershed segmentation is usually applied to a smoothed version of the gradient magnitude image, which also makes it usable with color images. As an alternative, the maximum oriented energy in a steerable filter (3.28–3.29) (Freeman and Adelson 1991) can be used as the basis of the *oriented watershed transform* developed by Arbeláez, Maire *et al.* (2011). Such techniques end up finding smooth regions separated by visible (higher gradient) boundaries. Since such boundaries are what active contours usually follow, active contour algorithms (Mortensen and Barrett 1999; Li, Sun *et al.* 2004) often precompute such a segmentation using either the watershed or the related *tobogganing* technique (Section 7.3.1).

¹⁶A related algorithm can be used to compute maximally stable extremal regions (MSERs) efficiently (Section 7.1.1) (Nistér and Stewénius 2008).

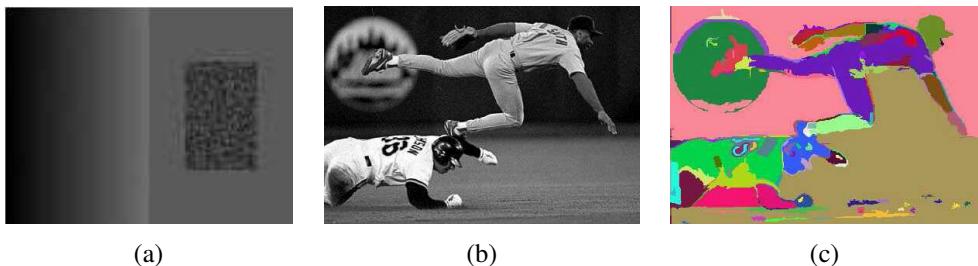


Figure 7.52 Graph-based merging segmentation (Felzenszwalb and Huttenlocher 2004)

© 2004 Springer: (a) input grayscale image that is successfully segmented into three regions even though the variation inside the smaller rectangle is larger than the variation across the middle edge; (b) input grayscale image; (c) resulting segmentation using an N_8 pixel neighborhood.

7.5.1 Graph-based segmentation

While many merging algorithms simply apply a fixed rule that groups pixels and regions together, Felzenszwalb and Huttenlocher (2004) present a merging algorithm that uses *relative dissimilarities* between regions to determine which ones should be merged; it produces an algorithm that provably optimizes a global grouping metric. They start with a pixel-to-pixel dissimilarity measure $w(e)$ that measures, for example, intensity differences between N_8 neighbors. Alternatively, they can use the *joint feature space* distances introduced by Comaniciu and Meer (2002), which we discuss in Sections 7.5.2 and 7.5.3. Figure 7.52 shows two examples of images segmented using their technique.

Probabilistic aggregation

Alpert, Galun *et al.* (2007) develop a probabilistic merging algorithm based on two cues, namely gray-level similarity and texture similarity. The gray-level similarity between regions R_i and R_j is based on the *minimal external difference* from other neighboring regions, which is compared to the *average intensity difference* to compute the likelihoods p_{ij} that two regions should be merged. Merging proceeds in a hierarchical fashion inspired by algebraic multigrid techniques (Brandt 1986; Briggs, Henson, and McCormick 2000) and previously used by Alpert, Galun *et al.* (2007) in their segmentation by weighted aggregation (SWA) algorithm (Sharon, Galun *et al.* 2006). Figure 7.56 shows the segmentations produced by this algorithm compared to other popular segmentation algorithms.

7.5.2 Mean shift

Mean-shift and mode finding techniques, such as k-means and mixtures of Gaussians, model the feature vectors associated with each pixel (e.g., color and position) as samples from an unknown probability density function and then try to find clusters (modes) in this distribution.

Consider the color image shown in Figure 7.53a. How would you segment this image based on color alone? Figure 7.53b shows the distribution of pixels in $L^*u^*v^*$ space, which is equivalent to what a vision algorithm that ignores spatial location would see. To make the visualization simpler, let us only consider the L^*u^* coordinates, as shown in Figure 7.53c. How many obvious (elongated) clusters do you see? How would you go about finding these clusters?

The k-means and mixtures of Gaussians techniques we studied in Section 5.2.2 use a *parametric* model of the density function to answer this question, i.e., they assume the density is the superposition of a small number of simpler distributions (e.g., Gaussians) whose locations (centers) and shape (covariance) can be estimated. Mean shift, on the other hand, smoothes the distribution and finds its peaks as well as the regions of feature space that correspond to each peak. Since a complete density is being modeled, this approach is called *non-parametric* (Bishop 2006).

The key to mean shift is a technique for efficiently finding peaks in this high-dimensional data distribution without ever computing the complete function explicitly (Fukunaga and Hostetler 1975; Cheng 1995; Comaniciu and Meer 2002). Consider once again the data points shown in Figure 7.53c, which can be thought of as having been drawn from some probability density function. If we could compute this density function, as visualized in Figure 7.53e, we could find its major peaks (*modes*) and identify regions of the input space that climb to the same peak as being part of the same region. This is the inverse of the *watershed* algorithm described in Section 7.5, which climbs downhill to find *basins of attraction*.

The first question, then, is how to estimate the density function given a sparse set of samples. One of the simplest approaches is to just smooth the data, e.g., by convolving it with a fixed kernel of width h , which, as we saw in Section 4.1.1, is the *Parzen window* approach to density estimation (Duda, Hart, and Stork 2001, Section 4.3; Bishop 2006, Section 2.5.1). Once we have computed $f(\mathbf{x})$, as shown in Figure 7.53e, we can find its local maxima using gradient ascent or some other optimization technique.

The problem with this “brute force” approach is that, for higher dimensions, it becomes computationally prohibitive to evaluate $f(\mathbf{x})$ over the complete search space. Instead, mean shift uses a variant of what is known in the optimization literature as *multiple restart gradient descent*. Starting at some guess for a local maximum, \mathbf{y}_k , which can be a random input data point \mathbf{x}_i , mean shift computes the gradient of the density estimate $f(\mathbf{x})$ at \mathbf{y}_k and takes an

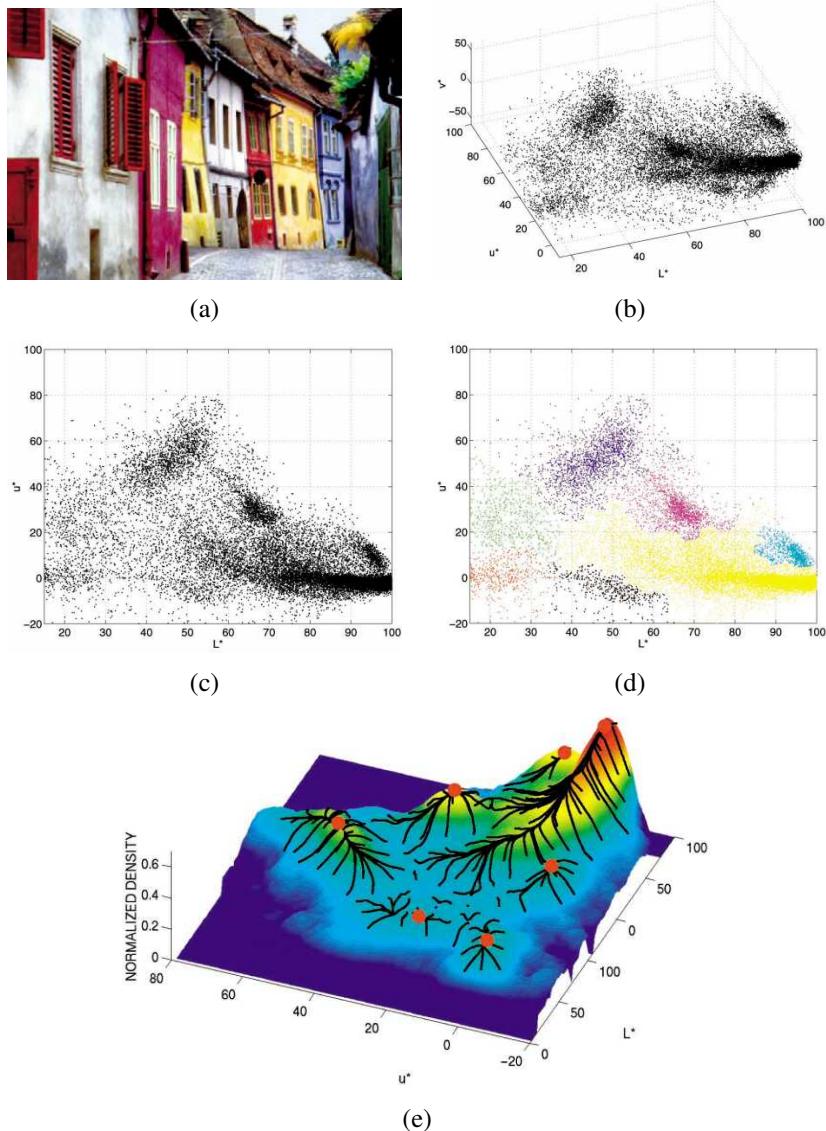


Figure 7.53 Mean-shift image segmentation (Comaniciu and Meer 2002) © 2002 IEEE:
 (a) input color image; (b) pixels plotted in $L^*u^*v^*$ space; (c) L^*u^* space distribution; (d) clustered results after 159 mean-shift procedures; (e) corresponding trajectories with peaks marked as red dots.

uphill step in that direction. Details on how this can be done efficiently can be found in papers on mean shift (Comaniciu and Meer 2002; Paris and Durand 2007) as well as the first edition of this book (Szeliski 2010, Section 5.3.2).

The color-based segmentation shown in Figure 7.53 only looks at pixel colors when determining the best clustering. It may therefore cluster together small isolated pixels that happen to have the same color, which may not correspond to a semantically meaningful segmentation of the image. Better results can usually be obtained by clustering in the *joint domain* of color and location. In this approach, the spatial coordinates of the image $\mathbf{x}_s = (x, y)$, which are called the *spatial domain*, are concatenated with the color values \mathbf{x}_r , which are known as the *range domain*, and mean-shift clustering is applied in this five-dimensional space \mathbf{x}_j . Since location and color may have different scales, the kernels are adjusted separately, just as in the bilateral filter kernel (3.34–3.37) discussed in Section 3.3.2. The difference between mean shift and bilateral filtering, however, is that in mean shift, the spatial coordinates of each pixel are adjusted along with its color values, so that the pixel migrates more quickly towards other pixels with similar colors, and can therefore later be used for clustering and segmentation.

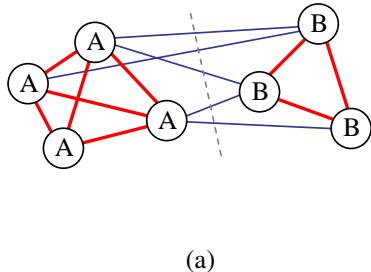
Mean shift has been applied to a number of different problems in computer vision, including face tracking, 2D shape extraction, and texture segmentation (Comaniciu and Meer 2002), stereo matching (Wei and Quan 2004), non-photorealistic rendering (Section 10.5.2) (DeCarlo and Santella 2002), and video editing (Section 10.4.5) (Wang, Bhat *et al.* 2005). Paris and Durand (2007) provide a nice review of such applications, as well as techniques for more efficiently solving the mean-shift equations and producing hierarchical segmentations.

7.5.3 Normalized cuts

While bottom-up merging techniques aggregate regions into coherent wholes and mean-shift techniques try to find clusters of similar pixels using mode finding, the normalized cuts technique introduced by Shi and Malik (2000) examines the *affinities* (similarities) between nearby pixels and tries to separate groups that are connected by weak affinities.

Consider the simple graph shown in Figure 7.54a. The pixels in group *A* are all strongly connected with high affinities, shown as thick red lines, as are the pixels in group *B*. The connections between these two groups, shown as thinner blue lines, are much weaker. A *normalized cut* between the two groups, shown as a dashed line, separates them into two clusters.

The cut between two groups *A* and *B* is defined as the sum of all the weights being cut, where the weights between two pixels (or regions) *i* and *j* measure their similarity. Using a minimum cut as a segmentation criterion, however, does not result in reasonable clusters, since the smallest cuts usually involve isolating a single pixel.



| | A | B | sum |
|-----|---------------|---------------|---------------|
| A | $assoc(A, A)$ | $cut(A, B)$ | $assoc(A, V)$ |
| B | $cut(B, A)$ | $assoc(B, B)$ | $assoc(B, V)$ |
| sum | $assoc(A, V)$ | $assoc(B, v)$ | |

(b)

Figure 7.54 Sample weighted graph and its normalized cut: (a) a small sample graph and its smallest normalized cut; (b) tabular form of the associations and cuts for this graph. The $assoc$ and cut entries are computed as area sums of the associated weight matrix \mathbf{W} . Normalizing the table entries by the row or column sums produces normalized associations and cuts $Nassoc$ and $Ncut$.

A better measure of segmentation is the normalized cut, which is defined as

$$Ncut(A, B) = \frac{cut(A, B)}{assoc(A, V)} + \frac{cut(A, B)}{assoc(B, V)}, \quad (7.41)$$

where $assoc(A, A) = \sum_{i \in A, j \in A} w_{ij}$ is the *association* (sum of all the weights) within a cluster and $assoc(A, V) = assoc(A, A) + cut(A, B)$ is the sum of *all* the weights associated with nodes in A . Figure 7.54b shows how the cuts and associations can be thought of as area sums in the weight matrix $\mathbf{W} = [w_{ij}]$, where the entries of the matrix have been arranged so that the nodes in A come first and the nodes in B come second. Dividing each of these areas by the corresponding row sum (the rightmost column of Figure 7.54b) results in the normalized cut and association values. These normalized values better reflect the fitness of a particular segmentation, since they look for collections of edges that are weak relative to all of the edges both inside and emanating from a particular region.

Unfortunately, computing the optimal normalized cut is NP-complete. Instead, Shi and Malik (2000) suggest computing a real-valued assignment of nodes to groups, using a generalized eigenvalue analysis of the *normalized* affinity matrix (Weiss 1999), as described in more detail in the normalized cuts paper and (Szeliski 2010, Section 5.4). Because these eigenvectors can be interpreted as the large modes of vibration in a spring-mass system, normalized cuts is an example of a *spectral method* for image segmentation. After the real-valued eigenvector is computed, the variables corresponding to positive and negative eigenvector values are associated with the two cut components. This process can be further repeated to hierarchically subdivide an image, as shown in Figure 7.55.

The original algorithm proposed by Shi and Malik (2000) used spatial position and image

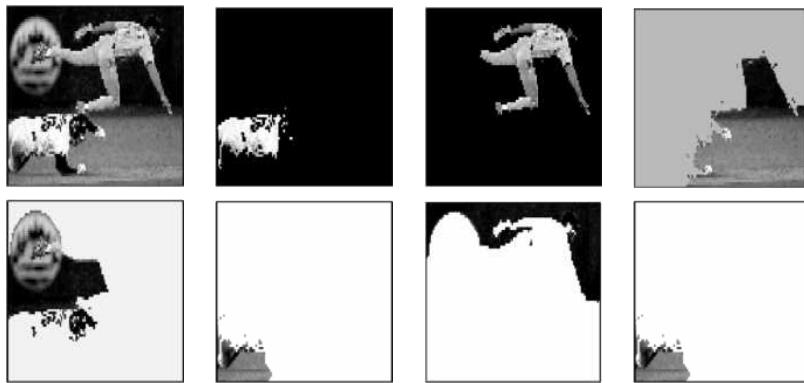


Figure 7.55 Normalized cuts segmentation (Shi and Malik 2000) © 2000 IEEE: The input image and the components returned by the normalized cuts algorithm.

feature differences to compute the pixel-wise affinities. In subsequent work, Malik, Belongie *et al.* (2001) look for *intervening contours* between pixels i and j to define intervening contour weights and then multiply these weights with a texton-based texture similarity metric. They then use an initial over-segmentation based purely on local pixel-wise features to re-estimate intervening contours and texture statistics in a region-based manner. Figure 7.56 shows the results of running this improved algorithm on a number of test images.

Because it requires the solution of large sparse eigenvalue problems, normalized cuts can be quite slow. Sharon, Galun *et al.* (2006) present a way to accelerate the computation of the normalized cuts using an approach inspired by algebraic multigrid (Brandt 1986; Briggs, Henson, and McCormick 2000).

An example of the segmentation produced by weighted aggregation (SWA) is shown in Figure 7.56, along with the most recent probabilistic bottom-up merging algorithm by Alpert, Galun *et al.* (2007). In more recent work, Pont-Tuset, Arbeláez *et al.* (2017) speed up normalized cuts and extend it to multiple scales to obtain state-of-the-art results on both the Berkeley Segmentation Dataset as well as (at the time) object proposals on the VOC and COCO datasets.

7.6 Additional reading

One of the seminal papers on feature detection, description, and matching is by Lowe (2004). Comprehensive surveys and evaluations of such techniques have been made by Schmid, Mohr, and Bauckhage (2000), Mikolajczyk and Schmid (2005), Mikolajczyk, Tuytelaars *et*

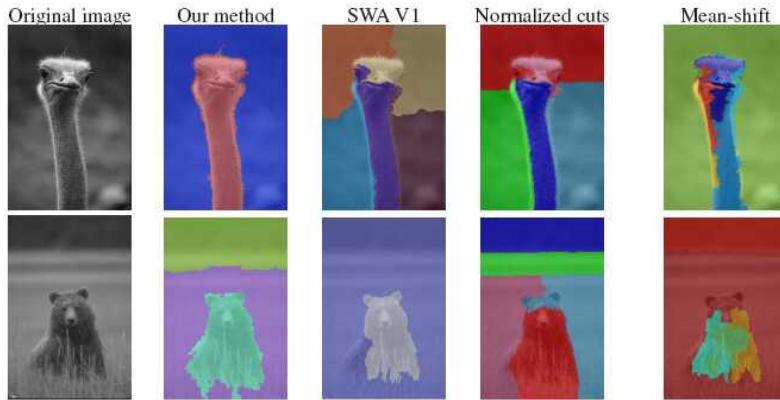


Figure 7.56 Comparative segmentation results (Alpert, Galun et al. 2007) © 2007 IEEE. “Our method” refers to the probabilistic bottom-up merging algorithm developed by Alpert et al.

al. (2005), and Tuytelaars and Mikolajczyk (2008), while Shi and Tomasi (1994) and Triggs (2004) also provide nice reviews.

In the area of feature detectors (Mikolajczyk, Tuytelaars *et al.* 2005), in addition to such classic approaches as Förstner–Harris (Förstner 1986; Harris and Stephens 1988) and difference of Gaussians (Lindeberg 1993, 1998b; Lowe 2004), maximally stable extremal regions (MSERs) are widely used for applications that require affine invariance (Matas, Chum *et al.* 2004; Nistér and Stewénius 2008). More recent interest point detectors are discussed by Xiao and Shah (2003), Koethe (2003), Carneiro and Jepson (2005), Kenney, Zuliani, and Manjunath (2005), Bay, Ess *et al.* (2008), Platel, Balmachnova *et al.* (2006), and Rosten, Porter, and Drummond (2010), as are techniques based on line matching (Zoghliami, Faugeras, and Deriche 1997; Bartoli, Coquerelle, and Sturm 2004) and region detection (Kadir, Zisserman, and Brady 2004; Matas, Chum *et al.* 2004; Tuytelaars and Van Gool 2004; Corso and Hager 2005). Three recent papers with nice reviews of DNN-based feature detectors are Balntas, Lenc *et al.* (2020), Barroso-Laguna, Riba *et al.* (2019), and Tian, Balntas *et al.* (2020).

A variety of local feature descriptors (and matching heuristics) are surveyed and compared by Mikolajczyk and Schmid (2005). More recent publications in this area include those by van de Weijer and Schmid (2006), Abdel-Hakim and Farag (2006), Winder and Brown (2007), and Hua, Brown, and Winder (2007) and the recent evaluations by Balntas, Lenc *et al.* (2020) and Jin, Mishkin *et al.* (2021). Techniques for efficiently matching features include k-d trees (Beis and Lowe 1999; Lowe 2004; Muja and Lowe 2009), pyramid matching kernels (Grauman and Darrell 2005), metric (vocabulary) trees (Nistér and Stewénius 2006),

variety of multi-dimensional hashing techniques (Shakhnarovich, Viola, and Darrell 2003; Torralba, Weiss, and Fergus 2008; Weiss, Torralba, and Fergus 2008; Kulis and Grauman 2009; Raginsky and Lazebnik 2009), and product quantization (Jégou, Douze, and Schmid 2010; Johnson, Douze, and Jégou 2021). A good review of large-scale systems for instance retrieval is Zheng, Yang, and Tian (2018).

The classic reference on feature detection and tracking is Shi and Tomasi (1994). More recent work in this field has focused on learning better matching functions for specific features (Avidan 2001; Jurie and Dhome 2002; Williams, Blake, and Cipolla 2003; Lepetit and Fua 2005; Lepetit, Pilet, and Fua 2006; Hinterstoisser, Benhimane *et al.* 2008; Rogez, Rihan *et al.* 2008; Özysal, Calonder *et al.* 2010).

A highly cited and widely used edge detector is the one developed by Canny (1986). Alternative edge detectors as well as experimental comparisons can be found in publications by Nalwa and Binford (1986), Nalwa (1987), Deriche (1987), Freeman and Adelson (1991), Nalwa (1993), Heath, Sarkar *et al.* (1998), Crane (1997), Ritter and Wilson (2000), Bowyer, Kranenburg, and Dougherty (2001), Arbeláez, Maire *et al.* (2011), and Pont-Tuset, Arbeláez *et al.* (2017). The topic of scale selection in edge detection is nicely treated by Elder and Zucker (1998), while approaches to color and texture edge detection can be found in Ruzon and Tomasi (2001), Martin, Fowlkes, and Malik (2004), and Gevers, van de Weijer, and Stokman (2006). Edge detectors have also been combined with region segmentation techniques to further improve the detection of semantically salient boundaries (Maire, Arbelaez *et al.* 2008; Arbeláez, Maire *et al.* 2011; Xiaofeng and Bo 2012; Pont-Tuset, Arbeláez *et al.* 2017). Edges linked into contours can be smoothed and manipulated for artistic effect (Lowe 1989; Finkelstein and Salesin 1994; Taubin 1995) and used for recognition (Belongie, Malik, and Puzicha 2002; Tek and Kimia 2003; Sebastian and Kimia 2005).

The topic of active contours has a long history, beginning with the seminal work on snakes and other energy-minimizing variational methods (Kass, Witkin, and Terzopoulos 1988; Cootes, Cooper *et al.* 1995; Blake and Isard 1998), continuing through techniques such as intelligent scissors (Mortensen and Barrett 1995, 1999; Pérez, Blake, and Gangnet 2001), and culminating in level sets (Malladi, Sethian, and Vemuri 1995; Caselles, Kimmel, and Sapiro 1997; Sethian 1999; Paragios and Deriche 2000; Sapiro 2001; Osher and Paragios 2003; Paragios, Faugeras *et al.* 2005; Cremers, Rousson, and Deriche 2007; Rousson and Paragios 2008; Paragios and Sgallari 2009), which are currently the most widely used active contour methods.

An early, well-regarded paper on straight line extraction in images was written by Burns, Hanson, and Riseman (1986). Their idea of bottom-up *line-support regions* was extended by Grompone von Gioi, Jakubowicz *et al.* (2008) to construct the popular LSD line segment

detector. The literature on vanishing point detection is quite vast and still evolving (Quan and Mohr 1989; Collins and Weiss 1990; Brillaut-O’Mahoney 1991; McLean and Kotturi 1995; Becker and Bove 1995; Shufelt 1999; Tuytelaars, Van Gool, and Proesmans 1997; Schaffalitzky and Zisserman 2000; Antone and Teller 2002; Rother 2002; Košecká and Zhang 2005; Denis, Elder, and Estrada 2008; Pflugfelder 2008; Tardif 2009; Bazin, Seo *et al.* 2012; Antunes and Barreto 2013; Zhou, Qi *et al.* 2019a). Simultaneous line and junction detection techniques have also been developed (Huang, Wang *et al.* 2018; Zhang, Li *et al.* 2019).

The topic of image segmentation is closely related to clustering techniques, which are treated in a number of monographs and review articles (Jain and Dubes 1988; Kaufman and Rousseeuw 1990; Jain, Duin, and Mao 2000; Jain, Topchy *et al.* 2004). Some early segmentation techniques include those described by Brice and Fennema (1970), Pavlidis (1977), Riseman and Arbib (1977), Ohlander, Price, and Reddy (1978), Rosenfeld and Davis (1979), and Haralick and Shapiro (1985), while examples of newer techniques are developed by Leclerc (1989), Mumford and Shah (1989), Shi and Malik (2000), and Felzenszwalb and Huttenlocher (2004).

Arbeláez, Maire *et al.* (2011) and Pont-Tuset, Arbeláez *et al.* (2017) provide good reviews of automatic segmentation techniques and compare their performance on the Berkeley Segmentation Dataset and Benchmark (Martin, Fowlkes *et al.* 2001).¹⁷ Additional comparison papers and databases include those by Unnikrishnan, Pantofaru, and Hebert (2007), Alpert, Galun *et al.* (2007), and Estrada and Jepson (2009).

Techniques for segmenting images based on local pixel similarities combined with aggregation or splitting methods include watersheds (Vincent and Soille 1991; Beare 2006; Arbeláez, Maire *et al.* 2011), region splitting (Ohlander, Price, and Reddy 1978), region merging (Brice and Fennema 1970; Pavlidis and Liow 1990; Jain, Topchy *et al.* 2004), as well as graph-based and probabilistic multi-scale approaches (Felzenszwalb and Huttenlocher 2004; Alpert, Galun *et al.* 2007).

Mean-shift algorithms, which find modes (peaks) in a density function representation of the pixels, are presented by Comaniciu and Meer (2002) and Paris and Durand (2007). Parametric mixtures of Gaussians can also be used to represent and segment such pixel densities (Bishop 2006; Ma, Derksen *et al.* 2007).

The seminal work on spectral (eigenvalue) methods for image segmentation is the *normalized cut* algorithm of Shi and Malik (2000). Related work includes that by Weiss (1999), Meilă and Shi (2000), Meilă and Shi (2001), Malik, Belongie *et al.* (2001), Ng, Jordan, and Weiss (2001), Yu and Shi (2003), Cour, Bénézit, and Shi (2005), Sharon, Galun *et al.* (2006), Tolliver and Miller (2006), and Wang and Oliensis (2010).

¹⁷<http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/segbench>.

7.7 Exercises

Ex 7.1: Interest point detector. Implement one or more keypoint detectors and compare their performance (with your own or with a classmate’s detector).

Possible detectors:

- Laplacian or Difference of Gaussian;
- Förstner–Harris Hessian (try different formula variants given in (7.9–7.11));
- oriented/steerable filter, looking for either second-order high second response or two edges in a window (Koethe 2003), as discussed in Section 7.1.1.
- any of the newer DNN-based detectors.

Other detectors are described in Mikolajczyk, Tuytelaars *et al.* (2005), Tuytelaars and Mikolajczyk (2008), and Balntas, Lenc *et al.* (2020). Additional optional steps could include:

1. Compute the detections on a sub-octave pyramid and find 3D maxima.
2. Find local orientation estimates using steerable filter responses or a gradient histogramming method.
3. Implement non-maximal suppression, such as the adaptive technique of Brown, Szeliski, and Winder (2005).
4. Vary the window shape and size (prefilter and aggregation).

To test for repeatability, download the code from <https://www.robots.ox.ac.uk/~vgg/research/affine> (Mikolajczyk, Tuytelaars *et al.* 2005; Tuytelaars and Mikolajczyk 2008) or simply rotate or shear your own test images. (Pick a domain you may want to use later, e.g., for outdoor stitching.)

Be sure to measure and report the stability of your scale and orientation estimates.

Ex 7.2: Interest point descriptor. Implement two or more descriptors from Section 7.1.2 (steered to local scale and orientation estimates, if appropriate) and compare their performance on some images of your own choosing.

You can either use the evaluation methodologies (and optionally software) described in Mikolajczyk and Schmid (2005), Balntas, Lenc *et al.* (2020), or Jin, Mishkin *et al.* (2021).

Ex 7.3: ROC curve computation. Given a pair of curves (histograms) plotting the number of matching and non-matching features as a function of Euclidean distance d as shown in

Figure 7.22b, derive an algorithm for plotting a ROC curve (Figure 7.22a). In particular, let $t(d)$ be the distribution of true matches and $f(d)$ be the distribution of (false) non-matches. Write down the equations for the ROC, i.e., TPR(FPR), and the AUC.

(Hint: Plot the cumulative distributions $T(d) = \int t(d)$ and $F(d) = \int f(d)$ and see if these help you derive the TPR and FPR at a given threshold θ .)

Ex 7.4: Feature matcher. After extracting features from a collection of overlapping or distorted images,¹⁸ match them up by their descriptors either using nearest neighbor matching or a more efficient matching strategy such as a k-d tree.

See whether you can improve the accuracy of your matches using techniques such as the nearest neighbor distance ratio.

Ex 7.5: Feature tracker. Instead of finding feature points independently in multiple images and then matching them, find features in the first image of a video or image sequence and then re-locate the corresponding points in the next frames using either search and gradient descent (Shi and Tomasi 1994) or learned feature detectors (Lepetit, Pilet, and Fua 2006; Fossati, Dimitrijevic *et al.* 2007). When the number of tracked points drops below a threshold or new regions in the image become visible, find additional points to track.

(Optional) Winnow out incorrect matches by estimating a homography (8.19–8.23) or fundamental matrix (Section 11.3.3).

(Optional) Refine the accuracy of your matches using the iterative registration algorithm described in Section 9.2 and Exercise 9.2.

Ex 7.6: Facial feature tracker. Apply your feature tracker to tracking points on a person's face, either manually initialized to interesting locations such as eye corners or automatically initialized at interest points.

(Optional) Match features between two people and use these features to perform image morphing (Exercise 3.25).

Ex 7.7: Edge detector. Implement an edge detector of your choice. Compare its performance to that of your classmates' detectors or code downloaded from the internet.

A simple but well-performing sub-pixel edge detector can be created as follows:

1. Blur the input image a little,

$$B_\sigma(\mathbf{x}) = G_\sigma(\mathbf{x}) * I(\mathbf{x}).$$

¹⁸<https://www.robots.ox.ac.uk/~vgg/research/affine>.

2. Construct a Gaussian pyramid (Exercise 3.17),

$$P = \text{Pyramid}\{B_\sigma(\mathbf{x})\}$$

3. Subtract an interpolated coarser-level pyramid image from the original resolution blurred image,

$$S(\mathbf{x}) = B_\sigma(\mathbf{x}) - P.\text{InterpolatedLevel}(L).$$

4. For each quad of pixels, $\{(i, j), (i + 1, j), (i, j + 1), (i + 1, j + 1)\}$, count the number of zero crossings along the four edges.
5. When there are exactly two zero crossings, compute their locations using (7.25) and store these edgel endpoints along with the midpoint in the edgel structure.
6. For each edgel, compute the local gradient by taking the horizontal and vertical differences between the values of S along the zero crossing edges.
7. Store the magnitude of this gradient as the edge strength and either its orientation or that of the segment joining the edgel endpoints as the edge orientation.
8. Add the edgel to a list of edgels or store it in a 2D array of edgels (addressed by pixel coordinates).

Ex 7.8: Edge linking and thresholding. Link up the edges computed in the previous exercise into chains and optionally perform thresholding with hysteresis.

The steps may include:

1. Store the edgels either in a 2D array (say, an integer image with indices into the edgel list) or pre-sort the edgel list first by (integer) x coordinates and then y coordinates, for faster neighbor finding.
2. Pick up an edgel from the list of unlinked edgels and find its neighbors in both directions until no neighbor is found or a closed contour is obtained. Flag edgels as linked as you visit them and push them onto your list of linked edgels.
3. (Optional) Perform hysteresis-based thresholding (Canny 1986). Use two thresholds “hi” and “lo” for the edge strength. A candidate edgel is considered an edge if either its strength is above the “hi” threshold or its strength is above the “lo” threshold and it is (recursively) connected to a previously detected edge.
4. (Optional) Link together contours that have small gaps but whose endpoints have similar orientations.

5. (Optional) Find junctions between adjacent contours, e.g., using some of the ideas (or references) from Maire, Arbelaez *et al.* (2008).

Ex 7.9: Contour matching. Convert a closed contour (linked edgel list) into its arc-length parameterization and use this to match object outlines.

The steps may include:

1. Walk along the contour and create a list of (x_i, y_i, s_i) triplets, using the arc-length formula

$$s_{i+1} = s_i + \|\mathbf{x}_{i+1} - \mathbf{x}_i\|. \quad (7.42)$$

2. Resample this list onto a regular set of (x_j, y_j, j) samples using linear interpolation of each segment.
3. Compute the average values of x and y , i.e., \bar{x} and \bar{y} and subtract them from your sampled curve points.
4. Resample the original (x_i, y_i, s_i) piecewise-linear function onto a length-independent set of samples, say $j \in [0, 1023]$. (Using a length which is a power of two makes subsequent Fourier transforms more convenient.)
5. Compute the Fourier transform of the curve, treating each (x, y) pair as a complex number.
6. To compare two curves, fit a linear equation to the phase difference between the two curves. (Careful: phase wraps around at 360° . Also, you may wish to weight samples by their Fourier spectrum magnitude—see Section 9.1.2.)
7. (Optional) Prove that the constant phase component corresponds to the temporal shift in s , while the linear component corresponds to rotation.

Of course, feel free to try any other curve descriptor and matching technique from the computer vision literature (Tek and Kimia 2003; Sebastian and Kimia 2005).

Ex 7.10: Jigsaw puzzle solver—challenging. Write a program to automatically solve a jigsaw puzzle from a set of scanned puzzle pieces. Your software may include the following components:

1. Scan the pieces (either face up or face down) on a flatbed scanner with a distinctively colored background.
2. (Optional) Scan in the box top to use as a low-resolution reference image.

3. Use color-based thresholding to isolate the pieces.
4. Extract the contour of each piece using edge finding and linking.
5. (Optional) Re-represent each contour using an arc-length or some other re-parameterization.
Break up the contours into meaningful matchable pieces. (Is this hard?)
6. (Optional) Associate color values with each contour to help in the matching.
7. (Optional) Match pieces to the reference image using some rotationally invariant feature descriptors.
8. Solve a global optimization or (backtracking) search problem to snap pieces together and place them in the correct location relative to the reference image.
9. Test your algorithm on a succession of more difficult puzzles and compare your results with those of others.

For some additional ideas, have a look at Cho, Avidan, and Freeman (2010).

Ex 7.11: Successive approximation line detector. Implement a line simplification algorithm (Section 7.4.1) (Ramer 1972; Douglas and Peucker 1973) to convert a hand-drawn curve (or linked edge image) into a small set of polylines.

(Optional) Re-render this curve using either an approximating or interpolating spline or Bezier curve (Szeliski and Ito 1986; Bartels, Beatty, and Barsky 1987; Farin 2002).

Ex 7.12: Line fitting uncertainty. Estimate the uncertainty (covariance) in your line fit using uncertainty analysis.

1. After determining which edgels belong to the line segment (using either successive approximation or Hough transform), re-fit the line segment using total least squares (Van Huffel and Vandewalle 1991; Van Huffel and Lemmerling 2002), i.e., find the mean or centroid of the edgels and then use eigenvalue analysis to find the dominant orientation.
2. Compute the perpendicular errors (deviations) to the line and robustly estimate the variance of the fitting noise using an estimator such as MAD (Appendix B.3).
3. (Optional) re-fit the line parameters by throwing away outliers or using a robust norm or influence function.
4. Estimate the error in the perpendicular location of the line segment and its orientation.

Ex 7.13: Vanishing points. Compute the vanishing points in an image using one of the techniques described in Section 7.4.3 and optionally refine the original line equations associated with each vanishing point. Your results can be used later to track a target or reconstruct architecture (Section 13.6.1).

Ex 7.14: Vanishing point uncertainty. Perform an uncertainty analysis on your estimated vanishing points. You will need to decide how to represent your vanishing point, e.g., homogeneous coordinates on a sphere, to handle vanishing points near infinity.

See the discussion of Bingham distributions by Collins and Weiss (1990) for some ideas.

Ex 7.15: Region segmentation. Implement one of the region segmentation algorithms described in this chapter. Some popular segmentation algorithms include:

- k-means (Section 5.2.2);
- mixtures of Gaussians (Section 5.2.2);
- mean shift (Section 7.5.2);
- normalized cuts (Section 7.5.3);
- similarity graph-based segmentation (Section 7.5.1);
- binary Markov random fields solved using graph cuts (Section 4.3.2).

Apply your region segmentation to a video sequence and use it to track moving regions from frame to frame.

Alternatively, test out your segmentation algorithm on the Berkeley segmentation database (Martin, Fowlkes *et al.* 2001).

Chapter 8

Image alignment and stitching

| | | |
|-------|---|-----|
| 8.1 | Pairwise alignment | 503 |
| 8.1.1 | 2D alignment using least squares | 504 |
| 8.1.2 | <i>Application:</i> Panography | 506 |
| 8.1.3 | Iterative algorithms | 507 |
| 8.1.4 | Robust least squares and RANSAC | 510 |
| 8.1.5 | 3D alignment | 513 |
| 8.2 | Image stitching | 514 |
| 8.2.1 | Parametric motion models | 516 |
| 8.2.2 | <i>Application:</i> Whiteboard and document scanning | 517 |
| 8.2.3 | Rotational panoramas | 519 |
| 8.2.4 | Gap closing | 520 |
| 8.2.5 | <i>Application:</i> Video summarization and compression | 522 |
| 8.2.6 | Cylindrical and spherical coordinates | 523 |
| 8.3 | Global alignment | 526 |
| 8.3.1 | Bundle adjustment | 527 |
| 8.3.2 | Parallax removal | 531 |
| 8.3.3 | Recognizing panoramas | 533 |
| 8.4 | Compositing | 536 |
| 8.4.1 | Choosing a compositing surface | 536 |
| 8.4.2 | Pixel selection and weighting (deghosting) | 538 |
| 8.4.3 | <i>Application:</i> Photomontage | 544 |
| 8.4.4 | Blending | 544 |
| 8.5 | Additional reading | 547 |
| 8.6 | Exercises | 549 |



Figure 8.1 Image stitching: (a) geometric alignment of 2D images for stitching (Szeliski and Shum 1997) © 1997 ACM; (b) a spherical panorama constructed from 54 photographs (Szeliski and Shum 1997) © 1997 ACM; (c) a multi-image panorama automatically assembled from an unordered photo collection; a multi-image stitch (d) without and (e) with moving object removal (Uyttendaele, Eden, and Szeliski 2001) © 2001 IEEE.

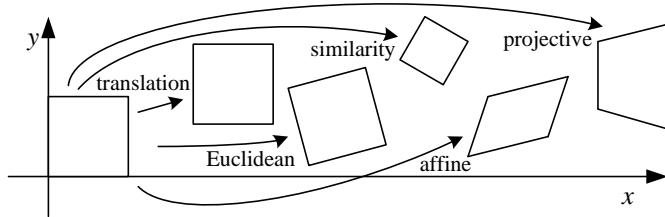


Figure 8.2 Basic set of 2D planar transformations

Once we have extracted features from images, the next stage in many vision algorithms is to match these features across different images (Section 7.1.3). An important component of this matching is to verify whether the set of matching features is geometrically consistent, e.g., whether the feature displacements can be described by a simple 2D or 3D geometric transformation. The computed motions can then be used in other applications such as image stitching (Section 8.2) or augmented reality (Section 11.2.2).

In this chapter, we look at the topic of geometric image registration, i.e., the computation of 2D and 3D transformations that map features in one image to another (Section 8.1). In Chapter 11, we look at the related problems of *pose estimation*, which is determining a camera's position relative to a known 3D object or scene, and *structure from motion*, i.e., how to simultaneously estimate 3D geometry and camera motion.

8.1 Pairwise alignment

Feature-based alignment is the problem of estimating the motion between two or more sets of matched 2D or 3D points. In this section, we restrict ourselves to global *parametric* transformations, such as those described in Section 2.1.1 and shown in Table 2.1 and Figure 8.2, or higher order transformation for curved surfaces (Shashua and Toelg 1997; Can, Stewart *et al.* 2002). Applications to non-rigid or elastic deformations (Bookstein 1989; Kambhamettu, Goldgof *et al.* 1994; Szeliski and Lavallée 1996; Torresani, Hertzmann, and Bregler 2008) are examined in Sections 9.2.2 and 13.6.4.

| Transform | Matrix | Parameters \mathbf{p} | Jacobian \mathbf{J} |
|-------------|---|--|---|
| translation | $\begin{bmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \end{bmatrix}$ | (t_x, t_y) | $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ |
| Euclidean | $\begin{bmatrix} c_\theta & -s_\theta & t_x \\ s_\theta & c_\theta & t_y \end{bmatrix}$ | (t_x, t_y, θ) | $\begin{bmatrix} 1 & 0 & -s_\theta x - c_\theta y \\ 0 & 1 & c_\theta x - s_\theta y \end{bmatrix}$ |
| similarity | $\begin{bmatrix} 1+a & -b & t_x \\ b & 1+a & t_y \end{bmatrix}$ | (t_x, t_y, a, b) | $\begin{bmatrix} 1 & 0 & x & -y \\ 0 & 1 & y & x \end{bmatrix}$ |
| affine | $\begin{bmatrix} 1+a_{00} & a_{01} & t_x \\ a_{10} & 1+a_{11} & t_y \end{bmatrix}$ | $(t_x, t_y, a_{00}, a_{01}, a_{10}, a_{11})$ | $\begin{bmatrix} 1 & 0 & x & y & 0 & 0 \\ 0 & 1 & 0 & 0 & x & y \end{bmatrix}$ |
| projective | $\begin{bmatrix} 1+h_{00} & h_{01} & h_{02} \\ h_{10} & 1+h_{11} & h_{12} \\ h_{20} & h_{21} & 1 \end{bmatrix}$ | $(h_{00}, h_{01}, \dots, h_{21})$ | (see Section 8.1.3) |

Table 8.1 Jacobians of the 2D coordinate transformations $\mathbf{x}' = \mathbf{f}(\mathbf{x}; \mathbf{p})$ shown in Table 2.1, where we have re-parameterized the motions so that they are identity for $\mathbf{p} = 0$.

8.1.1 2D alignment using least squares

Given a set of matched feature points $\{(\mathbf{x}_i, \mathbf{x}'_i)\}$ and a planar parametric transformation¹ of the form

$$\mathbf{x}' = \mathbf{f}(\mathbf{x}; \mathbf{p}), \quad (8.1)$$

how can we produce the best estimate of the motion parameters \mathbf{p} ? The usual way to do this is to use least squares, i.e., to minimize the sum of squared residuals

$$E_{\text{LS}} = \sum_i \|\mathbf{r}_i\|^2 = \sum_i \|\mathbf{f}(\mathbf{x}_i; \mathbf{p}) - \mathbf{x}'_i\|^2, \quad (8.2)$$

where

$$\mathbf{r}_i = \mathbf{x}'_i - \mathbf{f}(\mathbf{x}_i; \mathbf{p}) = \hat{\mathbf{x}}'_i - \tilde{\mathbf{x}}'_i \quad (8.3)$$

is the *residual* between the measured location $\hat{\mathbf{x}}'_i$ and its corresponding current *predicted* location $\tilde{\mathbf{x}}'_i = \mathbf{f}(\mathbf{x}_i; \mathbf{p})$. (See Appendix A.2 for more on least squares and Appendix B.2 for a statistical justification.)

¹For examples of non-planar parametric models, such as quadrics, see the work of Shashua and Toelg (1997) and Shashua and Wexler (2001).

Many of the motion models presented in Section 2.1.1 and Table 2.1, i.e., translation, similarity, and affine, have a *linear* relationship between the amount of motion $\Delta\mathbf{x} = \mathbf{x}' - \mathbf{x}$ and the unknown parameters \mathbf{p} ,

$$\Delta\mathbf{x} = \mathbf{x}' - \mathbf{x} = \mathbf{J}(\mathbf{x})\mathbf{p}, \quad (8.4)$$

where $\mathbf{J} = \partial\mathbf{f}/\partial\mathbf{p}$ is the *Jacobian* of the transformation \mathbf{f} with respect to the motion parameters \mathbf{p} (see Table 8.1). In this case, a simple *linear* regression (linear least squares problem) can be formulated as

$$E_{\text{LLS}} = \sum_i \|\mathbf{J}(\mathbf{x}_i)\mathbf{p} - \Delta\mathbf{x}_i\|^2 \quad (8.5)$$

$$= \mathbf{p}^T \left[\sum_i \mathbf{J}^T(\mathbf{x}_i) \mathbf{J}(\mathbf{x}_i) \right] \mathbf{p} - 2\mathbf{p}^T \left[\sum_i \mathbf{J}^T(\mathbf{x}_i) \Delta\mathbf{x}_i \right] + \sum_i \|\Delta\mathbf{x}_i\|^2 \quad (8.6)$$

$$= \mathbf{p}^T \mathbf{A} \mathbf{p} - 2\mathbf{p}^T \mathbf{b} + c. \quad (8.7)$$

The minimum can be found by solving the symmetric positive definite (SPD) system of *normal equations*²

$$\mathbf{A}\mathbf{p} = \mathbf{b}, \quad (8.8)$$

where

$$\mathbf{A} = \sum_i \mathbf{J}^T(\mathbf{x}_i) \mathbf{J}(\mathbf{x}_i) \quad (8.9)$$

is called the *Hessian* and $\mathbf{b} = \sum_i \mathbf{J}^T(\mathbf{x}_i) \Delta\mathbf{x}_i$. For the case of pure translation, the resulting equations have a particularly simple form, i.e., the translation is the average translation between corresponding points or, equivalently, the translation of the point centroids.

Uncertainty weighting. The above least squares formulation assumes that all feature points are matched with the same accuracy. This is often not the case, since certain points may fall into more textured regions than others. If we associate a scalar variance estimate σ_i^2 with each correspondence, we can minimize the *weighted least squares* problem instead,³

$$E_{\text{WLS}} = \sum_i \sigma_i^{-2} \|\mathbf{r}_i\|^2. \quad (8.10)$$

²For poorly conditioned problems, it is better to use QR decomposition on the set of linear equations $\mathbf{J}(\mathbf{x}_i)\mathbf{p} = \Delta\mathbf{x}_i$ instead of the normal equations (Björck 1996; Golub and Van Loan 1996). However, such conditions rarely arise in image registration.

³Problems where each measurement can have a different variance or uncertainty are called *heteroscedastic models*.



Figure 8.3 A simple panograph consisting of three images automatically aligned with a translational model and then averaged together.

As shown in Section 9.1.3, a covariance estimate for patch-based matching can be obtained by multiplying the inverse of the *patch Hessian* \mathbf{A}_i (9.48) with the per-pixel noise covariance σ_n^2 (9.37). Weighting each squared residual by its inverse covariance $\Sigma_i^{-1} = \sigma_n^{-2} \mathbf{A}_i$ (which is called the *information matrix*), we obtain

$$E_{\text{CWLS}} = \sum_i \|\mathbf{r}_i\|_{\Sigma_i^{-1}}^2 = \sum_i \mathbf{r}_i^T \Sigma_i^{-1} \mathbf{r}_i = \sum_i \sigma_n^{-2} \mathbf{r}_i^T \mathbf{A}_i \mathbf{r}_i. \quad (8.11)$$

8.1.2 Application: Panography

One of the simplest (and most fun) applications of image alignment is a special form of image stitching called *panography*. In a panograph, images are translated and optionally rotated and scaled before being blended with simple averaging (Figure 8.3). This process mimics the photographic collages created by artist David Hockney, although his compositions use an opaque overlay model, being created out of regular photographs.

In most of the examples seen on the web, the images are aligned by hand for best artistic effect.⁴ However, it is also possible to use feature matching and alignment techniques to perform the registration automatically (Nomura, Zhang, and Nayar 2007; Zelnik-Manor and Perona 2007).

Consider a simple translational model. We want all the corresponding features in different images to line up as best as possible. Let \mathbf{t}_j be the location of the j th image coordinate frame in the global composite frame and \mathbf{x}_{ij} be the location of the i th matched feature in the j th

⁴<https://www.flickr.com/groups/panography/>.

image. In order to align the images, we wish to minimize the least squares error

$$E_{\text{PLS}} = \sum_{ij} \|(\mathbf{t}_j + \mathbf{x}_{ij}) - \mathbf{x}_i\|^2, \quad (8.12)$$

where \mathbf{x}_i is the consensus (average) position of feature i in the global coordinate frame. (An alternative approach is to register each pair of overlapping images separately and then compute a consensus location for each frame—see Exercise 8.2.)

The above least squares problem is indeterminate (you can add a constant offset to all the frame and point locations \mathbf{t}_j and \mathbf{x}_i). To fix this, either pick one frame as being at the origin or add a constraint to make the average frame offsets be 0.

The formulas for adding rotation and scale transformations are straightforward and are left as an exercise (Exercise 8.2). See if you can create some collages that you would be happy to share with others on the web.

8.1.3 Iterative algorithms

While linear least squares is the simplest method for estimating parameters, most problems in computer vision do not have a simple linear relationship between the measurements and the unknowns. In this case, the resulting problem is called *non-linear least squares* or *non-linear regression*.

Consider, for example, the problem of estimating a rigid Euclidean 2D transformation (translation plus rotation) between two sets of points. If we parameterize this transformation by the translation amount (t_x, t_y) and the rotation angle θ , as in Table 2.1, the Jacobian of this transformation, given in Table 8.1, depends on the current value of θ . Notice how in Table 8.1, we have re-parameterized the motion matrices so that they are always the identity at the origin $\mathbf{p} = 0$, which makes it easier to initialize the motion parameters.

To minimize the non-linear least squares problem, we iteratively find an update $\Delta\mathbf{p}$ to the current parameter estimate \mathbf{p} by minimizing

$$E_{\text{NLS}}(\Delta\mathbf{p}) = \sum_i \|\mathbf{f}(\mathbf{x}_i; \mathbf{p} + \Delta\mathbf{p}) - \mathbf{x}'_i\|^2 \quad (8.13)$$

$$\approx \sum_i \|\mathbf{J}(\mathbf{x}_i; \mathbf{p})\Delta\mathbf{p} - \mathbf{r}_i\|^2 \quad (8.14)$$

$$= \Delta\mathbf{p}^T \left[\sum_i \mathbf{J}^T \mathbf{J} \right] \Delta\mathbf{p} - 2\Delta\mathbf{p}^T \left[\sum_i \mathbf{J}^T \mathbf{r}_i \right] + \sum_i \|\mathbf{r}_i\|^2 \quad (8.15)$$

$$= \Delta\mathbf{p}^T \mathbf{A} \Delta\mathbf{p} - 2\Delta\mathbf{p}^T \mathbf{b} + c, \quad (8.16)$$

where the “Hessian”⁵ \mathbf{A} is the same as Equation (8.9) and the right-hand side vector

$$\mathbf{b} = \sum_i \mathbf{J}^T(\mathbf{x}_i) \mathbf{r}_i \quad (8.17)$$

is now a Jacobian-weighted sum of residual vectors. This makes intuitive sense, as the parameters are pulled in the direction of the prediction error with a strength proportional to the Jacobian.

Once \mathbf{A} and \mathbf{b} have been computed, we solve for $\Delta\mathbf{p}$ using

$$(\mathbf{A} + \lambda \text{diag}(\mathbf{A}))\Delta\mathbf{p} = \mathbf{b}, \quad (8.18)$$

and update the parameter vector $\mathbf{p} \leftarrow \mathbf{p} + \Delta\mathbf{p}$ accordingly. The parameter λ is an additional damping parameter used to ensure that the system takes a “downhill” step in energy (squared error) and is an essential component of the Levenberg–Marquardt algorithm (described in more detail in Appendix A.3). In many applications, it can be set to 0 if the system is successfully converging.

For the case of our 2D translation+rotation, we end up with a 3×3 set of normal equations in the unknowns $(\delta t_x, \delta t_y, \delta\theta)$. An initial guess for (t_x, t_y, θ) can be obtained by fitting a four-parameter similarity transform in (t_x, t_y, c, s) and then setting $\theta = \tan^{-1}(s/c)$. An alternative approach is to estimate the translation parameters using the centroids of the 2D points and to then estimate the rotation angle using polar coordinates (Exercise 8.3).

For the other 2D motion models, the derivatives in Table 8.1 are all fairly straightforward, except for the projective 2D motion (homography), which arises in image-stitching applications (Section 8.2). These equations can be re-written from (2.21) in their new parametric form as

$$x' = \frac{(1 + h_{00})x + h_{01}y + h_{02}}{h_{20}x + h_{21}y + 1} \quad \text{and} \quad y' = \frac{h_{10}x + (1 + h_{11})y + h_{12}}{h_{20}x + h_{21}y + 1}. \quad (8.19)$$

The Jacobian is therefore

$$\mathbf{J} = \frac{\partial \mathbf{f}}{\partial \mathbf{p}} = \frac{1}{D} \begin{bmatrix} x & y & 1 & 0 & 0 & 0 & -x'x & -x'y \\ 0 & 0 & 0 & x & y & 1 & -y'x & -y'y \end{bmatrix}, \quad (8.20)$$

where $D = h_{20}x + h_{21}y + 1$ is the denominator in (8.19), which depends on the current parameter settings (as do x' and y').

An initial guess for the eight unknowns $\{h_{00}, h_{01}, \dots, h_{21}\}$ can be obtained by multiplying both sides of the equations in (8.19) through by the denominator, which yields the linear

⁵The “Hessian” \mathbf{A} is not the true Hessian (second derivative) of the non-linear least squares problem (8.13). Instead, it is the approximate Hessian, which neglects second (and higher) order derivatives of $\mathbf{f}(\mathbf{x}_i; \mathbf{p} + \Delta\mathbf{p})$.

set of equations,

$$\begin{bmatrix} \hat{x}' - x \\ \hat{y}' - y \end{bmatrix} = \begin{bmatrix} x & y & 1 & 0 & 0 & 0 & -\hat{x}'x & -\hat{x}'y \\ 0 & 0 & 0 & x & y & 1 & -\hat{y}'x & -\hat{y}'y \end{bmatrix} \begin{bmatrix} h_{00} \\ \vdots \\ h_{21} \end{bmatrix}. \quad (8.21)$$

However, this is not optimal from a statistical point of view, since the denominator D , which was used to multiply each equation, can vary quite a bit from point to point.⁶

One way to compensate for this is to *reweight* each equation by the inverse of the current estimate of the denominator, D ,

$$\frac{1}{D} \begin{bmatrix} \hat{x}' - x \\ \hat{y}' - y \end{bmatrix} = \frac{1}{D} \begin{bmatrix} x & y & 1 & 0 & 0 & 0 & -\hat{x}'x & -\hat{x}'y \\ 0 & 0 & 0 & x & y & 1 & -\hat{y}'x & -\hat{y}'y \end{bmatrix} \begin{bmatrix} h_{00} \\ \vdots \\ h_{21} \end{bmatrix}. \quad (8.22)$$

While this may at first seem to be the exact same set of equations as (8.21), because least squares is being used to solve the over-determined set of equations, the weightings *do* matter and produce a different set of normal equations that performs better in practice.

The most principled way to do the estimation, however, is to directly minimize the squared residual Equations (8.13) using the Gauss–Newton approximation, i.e., performing a first-order Taylor series expansion in \mathbf{p} , as shown in (8.14), which yields the set of equations

$$\begin{bmatrix} \hat{x}' - \tilde{x}' \\ \hat{y}' - \tilde{y}' \end{bmatrix} = \frac{1}{D} \begin{bmatrix} x & y & 1 & 0 & 0 & 0 & -\tilde{x}'x & -\tilde{x}'y \\ 0 & 0 & 0 & x & y & 1 & -\tilde{y}'x & -\tilde{y}'y \end{bmatrix} \begin{bmatrix} \Delta h_{00} \\ \vdots \\ \Delta h_{21} \end{bmatrix}. \quad (8.23)$$

While these look similar to (8.22), they differ in two important respects. First, the left-hand side consists of unweighted *prediction errors* rather than point displacements and the solution vector is a *perturbation* to the parameter vector \mathbf{p} . Second, the quantities inside \mathbf{J} involve *predicted* feature locations (\tilde{x}', \tilde{y}') instead of *sensed* feature locations (\hat{x}', \hat{y}') . Both of these differences are subtle and yet they lead to an algorithm that, when combined with proper checking for downhill steps (as in the Levenberg–Marquardt algorithm), will converge to a local minimum. Note that iterating Equations (8.22) is not guaranteed to converge, since it is not minimizing a well-defined energy function.

⁶Hartley and Zisserman (2004) call this strategy of forming linear equations from rational equations the *direct linear transform*, but that term is more commonly associated with pose estimation (Section 11.2). Note also that our definition of the h_{ij} parameters differs from that used in their book, since we define h_{ii} to be the *difference* from unity and we do not leave h_{22} as a free parameter, which means that we cannot handle certain extreme homographies.

Equation (8.23) is analogous to the *additive* algorithm for direct intensity-based registration (Section 9.2), since the change to the full transformation is being computed. If we prepend an incremental homography to the current homography instead, i.e., we use a *compositional* algorithm (described in Section 9.2), we get $D = 1$ (since $\mathbf{p} = 0$) and the above formula simplifies to

$$\begin{bmatrix} \tilde{x}' - x \\ \tilde{y}' - y \end{bmatrix} = \begin{bmatrix} x & y & 1 & 0 & 0 & 0 & -x^2 & -xy \\ 0 & 0 & 0 & x & y & 1 & -xy & -y^2 \end{bmatrix} \begin{bmatrix} \Delta h_{00} \\ \vdots \\ \Delta h_{21} \end{bmatrix}, \quad (8.24)$$

where we have replaced (\tilde{x}', \tilde{y}') with (x, y) for conciseness.

8.1.4 Robust least squares and RANSAC

While regular least squares is the method of choice for measurements where the noise follows a normal (Gaussian) distribution, more robust versions of least squares are required when there are outliers among the correspondences (as there almost always are). In this case, it is preferable to use an *M-estimator* (Huber 1981; Hampel, Ronchetti *et al.* 1986; Black and Rangarajan 1996; Stewart 1999), which involves applying a robust penalty function $\rho(r)$ to the residuals

$$E_{\text{RLS}}(\Delta \mathbf{p}) = \sum_i \rho(\|\mathbf{r}_i\|) \quad (8.25)$$

instead of squaring them.⁷

We can take the derivative of this function with respect to \mathbf{p} and set it to 0,

$$\sum_i \psi(\|\mathbf{r}_i\|) \frac{\partial \|\mathbf{r}_i\|}{\partial \mathbf{p}} = \sum_i \frac{\psi(\|\mathbf{r}_i\|)}{\|\mathbf{r}_i\|} \mathbf{r}_i^T \frac{\partial \mathbf{r}_i}{\partial \mathbf{p}} = 0, \quad (8.26)$$

where $\psi(r) = \rho'(r)$ is the derivative of ρ and is called the *influence function*. If we introduce a *weight function*, $w(r) = \psi(r)/r$, we observe that finding the stationary point of (8.25) using (8.26) is equivalent to minimizing the *iteratively reweighted least squares* (IRLS) problem

$$E_{\text{IRLS}} = \sum_i w(\|\mathbf{r}_i\|) \|\mathbf{r}_i\|^2, \quad (8.27)$$

where the $w(\|\mathbf{r}_i\|)$ play the same local weighting role as σ_i^{-2} in (8.10). The IRLS algorithm alternates between computing the influence functions $w(\|\mathbf{r}_i\|)$ and solving the resulting weighted least squares problem (with fixed w values). Other incremental robust least squares

⁷The plots for some commonly used robust penalty functions ρ can be found in Figure 4.7.

algorithms can be found in the work of Sawhney and Ayer (1996), Black and Anandan (1996), Black and Rangarajan (1996), and Baker, Gross *et al.* (2003) and in textbooks and tutorials on robust statistics (Huber 1981; Hampel, Ronchetti *et al.* 1986; Rousseeuw and Leroy 1987; Stewart 1999).

While M-estimators can definitely help reduce the influence of outliers, in some cases, starting with too many outliers will prevent IRLS (or other gradient descent algorithms) from converging to the global optimum. A better approach is often to find a starting set of *inlier* correspondences, i.e., points that are consistent with a dominant motion estimate.⁸

Two widely used approaches to this problem are called RANdom SAmple Consensus, or RANSAC for short (Fischler and Bolles 1981), and *least median of squares* (LMS) (Rousseeuw 1984). Both techniques start by selecting (at random) a subset of k correspondences, which is then used to compute an initial estimate for \mathbf{p} . The *residuals* of the full set of correspondences are then computed as

$$\mathbf{r}_i = \hat{\mathbf{x}}'_i(\mathbf{x}_i; \mathbf{p}) - \hat{\mathbf{x}}'_i, \quad (8.28)$$

where $\hat{\mathbf{x}}'_i$ are the *estimated* (mapped) locations and $\hat{\mathbf{x}}'_i$ are the sensed (detected) feature point locations.⁹

The RANSAC technique then counts the number of *inliers* that are within ϵ of their predicted location, i.e., whose $\|\mathbf{r}_i\| \leq \epsilon$. (The ϵ value is application dependent but is often around 1–3 pixels.) Least median of squares finds the median value of the $\|\mathbf{r}_i\|^2$ values. The random selection process is repeated S times and the sample set with the largest number of inliers (or with the smallest median residual) is kept as the final solution. Either the initial parameter guess \mathbf{p} or the full set of computed inliers is then passed on to the next data fitting stage.

When the number of measurements is quite large, it may be preferable to only score a subset of the measurements in an initial round that selects the most plausible hypotheses for additional scoring and selection. This modification of RANSAC, which can significantly speed up its performance, is called *Preemptive RANSAC* (Nistér 2003). In another variant on RANSAC called PROSAC (PROgressive SAmple Consensus), random samples are initially added from the most “confident” matches, thereby speeding up the process of finding a (statistically) likely good set of inliers (Chum and Matas 2005). Raguram, Chum *et al.* (2012) provide a unified framework from which most of these techniques can be derived as well as a nice experimental comparison.

Additional variants on RANSAC include MLESAC (Torr and Zisserman 2000), DSAC

⁸For pixel-based alignment methods (Section 9.1.1), hierarchical (coarse-to-fine) techniques are often used to lock onto the *dominant motion* in a scene.

⁹For problems such as epipolar geometry estimation, the residual may be the distance between a point and a line.

| k | p | S |
|---|-----|-----|
| 3 | 0.5 | 35 |
| 6 | 0.6 | 97 |
| 6 | 0.5 | 293 |

Table 8.2 Number of trials S to attain a 99% probability of success (Stewart 1999).

(Brachmann, Krull *et al.* 2017), Graph-Cut RANSAC (Barath and Matas 2018), MAGSAC (Barath, Matas, and Noskova 2019), and ESAC (Brachmann and Rother 2019). Some of these algorithms, such as DSAC (Differentiable RANSAC), are designed to be differentiable so they can be used in end-to-end training of feature detection and matching pipelines (Section 7.1). The MAGSAC++ paper by Barath, Noskova *et al.* (2020) compares many of these variants. Yang, Antonante *et al.* (2020) claim that using a robust penalty function with a decreasing outlier parameter, i.e., *graduated non-convexity* (Blake and Zisserman 1987; Barron 2019), can outperform RANSAC in many geometric correspondence and pose estimation problems. To ensure that the random sampling has a good chance of finding a true set of inliers, a sufficient number of trials S must be evaluated. Let p be the probability that any given correspondence is valid and P be the probability of success after S trials. The likelihood in one trial that all k random samples are inliers is p^k . Therefore, the likelihood that S such trials will all fail is

$$1 - P = (1 - p^k)^S \quad (8.29)$$

and the required minimum number of trials is

$$S = \frac{\log(1 - P)}{\log(1 - p^k)}. \quad (8.30)$$

Stewart (1999) gives examples of the required number of trials S to attain a 99% probability of success. As you can see from Table 8.2, the number of trials grows quickly with the number of sample points used. This provides a strong incentive to use the *minimum* number of sample points k possible for any given trial, which is how RANSAC is normally used in practice.

Uncertainty modeling

In addition to robustly computing a good alignment, some applications require the computation of uncertainty (see Appendix B.6). For linear problems, this estimate can be obtained by inverting the Hessian matrix (8.9) and multiplying it by the feature position noise, if these

have not already been used to weight the individual measurements, as in Equations (8.10) and (8.11). In statistics, the Hessian, which is the inverse covariance, is sometimes called the (Fisher) *information matrix* (Appendix B.1).

When the problem involves non-linear least squares, the inverse of the Hessian matrix provides the *Cramer–Rao lower bound* on the covariance matrix, i.e., it provides the *minimum* amount of covariance in a given solution, which can actually have a wider spread (“longer tails”) if the energy flattens out away from the local minimum where the optimal solution is found.

8.1.5 3D alignment

Instead of aligning 2D sets of image features, many computer vision applications require the alignment of 3D points. In the case where the 3D transformations are linear in the motion parameters, e.g., for translation, similarity, and affine, regular least squares (8.5) can be used.

The case of rigid (Euclidean) motion,

$$E_{\text{R3D}} = \sum_i \| \mathbf{x}'_i - \mathbf{Rx}_i - \mathbf{t} \|^2, \quad (8.31)$$

which arises more frequently and is often called the *absolute orientation* problem (Horn 1987), requires slightly different techniques. If only scalar weightings are being used (as opposed to full 3D per-point anisotropic covariance estimates), the weighted centroids of the two point clouds \mathbf{c} and \mathbf{c}' can be used to estimate the translation $\mathbf{t} = \mathbf{c}' - \mathbf{R}\mathbf{c}$.¹⁰ We are then left with the problem of estimating the rotation between two sets of points $\{\hat{\mathbf{x}}_i = \mathbf{x}_i - \mathbf{c}\}$ and $\{\hat{\mathbf{x}}'_i = \mathbf{x}'_i - \mathbf{c}'\}$ that are both centered at the origin.

One commonly used technique is called the *orthogonal Procrustes algorithm* (Golub and Van Loan 1996, p. 601) and involves computing the singular value decomposition (SVD) of the 3×3 correlation matrix

$$\mathbf{C} = \sum_i \hat{\mathbf{x}}' \hat{\mathbf{x}}^T = \mathbf{U} \boldsymbol{\Sigma} \mathbf{V}^T. \quad (8.32)$$

The rotation matrix is then obtained as $\mathbf{R} = \mathbf{U} \boldsymbol{\Sigma} \mathbf{V}^T$. (Verify this for yourself when $\hat{\mathbf{x}}' = \mathbf{R}\hat{\mathbf{x}}$.)

Another technique is the absolute orientation algorithm (Horn 1987) for estimating the unit quaternion corresponding to the rotation matrix \mathbf{R} , which involves forming a 4×4 matrix from the entries in \mathbf{C} and then finding the eigenvector associated with its largest positive eigenvalue.

¹⁰When full covariances are used, they are transformed by the rotation, so a closed-form solution for translation is not possible.

Lorusso, Eggert, and Fisher (1995) experimentally compare these two techniques to two additional techniques proposed in the literature, but find that the difference in accuracy is negligible (well below the effects of measurement noise).

In situations where these closed-form algorithms are not applicable, e.g., when full 3D covariances are being used or when the 3D alignment is part of some larger optimization, the incremental rotation update introduced in Section 2.1.3 (2.35–2.36), which is parameterized by an instantaneous rotation vector ω , can be used (See Section 8.2.3 for an application to image stitching.)

In some situations, e.g., when merging range data maps, the correspondence between data points is not known *a priori*. In this case, iterative algorithms that start by matching nearby points and then update the most likely correspondence can be used (Besl and McKay 1992; Zhang 1994; Szeliski and Lavallée 1996; Gold, Rangarajan *et al.* 1998; David, DeMenthon *et al.* 2004; Li and Hartley 2007; Enqvist, Josephson, and Kahl 2009). These techniques are discussed in more detail in Section 13.2.1.

8.2 Image stitching

Algorithms for aligning images and stitching them into seamless photo-mosaics are among the oldest and most widely used in computer vision (Milgram 1975; Peleg 1981). Image stitching algorithms create the high-resolution photo-mosaics used to produce today’s digital maps and satellite photos. They are also now a standard mode in smartphone cameras and can be used to create beautiful ultra wide-angle panoramas.

Image stitching originated in the photogrammetry community, where more manually intensive methods based on surveyed *ground control points* or manually registered *tie points* have long been used to register aerial photos into large-scale photo-mosaics (Slama 1980). One of the key advances in this community was the development of *bundle adjustment* algorithms (Section 11.4.2), which could simultaneously solve for the locations of all of the camera positions, thus yielding globally consistent solutions (Triggs, McLauchlan *et al.* 1999). Another recurring problem in creating photo-mosaics is the elimination of visible seams, for which a variety of techniques have been developed over the years (Milgram 1975, 1977; Peleg 1981; Davis 1998; Agarwala, Dontcheva *et al.* 2004).

In film photography, special cameras were developed in the 1990s to take ultra-wide-angle panoramas, often by exposing the film through a vertical slit as the camera rotated on its axis (Meehan 1990). In the mid-1990s, image alignment techniques started being applied to the construction of wide-angle seamless panoramas from regular hand-held cameras (Mann and Picard 1994; Chen 1995; Szeliski 1996). Subsequent algorithms addressed the need to

compute globally consistent alignments (Szeliski and Shum 1997; Sawhney and Kumar 1999; Shum and Szeliski 2000), to remove “ghosts” due to parallax and object movement (Davis 1998; Shum and Szeliski 2000; Uyttendaele, Eden, and Szeliski 2001; Agarwala, Dontcheva *et al.* 2004), and to deal with varying exposures (Mann and Picard 1994; Uyttendaele, Eden, and Szeliski 2001; Levin, Zomet *et al.* 2004; Eden, Uyttendaele, and Szeliski 2006; Kopf, Uyttendaele *et al.* 2007).¹¹

While early techniques worked by directly minimizing pixel-to-pixel dissimilarities, today’s algorithms extract a sparse set of features and match them to each other, as described in Chapter 7. Such feature-based approaches (Zoghlaei, Faugeras, and Deriche 1997; Capel and Zisserman 1998; Cham and Cipolla 1998; Badra, Qumsieh, and Dudek 1998; McLauchlan and Jaenicke 2002; Brown and Lowe 2007) have the advantage of being more robust against scene movement and are usually faster.¹² Their biggest advantage, however, is the ability to “recognize panoramas”, i.e., to automatically discover the adjacency (overlap) relationships among an unordered set of images, which makes them ideally suited for fully automated stitching of panoramas taken by casual users (Brown and Lowe 2007).

What, then, are the essential problems in image stitching? As with image alignment, we must first determine the appropriate mathematical model relating pixel coordinates in one image to pixel coordinates in another; Section 8.2.1 reviews the basic models we have studied and presents some new motion models related specifically to panoramic image stitching. Next, we must somehow estimate the correct alignments relating various pairs (or collections) of images. Chapter 7 discusses how distinctive features can be found in each image and then efficiently matched to rapidly establish correspondences between pairs of images. Chapter 9 discusses how direct pixel-to-pixel comparisons combined with gradient descent (and other optimization techniques) can also be used to estimate these parameters. When multiple images exist in a panorama, global optimization techniques can be used to compute a globally consistent set of alignments and to efficiently discover which images overlap one another. In Section 8.3, we look at how each of these previously developed techniques can be modified to take advantage of the imaging setups commonly used to create panoramas.

Once we have aligned the images, we must choose a final compositing surface for warping the aligned images (Section 8.4.1). We also need algorithms to seamlessly cut and blend overlapping images, even in the presence of parallax, lens distortion, scene motion, and exposure differences (Section 8.4.2–8.4.4).

¹¹A collection of some of these papers was compiled by Benosman and Kang (2001) and they are surveyed by Szeliski (2006a).

¹²See a discussion of the pros and cons of direct vs. feature-based techniques in (Triggs, Zisserman, and Szeliski 2000) and in the first edition of this book (Szeliski 2010, Section 8.3.4).



(a) translation [2 dof] (b) affine [6 dof] (c) perspective [8 dof] (d) 3D rotation [3+ dof]

Figure 8.4 Two-dimensional motion models and how they can be used for image stitching.

8.2.1 Parametric motion models

Before we can register and align images, we need to establish the mathematical relationships that map pixel coordinates from one image to another. A variety of such *parametric motion models* are possible, from simple 2D transforms, to planar perspective models, 3D camera rotations, lens distortions, and mapping to non-planar (e.g., cylindrical) surfaces.

We already covered several of these models in Sections 2.1 and 8.1. In particular, we saw in Section 2.1.4 how the parametric motion describing the deformation of a planar surface as viewed from different positions can be described with an eight-parameter homography (2.71) (Mann and Picard 1994; Szeliski 1996). We also saw how a camera undergoing a pure rotation induces a different kind of homography (2.72).

In this section, we review both of these models and show how they can be applied to different stitching situations. We also introduce spherical and cylindrical compositing surfaces and show how, under favorable circumstances, they can be used to perform alignment using pure translations (Section 8.2.6). Deciding which alignment model is most appropriate for a given situation or set of data is a *model selection* problem (Torr 2002; Bishop 2006; Robert 2007; Hastie, Tibshirani, and Friedman 2009; Murphy 2012), an important topic we do not cover in this book.

Planar perspective motion

The simplest possible motion model to use when aligning images is to simply translate and rotate them in 2D (Figure 8.4a). This is exactly the same kind of motion that you would use if you had overlapping photographic prints. It is also the kind of technique favored by David Hockney to create the collages that he calls *joiners* (Zelnik-Manor and Perona 2007; Nomura, Zhang, and Nayar 2007). Creating such collages, which show visible seams and inconsistencies that add to the artistic effect, is popular on websites such as Flickr, where they

more commonly go under the name *panography* (Section 8.1.2). Translation and rotation are also usually adequate motion models to compensate for small camera motions in applications such as photo and video stabilization and merging (Exercise 8.1 and Section 9.2.1).

In Section 2.1.4, we saw how the mapping between two cameras viewing a common plane can be described using a 3×3 homography (2.71). Consider the matrix \mathbf{M}_{10} that arises when mapping a pixel in one image to a 3D point and then back onto a second image,

$$\tilde{\mathbf{x}}_1 \sim \tilde{\mathbf{P}}_1 \tilde{\mathbf{P}}_0^{-1} \tilde{\mathbf{x}}_0 = \mathbf{M}_{10} \tilde{\mathbf{x}}_0. \quad (8.33)$$

When the last row of the \mathbf{P}_0 matrix is replaced with a plane equation $\hat{\mathbf{n}}_0 \cdot \mathbf{p} + c_0$ and points are assumed to lie on this plane, i.e., their disparity is $d_0 = 0$, we can ignore the last column of \mathbf{M}_{10} and also its last row, since we do not care about the final z-buffer depth. The resulting homography matrix $\tilde{\mathbf{H}}_{10}$ (the upper left 3×3 sub-matrix of \mathbf{M}_{10}) describes the mapping between pixels in the two images,

$$\tilde{\mathbf{x}}_1 \sim \tilde{\mathbf{H}}_{10} \tilde{\mathbf{x}}_0. \quad (8.34)$$

This observation formed the basis of some of the earliest automated image stitching algorithms (Mann and Picard 1994; Szeliski 1994, 1996). Because reliable feature matching techniques had not yet been developed, these algorithms used direct pixel value matching, i.e., direct parametric motion estimation, as described in Section 9.2 and Equations (8.19–8.20).

More recent stitching algorithms first extract features and then match them up, often using robust techniques such as RANSAC (Section 8.1.4) to compute a good set of inliers. The final computation of the homography (8.34), i.e., the solution of the least squares fitting problem given pairs of corresponding features,

$$x_1 = \frac{(1 + h_{00})x_0 + h_{01}y_0 + h_{02}}{h_{20}x_0 + h_{21}y_0 + 1} \quad \text{and} \quad (8.35)$$

$$y_1 = \frac{h_{10}x_0 + (1 + h_{11})y_0 + h_{12}}{h_{20}x_0 + h_{21}y_0 + 1}, \quad (8.36)$$

uses iterative least squares, as described in Section 8.1.3 and Equations (8.21–8.23).

8.2.2 Application: Whiteboard and document scanning

The simplest image-stitching application is to stitch together a number of image scans taken on a flatbed scanner. Say you have a large map, or a piece of child's artwork, that is too large to fit on your scanner. Simply take multiple scans of the document, making sure to overlap the scans by a large enough amount to ensure that there are enough common features. Next, take successive pairs of images that you know overlap, extract features, match them up, and

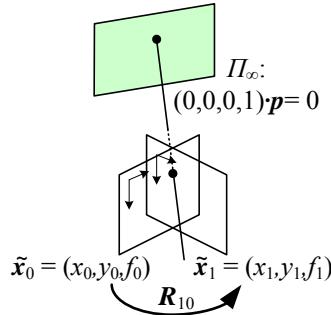


Figure 8.5 Pure 3D camera rotation. The form of the homography (mapping) is particularly simple and depends only on the 3D rotation matrix and focal lengths.

estimate the 2D rigid transform (2.16),

$$\mathbf{x}_{k+1} = \mathbf{R}_k \mathbf{x}_k + \mathbf{t}_k, \quad (8.37)$$

that best matches the features, using two-point RANSAC, if necessary, to find a good set of inliers. Then, on a final compositing surface (aligned with the first scan, for example), resample your images (Section 3.6.1) and average them together. Can you see any potential problems with this scheme?

One complication is that a 2D rigid transformation is non-linear in the rotation angle θ , so you will have to either use non-linear least squares or constrain \mathbf{R} to be orthonormal, as described in Section 8.1.3.

A bigger problem lies in the pairwise alignment process. As you align more and more pairs, the solution may drift so that it is no longer globally consistent. In this case, a global optimization procedure, as described in Section 8.3, may be required. Such global optimization often requires a large system of non-linear equations to be solved, although in some cases, such as linearized homographies (Section 8.2.3) or similarity transforms (Section 8.1.2), regular least squares may be an option.

A slightly more complex scenario is when you take multiple overlapping handheld pictures of a whiteboard or other large planar object (He and Zhang 2005; Zhang and He 2007). Here, the natural motion model to use is a homography, although a more complex model that estimates the 3D rigid motion relative to the plane (plus the focal length, if unknown), could in principle be used.

8.2.3 Rotational panoramas

The most typical case for panoramic image stitching is when the camera undergoes a pure rotation. Think of standing at the rim of the Grand Canyon. Relative to the distant geometry in the scene, as you snap away, the camera is undergoing a pure rotation, which is equivalent to assuming that all points are very far from the camera, i.e., on the *plane at infinity* (Figure 8.5).¹³ Setting $\mathbf{t}_0 = \mathbf{t}_1 = 0$, we get the simplified 3×3 homography

$$\tilde{\mathbf{H}}_{10} = \mathbf{K}_1 \mathbf{R}_1 \mathbf{R}_0^{-1} \mathbf{K}_0^{-1} = \mathbf{K}_1 \mathbf{R}_{10} \mathbf{K}_0^{-1}, \quad (8.38)$$

where $\mathbf{K}_k = \text{diag}(f_k, f_k, 1)$ is the simplified camera intrinsic matrix (2.59), assuming that $c_x = c_y = 0$, i.e., we are indexing the pixels starting from the image center (Szeliski 1996). This can also be re-written as

$$\begin{bmatrix} x_1 \\ y_1 \\ 1 \end{bmatrix} \sim \begin{bmatrix} f_1 & & \\ & f_1 & \\ & & 1 \end{bmatrix} \mathbf{R}_{10} \begin{bmatrix} f_0^{-1} & & \\ & f_0^{-1} & \\ & & 1 \end{bmatrix} \begin{bmatrix} x_0 \\ y_0 \\ 1 \end{bmatrix} \quad (8.39)$$

or

$$\begin{bmatrix} x_1 \\ y_1 \\ f_1 \end{bmatrix} \sim \mathbf{R}_{10} \begin{bmatrix} x_0 \\ y_0 \\ f_0 \end{bmatrix}, \quad (8.40)$$

which reveals the simplicity of the mapping equations and makes all of the motion parameters explicit. Thus, instead of the general eight-parameter homography relating a pair of images, we get the three-, four-, or five-parameter *3D rotation* motion models corresponding to the cases where the focal length f is known, fixed, or variable (Szeliski and Shum 1997).¹⁴ Estimating the 3D rotation matrix (and, optionally, focal length) associated with each image is intrinsically more stable than estimating a homography with a full eight degrees of freedom, which makes this the method of choice for large-scale image stitching algorithms (Szeliski and Shum 1997; Shum and Szeliski 2000; Brown and Lowe 2007).

Given this representation, how do we update the rotation matrices to best align two overlapping images? Given a current estimate for the homography $\tilde{\mathbf{H}}_{10}$ in (8.38), the best way to

¹³In a more general (e.g., indoor) scene, if we want to ensure that there is no *parallax* (visible relative movement between objects at different depths), we need to rotate the camera around the lens's front *no-parallax point* (Littlefield 2006). This can be achieved by using a specialized panoramic rotation head with a built-in translation stage (Houghton 2013) or by determining the front nodal point using observations of collinear points—see Debevec, Wenger *et al.* (2002) and Szeliski (2010, Figure 6.7).

¹⁴An initial estimate of the focal lengths can be obtained using the intrinsic calibration techniques described in Section 11.1.3 or from EXIF tags.

update \mathbf{R}_{10} is to prepend an *incremental* rotation matrix $\mathbf{R}(\boldsymbol{\omega})$ to the current estimate \mathbf{R}_{10} (Szeliski and Shum 1997; Shum and Szeliski 2000),

$$\tilde{\mathbf{H}}(\boldsymbol{\omega}) = \mathbf{K}_1 \mathbf{R}(\boldsymbol{\omega}) \mathbf{R}_{10} \mathbf{K}_0^{-1} = [\mathbf{K}_1 \mathbf{R}(\boldsymbol{\omega}) \mathbf{K}_1^{-1}] [\mathbf{K}_1 \mathbf{R}_{10} \mathbf{K}_0^{-1}] = \mathbf{D} \tilde{\mathbf{H}}_{10}. \quad (8.41)$$

Note that here we have written the update rule in the *compositional* form, where the incremental update \mathbf{D} is *prepended* to the current homography $\tilde{\mathbf{H}}_{10}$. Using the small-angle approximation to $\mathbf{R}(\boldsymbol{\omega})$ given in (2.35), we can write the incremental update matrix as

$$\mathbf{D} = \mathbf{K}_1 \mathbf{R}(\boldsymbol{\omega}) \mathbf{K}_1^{-1} \approx \mathbf{K}_1 (\mathbf{I} + [\boldsymbol{\omega}]_\times) \mathbf{K}_1^{-1} = \begin{bmatrix} 1 & -\omega_z & f_1 \omega_y \\ \omega_z & 1 & -f_1 \omega_x \\ -\omega_y/f_1 & \omega_x/f_1 & 1 \end{bmatrix}. \quad (8.42)$$

Notice how there is now a nice one-to-one correspondence between the entries in the \mathbf{D} matrix and the h_{00}, \dots, h_{21} parameters used in Table 8.1 and Equation (8.19), i.e.,

$$(h_{00}, h_{01}, h_{02}, h_{00}, h_{11}, h_{12}, h_{20}, h_{21}) = (0, -\omega_z, f_1 \omega_y, \omega_z, 0, -f_1 \omega_x, -\omega_y/f_1, \omega_x/f_1). \quad (8.43)$$

We can therefore apply the chain rule to Equations (8.24 and 8.43) to obtain

$$\begin{bmatrix} \hat{x}' - x \\ \hat{y}' - y \end{bmatrix} = \begin{bmatrix} -xy/f_1 & f_1 + x^2/f_1 & -y \\ -(f_1 + y^2/f_1) & xy/f_1 & x \end{bmatrix} \begin{bmatrix} \omega_x \\ \omega_y \\ \omega_z \end{bmatrix}, \quad (8.44)$$

which give us the linearized update equations needed to estimate $\boldsymbol{\omega} = (\omega_x, \omega_y, \omega_z)$.¹⁵ Notice that this update rule depends on the focal length f_1 of the *target* view and is independent of the focal length f_0 of the *template* view. This is because the compositional algorithm essentially makes small perturbations to the target. Once the incremental rotation vector $\boldsymbol{\omega}$ has been computed, the \mathbf{R}_1 rotation matrix can be updated using $\mathbf{R}_1 \leftarrow \mathbf{R}(\boldsymbol{\omega}) \mathbf{R}_1$.

The formulas for updating the focal length estimates are a little more involved and are given in Shum and Szeliski (2000). We will not repeat them here, since an alternative update rule, based on minimizing the difference between back-projected 3D rays, is given in Section 8.3.1. Figure 8.1a shows the alignment of four images under the 3D rotation motion model.

8.2.4 Gap closing

The techniques presented in this section can be used to estimate a series of rotation matrices and focal lengths, which can be chained together to create large panoramas. Unfortunately,

¹⁵This is the same as the rotational component of instantaneous rigid flow (Bergen, Anandan *et al.* 1992) and the update equations given by Szeliski and Shum (1997) and Shum and Szeliski (2000).

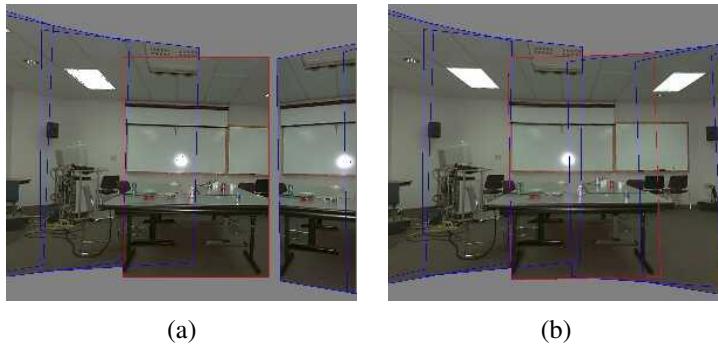


Figure 8.6 *Gap closing (Szeliski and Shum 1997) © 1997 ACM:* (a) A gap is visible when the focal length is wrong ($f = 510$). (b) No gap is visible for the correct focal length ($f = 468$).

because of accumulated errors, this approach will rarely produce a closed 360° panorama. Instead, there will invariably be either a gap or an overlap (Figure 8.6).

We can solve this problem by matching the first image in the sequence with the last one. The difference between the two rotation matrix estimates associated with the repeated first image indicates the amount of misregistration. This error can be distributed evenly across the whole sequence by taking the quotient of the two quaternions associated with these rotations and dividing this “error quaternion” by the number of images in the sequence (Szeliski and Shum 1997). We can also update the estimated focal length based on the amount of misregistration. To do this, we first convert the error quaternion into a *gap angle*, θ_g and then update the focal length using the equation $f' = f(1 - \theta_g/360^\circ)$.

Figure 8.6a shows the end of registered image sequence and the first image. There is a big gap between the last image and the first, which are in fact the same image. The gap is 32° because the wrong estimate of focal length ($f = 510$) was used. Figure 8.6b shows the registration after closing the gap with the correct focal length ($f = 468$). Notice that both mosaics show very little visual misregistration (except at the gap), yet Figure 8.6a has been computed using a focal length that has 9% error. Related approaches have been developed by Hartley (1994b), McMillan and Bishop (1995), Stein (1995), and Kang and Weiss (1997) to solve the focal length estimation problem using pure panning motion and cylindrical images.

Unfortunately, this gap-closing heuristic only works for the kind of “one-dimensional” panorama where the camera is continuously turning in the same direction. In Section 8.3, we describe a different approach to removing gaps and overlaps that works for arbitrary camera motions.

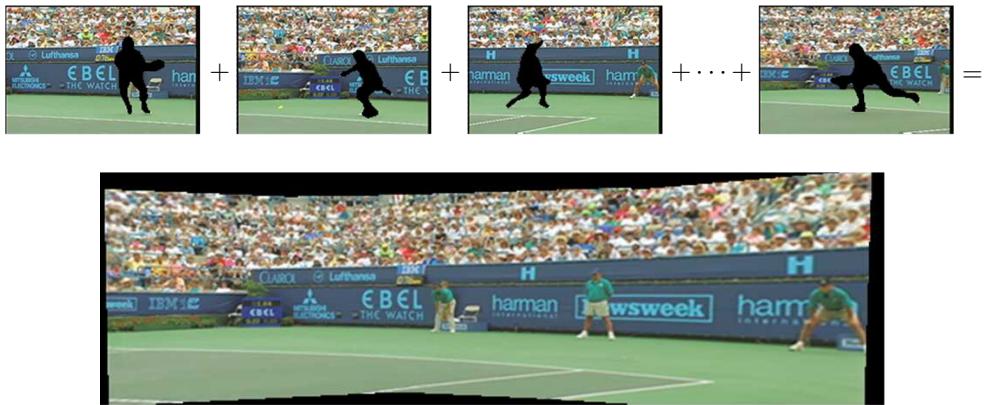


Figure 8.7 Video stitching the background scene to create a single sprite image that can be transmitted and used to re-create the background in each frame (Lee, Chen et al. 1997) © 1997 IEEE.

8.2.5 Application: Video summarization and compression

An interesting application of image stitching is the ability to summarize and compress videos taken with a panning camera. This application was first suggested by Teodosio and Bender (1993), who called their mosaic-based summaries *salient stills*. These ideas were then extended by Irani, Hsu, and Anandan (1995) and Irani and Anandan (1998) to additional applications, such as video compression and video indexing. While these early approaches used affine motion models and were therefore restricted to long focal lengths, the techniques were generalized by Lee, Chen et al. (1997) to full eight-parameter homographies and incorporated into the MPEG-4 video compression standard, where the stitched background layers were called *video sprites* (Figure 8.7).

While video stitching is in many ways a straightforward generalization of multiple-image stitching (Steedly, Pal, and Szeliski 2005; Baudisch, Tan et al. 2006), the potential presence of large amounts of independent motion, camera zoom, and the desire to visualize dynamic events impose additional challenges. For example, moving foreground objects can often be removed using *median filtering*. Alternatively, foreground objects can be extracted into a separate layer (Sawhney and Ayer 1996) and later composited back into the stitched panoramas, sometimes as multiple instances to give the impressions of a “Chronophotograph” (Massey and Bender 1996) and sometimes as video overlays (Irani and Anandan 1998). Videos can also be used to create animated *panoramic video textures* (Section 14.5.2), in which different portions of a panoramic scene are animated with independently moving video loops (Agar-

wala, Zheng *et al.* 2005; Rav-Acha, Pritch *et al.* 2005; Joshi, Mehta *et al.* 2012; Yan, Liu, and Furukawa 2017; He, Liao *et al.* 2017; Oh, Joo *et al.* 2017), or to shine “video flashlights” onto a composite mosaic of a scene (Sawhney, Arpa *et al.* 2002).

Video can also provide an interesting source of content for creating panoramas taken from moving cameras. While this invalidates the usual assumption of a single point of view (optical center), interesting results can still be obtained. For example, the VideoBrush system of Sawhney, Kumar *et al.* (1998) uses thin strips taken from the center of the image to create a panorama taken from a horizontally moving camera. This idea can be generalized to other camera motions and compositing surfaces using the concept of mosaics on an adaptive manifold (Peleg, Rousso *et al.* 2000), and also used to generate panoramic stereograms (Ishiguro, Yamamoto, and Tsuji 1992; Peleg, Ben-Ezra, and Pritch 2001).¹⁶ Related ideas have been used to create panoramic matte paintings for multiplane cel animation (Wood, Finkelstein *et al.* 1997), for creating stitched images of scenes with parallax (Kumar, Anandan *et al.* 1995), and as 3D representations of more complex scenes using *multiple-center-of-projection images* (Rademacher and Bishop 1998) and *multi-perspective panoramas* (Román, Garg, and Levoy 2004; Román and Lensch 2006; Agarwala, Agrawala *et al.* 2006; Kopf, Chen *et al.* 2010).

Another interesting variant on video-based panoramas is *concentric mosaics* (Section 14.3.3) (Shum and He 1999). Here, rather than trying to produce a single panoramic image, the complete original video is kept and used to re-synthesize views (from different camera origins) using ray remapping (light field rendering), thus endowing the panorama with a sense of 3D depth. The same dataset can also be used to explicitly reconstruct the depth using multi-baseline stereo (Ishiguro, Yamamoto, and Tsuji 1992; Peleg, Ben-Ezra, and Pritch 2001; Li, Shum *et al.* 2004; Zheng, Kang *et al.* 2007).

8.2.6 Cylindrical and spherical coordinates

An alternative to using homographies or 3D motions to align images is to first warp the images into *cylindrical* coordinates and then use a pure translational model to align them (Chen 1995; Szeliski 1996). Unfortunately, this only works if the images are all taken with a level camera or with a known tilt angle.

Assume for now that the camera is in its canonical position, i.e., its rotation matrix is the identity, $\mathbf{R} = \mathbf{I}$, so that the optical axis is aligned with the z -axis and the y -axis is aligned vertically. The 3D ray corresponding to an (x, y) pixel is therefore (x, y, f) .

¹⁶A similar technique was likely used in the Google Cardboard Camera, <https://blog.google/products/google-vr/cardboard-camera-ios>.

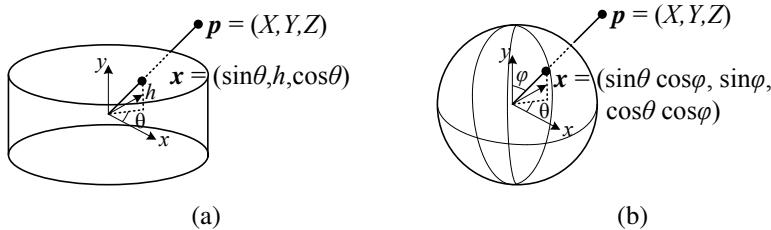


Figure 8.8 Projection from 3D to (a) cylindrical and (b) spherical coordinates.

We wish to project this image onto a *cylindrical surface* of unit radius (Szeliski 1996). Points on this surface are parameterized by an angle θ and a height h , with the 3D cylindrical coordinates corresponding to (θ, h) given by

$$(\sin \theta, h, \cos \theta) \propto (x, y, f), \quad (8.45)$$

as shown in Figure 8.8a. From this correspondence, we can compute the formula for the warped or mapped coordinates (Szeliski and Shum 1997),

$$x' = s\theta = s \tan^{-1} \frac{x}{f}, \quad (8.46)$$

$$y' = sh = s \frac{y}{\sqrt{x^2 + f^2}}, \quad (8.47)$$

where s is an arbitrary scaling factor (sometimes called the *radius* of the cylinder) that can be set to $s = f$ to minimize the distortion (scaling) near the center of the image.¹⁷ The inverse of this mapping equation is given by

$$x = f \tan \theta = f \tan \frac{x'}{s}, \quad (8.48)$$

$$y = h\sqrt{x^2 + f^2} = \frac{y'}{s}f\sqrt{1 + \tan^2 x'/s} = f\frac{y'}{s}\sec\frac{x'}{s}. \quad (8.49)$$

Images can also be projected onto a *spherical surface* (Szeliski and Shum 1997), which is useful if the final panorama includes a full sphere or hemisphere of views, instead of just a cylindrical strip. In this case, the sphere is parameterized by two angles (θ, ϕ) , with 3D spherical coordinates given by

$$(\sin \theta \cos \phi, \sin \phi, \cos \theta \cos \phi) \propto (x, y, f), \quad (8.50)$$

¹⁷The scale can also be set to a larger or smaller value for the final compositing surface, depending on the desired output panorama resolution—see Section 8.4.

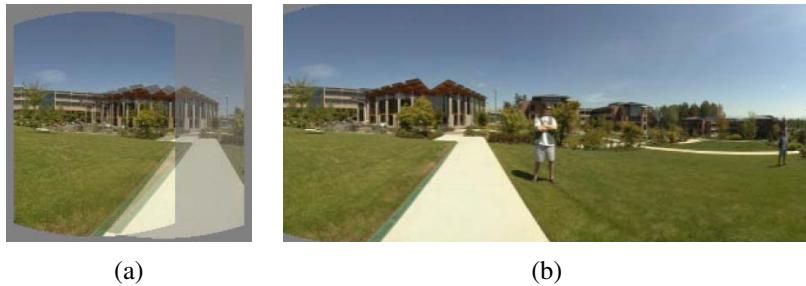


Figure 8.9 A cylindrical panorama (Szeliski and Shum 1997) © 1997 ACM: (a) two cylindrically warped images related by a horizontal translation; (b) part of a cylindrical panorama composited from a sequence of images.

as shown in Figure 8.8b.¹⁸ The correspondence between coordinates is now given by (Szeliski and Shum 1997):

$$x' = s\theta = s \tan^{-1} \frac{x}{f}, \quad (8.51)$$

$$y' = s\phi = s \tan^{-1} \frac{y}{\sqrt{x^2 + f^2}}, \quad (8.52)$$

while the inverse is given by

$$x = f \tan \theta = f \tan \frac{x'}{s}, \quad (8.53)$$

$$y = \sqrt{x^2 + f^2} \tan \phi = \tan \frac{y'}{s} f \sqrt{1 + \tan^2 x'/s} = f \tan \frac{y'}{s} \sec \frac{x'}{s}. \quad (8.54)$$

Note that it may be simpler to generate a scaled (x, y, z) direction from Equation (8.50) followed by a perspective division by z and a scaling by f .

Cylindrical image stitching algorithms are most commonly used when the camera is known to be level and only rotating around its vertical axis (Chen 1995). Under these conditions, images at different rotations are related by a pure horizontal translation.¹⁹ This makes it attractive as an initial class project in an introductory computer vision course, since the full complexity of the perspective alignment algorithm (Sections 8.1, 9.2, and 8.2.3) can be avoided. Figure 8.9 shows how two cylindrically warped images from a leveled rotational panorama are related by a pure translation (Szeliski and Shum 1997).

¹⁸Note that these are not the usual spherical coordinates, first presented in Equation (2.8). Here, the y -axis points at the north pole instead of the z -axis, since we are used to viewing images taken horizontally, i.e., with the y -axis pointing in the direction of the gravity vector.

¹⁹Small vertical tilts can sometimes be compensated for with vertical translations.

Professional panoramic photographers often use pan-tilt heads that make it easy to control the tilt and to stop at specific *detents* in the rotation angle. Motorized rotation heads are also sometimes used for the acquisition of larger panoramas (Kopf, Uyttendaele *et al.* 2007).²⁰ Not only do they ensure a uniform coverage of the visual field with a desired amount of image overlap but they also make it possible to stitch the images using cylindrical or spherical coordinates and pure translations. In this case, pixel coordinates (x, y, f) must first be rotated using the known tilt and panning angles before being projected into cylindrical or spherical coordinates (Chen 1995). Having a roughly known panning angle also makes it easier to compute the alignment, as the rough relative positioning of all the input images is known ahead of time, enabling a reduced search range for alignment. Figure 8.1b shows a full 3D rotational panorama unwrapped onto the surface of a sphere (Szeliski and Shum 1997).

One final coordinate mapping worth mentioning is the *polar* mapping, where the north pole lies along the optical axis rather than the vertical axis,

$$(\cos \theta \sin \phi, \sin \theta \sin \phi, \cos \phi) = s(x, y, z). \quad (8.55)$$

In this case, the mapping equations become

$$x' = s\phi \cos \theta = s \frac{x}{r} \tan^{-1} \frac{r}{z}, \quad (8.56)$$

$$y' = s\phi \sin \theta = s \frac{y}{r} \tan^{-1} \frac{r}{z}, \quad (8.57)$$

where $r = \sqrt{x^2 + y^2}$ is the *radial distance* in the (x, y) plane and $s\phi$ plays a similar role in the (x', y') plane. This mapping provides an attractive visualization surface for certain kinds of wide-angle panoramas and is also a good model for the distortion induced by *fisheye lenses*, as discussed in Section 2.1.5. Note how for small values of (x, y) , the mapping equations reduce to $x' \approx sx/z$, which suggests that s plays a role similar to the focal length f .

8.3 Global alignment

So far, we have discussed how to register pairs of images using a variety of motion models. In most applications, we are given more than a single pair of images to register. The goal is then to find a globally consistent set of alignment parameters that minimize the misregistration between all pairs of images (Szeliski and Shum 1997; Shum and Szeliski 2000; Sawhney and Kumar 1999; Coorg and Teller 2000).

²⁰See also <https://gigapan.org>.

In this section, we extend the pairwise matching criteria (8.2, 9.1, and 9.43) to a global energy function that involves all of the per-image pose parameters (Section 8.3.1). Once we have computed the global alignment, we often need to perform *local adjustments*, such as *parallax removal*, to reduce double images and blurring due to local misregistrations (Section 8.3.2). Finally, if we are given an unordered set of images to register, we need to discover which images go together to form one or more panoramas. This process of *panorama recognition* is described in Section 8.3.3.

8.3.1 Bundle adjustment

One way to register a large number of images is to add new images to the panorama one at a time, aligning the most recent image with the previous ones already in the collection (Szeliski and Shum 1997) and discovering, if necessary, which images it overlaps (Sawhney and Kumar 1999). In the case of 360° panoramas, accumulated error may lead to the presence of a gap (or excessive overlap) between the two ends of the panorama, which can be fixed by stretching the alignment of all the images using a process called *gap closing* (Section 8.2.4). However, a better alternative is to simultaneously align all the images using a least-squares framework to correctly distribute any misregistration errors.

The process of simultaneously adjusting pose parameters and 3D point locations for a large collection of overlapping images is called *bundle adjustment* in the photogrammetry community (Triggs, McLauchlan *et al.* 1999). In computer vision, it was first applied to the general structure from motion problem (Szeliski and Kang 1994) and then later specialized for panoramic image stitching (Shum and Szeliski 2000; Sawhney and Kumar 1999; Coorg and Teller 2000).

In this section, we formulate the problem of global alignment using a feature-based approach, since this results in a simpler system. An equivalent direct approach can be obtained either by dividing images into patches and creating a virtual feature correspondence for each one (Shum and Szeliski 2000) or by replacing the per-feature error metrics with per-pixel metrics (Irani and Anandan 1999).

Before we describe this in more details, we should mention that a simpler, although less accurate, approach is to compute pairwise rotation estimates between overlapping images, and to then use a *rotation averaging* approach to estimate a global rotation for each camera (Hartley, Trumpf *et al.* 2013). However, since the measurement errors in each feature point location are not being counted correctly, as is the case in bundle adjustment, the solution will not have the same theoretical optimality.

Consider the feature-based alignment problem given in Equation (8.2), i.e.,

$$E_{\text{pairwise-LS}} = \sum_i \|\mathbf{r}_i\|^2 = \|\tilde{\mathbf{x}}'_i(\mathbf{x}_i; \mathbf{p}) - \hat{\mathbf{x}}'_i\|^2. \quad (8.58)$$

For multi-image alignment, instead of having a single collection of pairwise feature correspondences, $\{(\mathbf{x}_i, \hat{\mathbf{x}}'_i)\}$, we have a collection of n features, with the location of the i th feature point in the j th image denoted by \mathbf{x}_{ij} and its scalar confidence (i.e., inverse variance) denoted by c_{ij} .²¹ Each image also has some associated pose parameters.

In this section, we assume that this pose consists of a rotation matrix \mathbf{R}_j and a focal length f_j , although formulations in terms of homographies are also possible (Szeliski and Shum 1997; Sawhney and Kumar 1999). The equation mapping a 3D point \mathbf{x}_i into a point \mathbf{x}_{ij} in frame j can be re-written from Equations (2.68) and (8.38) as

$$\tilde{\mathbf{x}}_{ij} \sim \mathbf{K}_j \mathbf{R}_j \mathbf{x}_i \quad \text{and} \quad \mathbf{x}_i \sim \mathbf{R}_j^{-1} \mathbf{K}_j^{-1} \tilde{\mathbf{x}}_{ij}, \quad (8.59)$$

where $\mathbf{K}_j = \text{diag}(f_j, f_j, 1)$ is the simplified form of the calibration matrix. The motion mapping a point \mathbf{x}_{ij} from frame j into a point \mathbf{x}_{ik} in frame k is similarly given by

$$\tilde{\mathbf{x}}_{ik} \sim \tilde{\mathbf{H}}_{kj} \tilde{\mathbf{x}}_{ij} = \mathbf{K}_k \mathbf{R}_k \mathbf{R}_j^{-1} \mathbf{K}_j^{-1} \tilde{\mathbf{x}}_{ij}. \quad (8.60)$$

Given an initial set of $\{(\mathbf{R}_j, f_j)\}$ estimates obtained from chaining pairwise alignments, how do we refine these estimates?

One approach is to directly extend the pairwise energy $E_{\text{pairwise-LS}}$ (8.58) to a multiview formulation,

$$E_{\text{all-pairs-2D}} = \sum_i \sum_{jk} c_{ij} c_{ik} \|\tilde{\mathbf{x}}_{ik}(\hat{\mathbf{x}}_{ij}; \mathbf{R}_j, f_j, \mathbf{R}_k, f_k) - \hat{\mathbf{x}}_{ik}\|^2, \quad (8.61)$$

where the $\tilde{\mathbf{x}}_{ik}$ function is the *predicted* location of feature i in frame k given by (8.60), $\hat{\mathbf{x}}_{ij}$ is the *observed* location, and the “2D” in the subscript indicates that an image-plane error is being minimized (Shum and Szeliski 2000). Note that since $\tilde{\mathbf{x}}_{ik}$ depends on the $\hat{\mathbf{x}}_{ij}$ observed value, we actually have an *errors-in-variable* problem, which in principle requires more sophisticated techniques than least squares to solve (Van Huffel and Lemmerling 2002; Matei and Meer 2006). However, in practice, if we have enough features, we can directly minimize the above quantity using regular non-linear least squares and obtain an accurate multi-frame alignment.

While this approach works pretty well, it suffers from two potential disadvantages. First, because a summation is taken over all pairs with corresponding features, features that are

²¹Features that are not seen in image j have $c_{ij} = 0$. We can also use 2×2 inverse covariance matrices Σ_{ij}^{-1} in place of c_{ij} , as shown in Equation (8.11).

observed many times are overweighted in the final solution. (In effect, a feature observed m times gets counted $\binom{m}{2}$ times instead of m times.) Second, the derivatives of $\tilde{\mathbf{x}}_{ik}$ with respect to the $\{(\mathbf{R}_j, f_j)\}$ are a little cumbersome, although using the incremental correction to \mathbf{R}_j introduced in Section 8.2.3 makes this more tractable.

An alternative way to formulate the optimization is to use true bundle adjustment, i.e., to solve not only for the pose parameters $\{(\mathbf{R}_j, f_j)\}$ but also for the 3D point positions $\{\mathbf{x}_i\}$,

$$E_{\text{BA-2D}} = \sum_i \sum_j c_{ij} \|\tilde{\mathbf{x}}_{ij}(\mathbf{x}_i; \mathbf{R}_j, f_j) - \hat{\mathbf{x}}_{ij}\|^2, \quad (8.62)$$

where $\tilde{\mathbf{x}}_{ij}(\mathbf{x}_i; \mathbf{R}_j, f_j)$ is given by (8.59). The disadvantage of full bundle adjustment is that there are more variables to solve for, so each iteration and also the overall convergence may be slower. (Imagine how the 3D points need to “shift” each time some rotation matrices are updated.) However, the computational complexity of each linearized Gauss–Newton step can be reduced using sparse matrix techniques (Section 11.4.3) (Szeliski and Kang 1994; Triggs, McLauchlan *et al.* 1999; Hartley and Zisserman 2004).

An alternative formulation is to minimize the error in 3D projected ray directions (Shum and Szeliski 2000), i.e.,

$$E_{\text{BA-3D}} = \sum_i \sum_j c_{ij} \|\tilde{\mathbf{x}}_i(\hat{\mathbf{x}}_{ij}; \mathbf{R}_j, f_j) - \mathbf{x}_i\|^2, \quad (8.63)$$

where $\tilde{\mathbf{x}}_i(\mathbf{x}_{ij}; \mathbf{R}_j, f_j)$ is given by the second half of (8.59). This has no particular advantage over (8.62). In fact, since errors are being minimized in 3D ray space, there is a bias towards estimating longer focal lengths, since the angles between rays become smaller as f increases.

However, if we eliminate the 3D rays \mathbf{x}_i , we can derive a pairwise energy formulated in 3D ray space (Shum and Szeliski 2000),

$$E_{\text{all-pairs-3D}} = \sum_i \sum_{jk} c_{ij} c_{ik} \|\tilde{\mathbf{x}}_i(\hat{\mathbf{x}}_{ij}; \mathbf{R}_j, f_j) - \tilde{\mathbf{x}}_i(\hat{\mathbf{x}}_{ik}; \mathbf{R}_k, f_k)\|^2. \quad (8.64)$$

This results in the simplest set of update equations (Shum and Szeliski 2000), since the f_k can be folded into the creation of the homogeneous coordinate vector as in Equation (8.40). Thus, even though this formula over-weights features that occur more frequently, it is the method used by Shum and Szeliski (2000) and Brown, Szeliski, and Winder (2005). To reduce the bias towards longer focal lengths, we multiply each residual (3D error) by $\sqrt{f_j f_k}$, which is similar to projecting the 3D rays into a “virtual camera” of intermediate focal length.

Up vector selection. As mentioned above, there exists a global ambiguity in the pose of the 3D cameras computed by the above methods. While this may not appear to matter, people

prefer that the final stitched image is “upright” rather than twisted or tilted. More concretely, people are used to seeing photographs displayed so that the vertical (gravity) axis points straight up in the image. Consider how you usually shoot photographs: while you may pan and tilt the camera any which way, you usually keep the horizontal edge of your camera (its x -axis) parallel to the ground plane (perpendicular to the world gravity direction).

Mathematically, this constraint on the rotation matrices can be expressed as follows. Recall from Equation (8.59) that the 3D to 2D projection is given by

$$\tilde{\mathbf{x}}_{ik} \sim \mathbf{K}_k \mathbf{R}_k \mathbf{x}_i. \quad (8.65)$$

We wish to post-multiply each rotation matrix \mathbf{R}_k by a global rotation \mathbf{R}_G such that the projection of the global y -axis, $\hat{\mathbf{j}} = (0, 1, 0)$ is perpendicular to the image x -axis, $\hat{\mathbf{i}} = (1, 0, 0)$.²²

This constraint can be written as

$$\hat{\mathbf{i}}^T \mathbf{R}_k \mathbf{R}_G \hat{\mathbf{j}} = 0 \quad (8.66)$$

(note that the scaling by the calibration matrix is irrelevant here). This is equivalent to requiring that the first row of \mathbf{R}_k , $\mathbf{r}_{k0} = \hat{\mathbf{i}}^T \mathbf{R}_k$ be perpendicular to the second column of \mathbf{R}_G , $\mathbf{r}_{G1} = \mathbf{R}_G \hat{\mathbf{j}}$. This set of constraints (one per input image) can be written as a least squares problem,

$$\mathbf{r}_{G1} = \arg \min_{\mathbf{r}} \sum_k (\mathbf{r}^T \mathbf{r}_{k0})^2 = \arg \min_{\mathbf{r}} \mathbf{r}^T \left[\sum_k \mathbf{r}_{k0} \mathbf{r}_{k0}^T \right] \mathbf{r}. \quad (8.67)$$

Thus, \mathbf{r}_{G1} is the smallest eigenvector of the *scatter* or *moment* matrix spanned by the individual camera rotation x -vectors, which should generally be of the form $(c, 0, s)$ when the cameras are upright.

To fully specify the \mathbf{R}_G global rotation, we need to specify one additional constraint. This is related to the *view selection* problem discussed in Section 8.4.1. One simple heuristic is to prefer the average z -axis of the individual rotation matrices, $\bar{\mathbf{k}} = \sum_k \hat{\mathbf{k}}^T \mathbf{R}_k$ to be close to the world z -axis, $\mathbf{r}_{G2} = \mathbf{R}_G \hat{\mathbf{k}}$. We can therefore compute the full rotation matrix \mathbf{R}_G in three steps:

1. $\mathbf{r}_{G1} = \min$ eigenvector $(\sum_k \mathbf{r}_{k0} \mathbf{r}_{k0}^T)$;
2. $\mathbf{r}_{G0} = \mathcal{N}((\sum_k \mathbf{r}_{k2}) \times \mathbf{r}_{G1})$;
3. $\mathbf{r}_{G2} = \mathbf{r}_{G0} \times \mathbf{r}_{G1}$,

where $\mathcal{N}(\mathbf{v}) = \mathbf{v} / \|\mathbf{v}\|$ normalizes a vector \mathbf{v} .

²²Note that here we use the convention common in computer graphics that the vertical world axis corresponds to y . This is a natural choice if we wish the rotation matrix associated with a “regular” image taken horizontally to be the identity, rather than a 90° rotation around the x -axis.

8.3.2 Parallax removal

Once we have optimized the global orientations and focal lengths of our cameras, we may find that the images are still not perfectly aligned, i.e., the resulting stitched image looks blurry or ghosted in some places. This can be caused by a variety of factors, including unmodeled radial distortion, 3D parallax (failure to rotate the camera around its front nodal point), small scene motions such as waving tree branches, and large-scale scene motions such as people moving in and out of pictures.

Each of these problems can be treated with a different approach. Radial distortion can be estimated (potentially ahead of time) using one of the techniques discussed in Section 2.1.5. For example, the *plumb-line method* (Brown 1971; Kang 2001; El-Melegy and Farag 2003) adjusts radial distortion parameters until slightly curved lines become straight, while mosaic-based approaches adjust them until misregistration is reduced in image overlap areas (Stein 1997; Sawhney and Kumar 1999).

3D parallax can be handled by doing a full 3D bundle adjustment, i.e., by replacing the projection Equation (8.59) used in Equation (8.62) with Equation (2.68), which models camera translations. The 3D positions of the matched feature points and cameras can then be simultaneously recovered, although this can be significantly more expensive than parallax-free image registration. Once the 3D structure has been recovered, the scene could (in theory) be projected to a single (central) viewpoint that contains no parallax. However, to do this, dense *stereo* correspondence needs to be performed (Section 12.3) (Li, Shum *et al.* 2004; Zheng, Kang *et al.* 2007), which may not be possible if the images contain only partial overlap. In that case, it may be necessary to correct for parallax only in the overlap areas, which can be accomplished using a *multi-perspective plane sweep* (MPPS) algorithm (Kang, Szeliski, and Uyttendaele 2004; Uyttendaele, Criminisi *et al.* 2004).

When the motion in the scene is very large, i.e., when objects appear and disappear completely, a sensible solution is to simply *select* pixels from only one image at a time as the source for the final composite (Milgram 1977; Davis 1998; Agarwala, Dontcheva *et al.* 2004), as discussed in Section 8.4.2. However, when the motion is reasonably small (on the order of a few pixels), general 2D motion estimation (optical flow) can be used to perform an appropriate correction before blending using a process called *local alignment* (Shum and Szeliski 2000; Kang, Uyttendaele *et al.* 2003). This same process can also be used to compensate for radial distortion and 3D parallax, although it uses a weaker motion model than explicitly modeling the source of error and may, therefore, fail more often or introduce unwanted distortions.

The local alignment technique introduced by Shum and Szeliski (2000) starts with the global bundle adjustment (8.64) used to optimize the camera poses. Once these have been

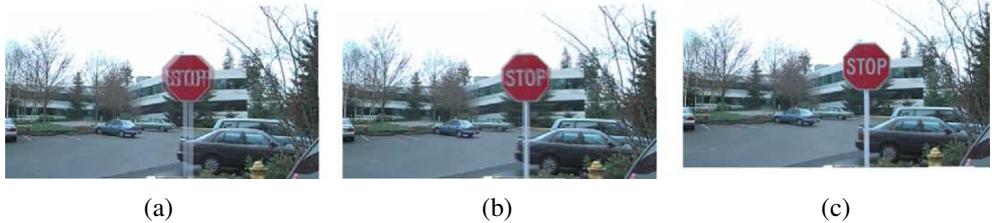


Figure 8.10 *Deghosting a mosaic with motion parallax (Shum and Szeliski 2000) © 2000 IEEE:* (a) composite with parallax; (b) after a single deghosting step (patch size 32); (c) after multiple steps (sizes 32, 16 and 8).

estimated, the *desired* location of a 3D point \mathbf{x}_i can be estimated as the *average* of the back-projected 3D locations,

$$\bar{\mathbf{x}}_i \sim \sum_j c_{ij} \tilde{\mathbf{x}}_i(\hat{\mathbf{x}}_{ij}; \mathbf{R}_j, f_j) \left/ \sum_j c_{ij} \right., \quad (8.68)$$

which can be projected into each image j to obtain a *target location* $\bar{\mathbf{x}}_{ij}$. The difference between the target locations $\bar{\mathbf{x}}_{ij}$ and the original features \mathbf{x}_{ij} provide a set of local motion estimates

$$\mathbf{u}_{ij} = \bar{\mathbf{x}}_{ij} - \mathbf{x}_{ij}, \quad (8.69)$$

which can be interpolated to form a dense correction field $\mathbf{u}_j(\mathbf{x}_j)$. In their system, Shum and Szeliski (2000) use an *inverse warping* algorithm where the sparse $-\mathbf{u}_{ij}$ values are placed at the new target locations $\bar{\mathbf{x}}_{ij}$, interpolated using bilinear kernel functions (Nielson 1993) and then added to the original pixel coordinates when computing the warped (corrected) image. To get a reasonably dense set of features to interpolate, Shum and Szeliski (2000) place a feature point at the center of each patch (the patch size controls the smoothness in the local alignment stage), rather than relying on features extracted using an interest operator (Figure 8.10).

An alternative approach to motion-based deghosting was proposed by Kang, Uyttendaele *et al.* (2003), who estimate dense optical flow between each input image and a central *reference* image. The accuracy of the flow vector is checked using a photo-consistency measure before a given warped pixel is considered valid and is used to compute a high dynamic range radiance estimate, which is the goal of their overall algorithm. The requirement for a reference image makes their approach less applicable to general image mosaicing, although an extension to this case could certainly be envisaged.

The idea of combining *global* parametric warps with *local* mesh-based warps or multiple motion models to compensate for parallax has been refined in a number of more recent papers

(Zaragoza, Chin *et al.* 2013; Zhang and Liu 2014; Lin, Pankanti *et al.* 2015; Lin, Jiang *et al.* 2016; Herrmann, Wang *et al.* 2018b; Lee and Sim 2020). Some of these papers use *content-preserving warps* (Liu, Gleicher *et al.* 2009) for their local deformations, while others include a rolling shutter model (Zhuang and Tran 2020).

8.3.3 Recognizing panoramas

The final piece needed to perform fully automated image stitching is a technique to recognize which images actually go together, which Brown and Lowe (2007) call *recognizing panoramas*. If the user takes images in sequence so that each image overlaps its predecessor and also specifies the first and last images to be stitched, bundle adjustment combined with the process of *topology inference* can be used to automatically assemble a panorama (Sawhney and Kumar 1999). However, users often jump around when taking panoramas, e.g., they may start a new row on top of a previous one, jump back to take a repeat shot, or create 360° panoramas where end-to-end overlaps need to be discovered. Furthermore, the ability to discover multiple panoramas taken by a user over an extended period of time can be a big convenience.

To recognize panoramas, Brown and Lowe (2007) first find all pairwise image overlaps using a feature-based method and then find connected components in the overlap graph to “recognize” individual panoramas (Figure 8.11). The feature-based matching stage first extracts scale invariant feature transform (SIFT) feature locations and feature descriptors (Lowe 2004) from all the input images and places them in an indexing structure, as described in Section 7.1.3. For each image pair under consideration, the nearest matching neighbor is found for each feature in the first image, using the indexing structure to rapidly find candidates and then comparing feature descriptors to find the best match. RANSAC is used to find a set of *inlier* matches; pairs of matches are used to hypothesize similarity motion models that are then used to count the number of inliers. A RANSAC algorithm tailored specifically for rotational panoramas is described by Brown, Hartley, and Nistér (2007).

In practice, the most difficult part of getting a fully automated stitching algorithm to work is deciding which pairs of images actually correspond to the same parts of the scene. Repeated structures such as windows (Figure 8.12) can lead to false matches when using a feature-based approach. One way to mitigate this problem is to perform a direct pixel-based comparison between the registered images to determine if they actually are different views of the same scene. Unfortunately, this heuristic may fail if there are moving objects in the scene (Figure 8.13). While there is no magic bullet for this problem, short of full scene understanding, further improvements can likely be made by applying domain-specific heuristics, such as priors on typical camera motions as well as machine learning techniques

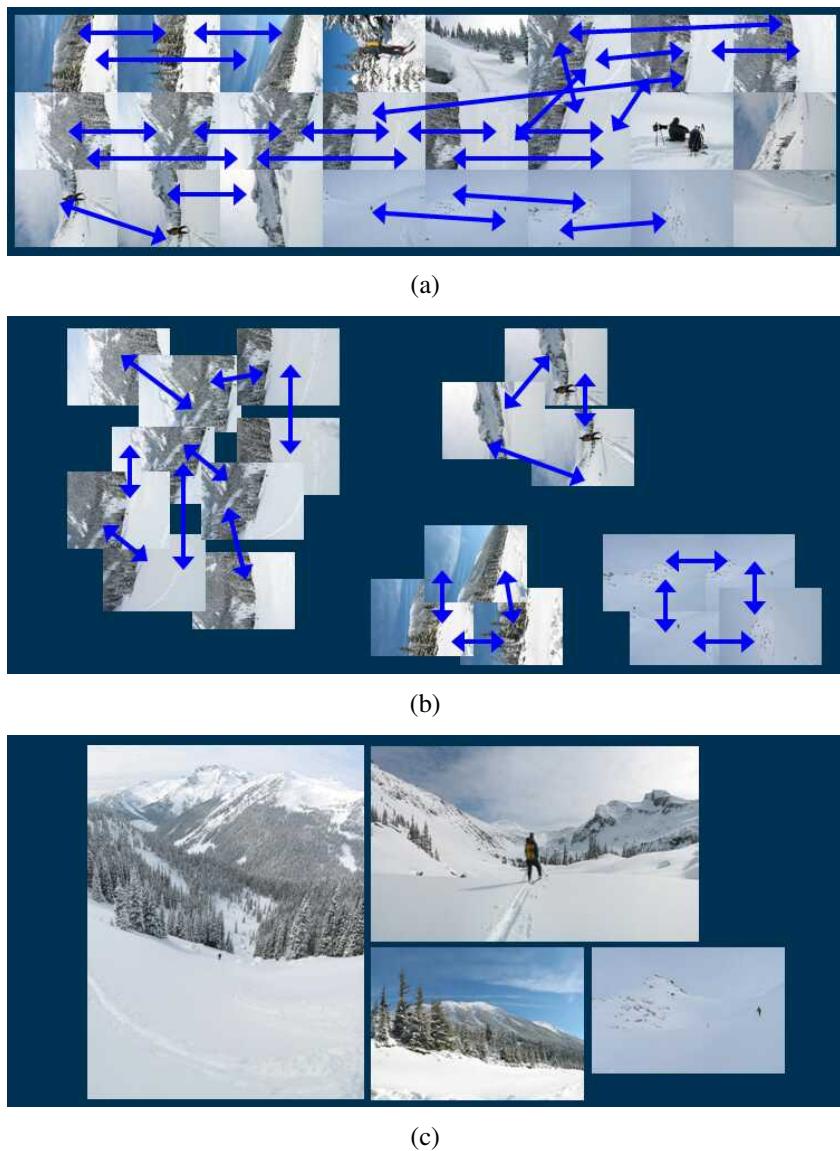


Figure 8.11 Recognizing panoramas (Brown, Szeliski, and Winder 2005), figures courtesy of Matthew Brown: (a) input images with pairwise matches; (b) images grouped into connected components (panoramas); (c) individual panoramas registered and blended into stitched composites.

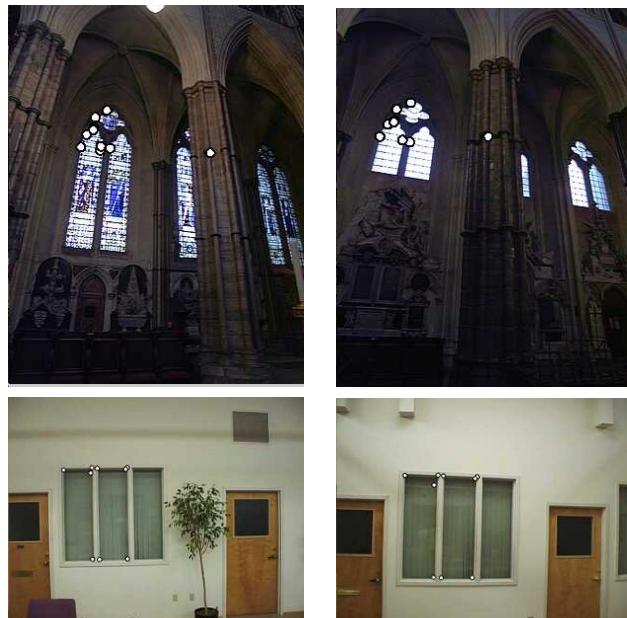


Figure 8.12 Matching errors (Brown, Szeliski, and Winder 2004): accidental matching of several features can lead to matches between pairs of images that do not actually overlap.

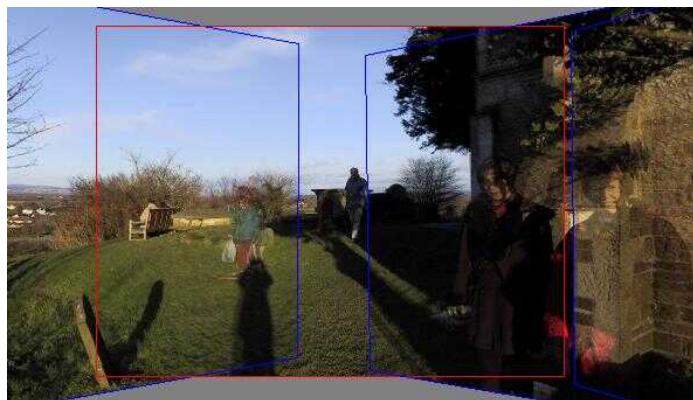


Figure 8.13 Validation of image matches by direct pixel error comparison can fail when the scene contains moving objects (Uyttendaele, Eden, and Szeliski 2001) © 2001 IEEE.

applied to the problem of match validation.

8.4 Compositing

Once we have registered all of the input images with respect to each other, we need to decide how to produce the final stitched mosaic image. This involves selecting a final compositing surface (flat, cylindrical, spherical, etc.) and view (reference image). It also involves selecting which pixels contribute to the final composite and how to optimally blend these pixels to minimize visible seams, blur, and ghosting.

In this section, we review techniques that address these problems, namely compositing surface parameterization, pixel and seam selection, blending, and exposure compensation. Our emphasis is on fully automated approaches to the problem. Since the creation of high-quality panoramas and composites is as much an artistic endeavor as a computational one, various interactive tools have been developed to assist this process (Agarwala, Dontcheva *et al.* 2004; Li, Sun *et al.* 2004; Rother, Kolmogorov, and Blake 2004). Some of these are covered in more detail in Section 10.4.

8.4.1 Choosing a compositing surface

The first choice to be made is how to represent the final image. If only a few images are stitched together, a natural approach is to select one of the images as the *reference* and to then warp all of the other images into its reference coordinate system. The resulting composite is sometimes called a *flat* panorama, since the projection onto the final surface is still a perspective projection, and hence straight lines remain straight (which is often a desirable attribute).²³

For larger fields of view, however, we cannot maintain a flat representation without excessively stretching pixels near the border of the image. (In practice, flat panoramas start to look severely distorted once the field of view exceeds 90° or so.) The usual choice for compositing larger panoramas is to use a cylindrical (Chen 1995; Szeliski 1996) or spherical (Szeliski and Shum 1997) projection, as described in Section 8.2.6. In fact, any surface used for *environment mapping* in computer graphics can be used, including a *cube map*, which represents the full viewing sphere with the six square faces of a cube (Greene 1986; Szeliski and Shum 1997). Cartographers have also developed a number of alternative methods for representing the globe (Bugayevskiy and Snyder 1995).

²³Techniques have also been developed to straighten curved lines in cylindrical and spherical panoramas (Carroll, Agrawala, and Agarwala 2009; Kopf, Lischinski *et al.* 2009; Carroll, Agarwala, and Agrawala 2010).

The choice of parameterization is somewhat application-dependent and involves a trade-off between keeping the local appearance undistorted (e.g., keeping straight lines straight) and providing a reasonably uniform sampling of the environment. Automatically making this selection and smoothly transitioning between representations based on the extent of the panorama is discussed in Kopf, Uyttendaele *et al.* (2007). A recent trend in panoramic photography has been the use of stereographic projections looking down at the ground (in an outdoor scene) to create “little planet” renderings.²⁴

View selection. Once we have chosen the output parameterization, we still need to determine which part of the scene will be *centered* in the final view. As mentioned above, for a flat composite, we can choose one of the images as a reference. Often, a reasonable choice is the one that is geometrically most central. For example, for rotational panoramas represented as a collection of 3D rotation matrices, we can choose the image whose z -axis is closest to the average z -axis (assuming a reasonable field of view). Alternatively, we can use the average z -axis (or quaternion, but this is trickier) to define the reference rotation matrix.

For larger, e.g., cylindrical or spherical, panoramas, we can use the same heuristic if a subset of the viewing sphere has been imaged. In the case of full 360° panoramas, a better choice is to choose the middle image from the sequence of inputs, or sometimes the first image, assuming this contains the object of greatest interest. In all of these cases, having the user control the final view is often highly desirable. If the “up vector” computation described in Section 8.3.1 is working correctly, this can be as simple as panning over the image or setting a vertical “center line” for the final panorama.

Coordinate transformations. After selecting the parameterization and reference view, we still need to compute the mappings between the input and output pixels coordinates.

If the final compositing surface is flat (e.g., a single plane or the face of a cube map) and the input images have no radial distortion, the coordinate transformation is the simple homography described by Equation (8.38). This kind of warping can be performed in graphics hardware by appropriately setting texture mapping coordinates and rendering a single quadrilateral.

If the final composite surface has some other analytic form (e.g., cylindrical or spherical), we need to convert every pixel in the final panorama into a viewing ray (3D point) and then map it back into each image according to the projection (and optionally radial distortion) equations. This process can be made more efficient by precomputing some lookup tables,

²⁴These are inspired by *The Little Prince* by Antoine De Saint-Exupéry. Go to <https://www.flickr.com> and search for “little planet projection”.

e.g., the partial trigonometric functions needed to map cylindrical or spherical coordinates to 3D coordinates or the radial distortion field at each pixel. It is also possible to accelerate this process by computing exact pixel mappings on a coarser grid and then interpolating these values.

When the final compositing surface is a texture-mapped polyhedron, a slightly more sophisticated algorithm must be used. Not only do the 3D and texture map coordinates have to be properly handled, but a small amount of *overdraw* outside the triangle footprints in the texture map is necessary, to ensure that the texture pixels being interpolated during 3D rendering have valid values (Szeliski and Shum 1997).

Sampling issues. While the above computations can yield the correct (fractional) pixel addresses in each input image, we still need to pay attention to sampling issues. For example, if the final panorama has a lower resolution than the input images, prefiltering the input images is necessary to avoid aliasing. These issues have been extensively studied in both the image processing and computer graphics communities. The basic problem is to compute the appropriate prefilter, which depends on the distance (and arrangement) between neighboring samples in a source image. As discussed in Sections 3.5.2 and 3.6.1, various approximate solutions, such as MIP mapping (Williams 1983) or elliptically weighted Gaussian averaging (Greene and Heckbert 1986) have been developed in the graphics community. For highest visual quality, a higher order (e.g., cubic) interpolator combined with a spatially adaptive prefilter may be necessary (Wang, Kang *et al.* 2001). Under certain conditions, it may also be possible to produce images with a higher resolution than the input images using the process of *super-resolution* (Section 10.3).

8.4.2 Pixel selection and weighting (deghosting)

Once the source pixels have been mapped onto the final composite surface, we must still decide how to blend them in order to create an attractive-looking panorama. If all of the images are in perfect registration and identically exposed, this is an easy problem, i.e., any pixel or combination will do. However, for real images, visible seams (due to exposure differences), blurring (due to misregistration), or ghosting (due to moving objects) can occur.

Creating clean, pleasing-looking panoramas involves both deciding which pixels to use and how to weight or blend them. The distinction between these two stages is a little fluid, since per-pixel weighting can be thought of as a combination of selection and blending. In this section, we discuss spatially varying weighting, pixel selection (seam placement), and then more sophisticated blending.

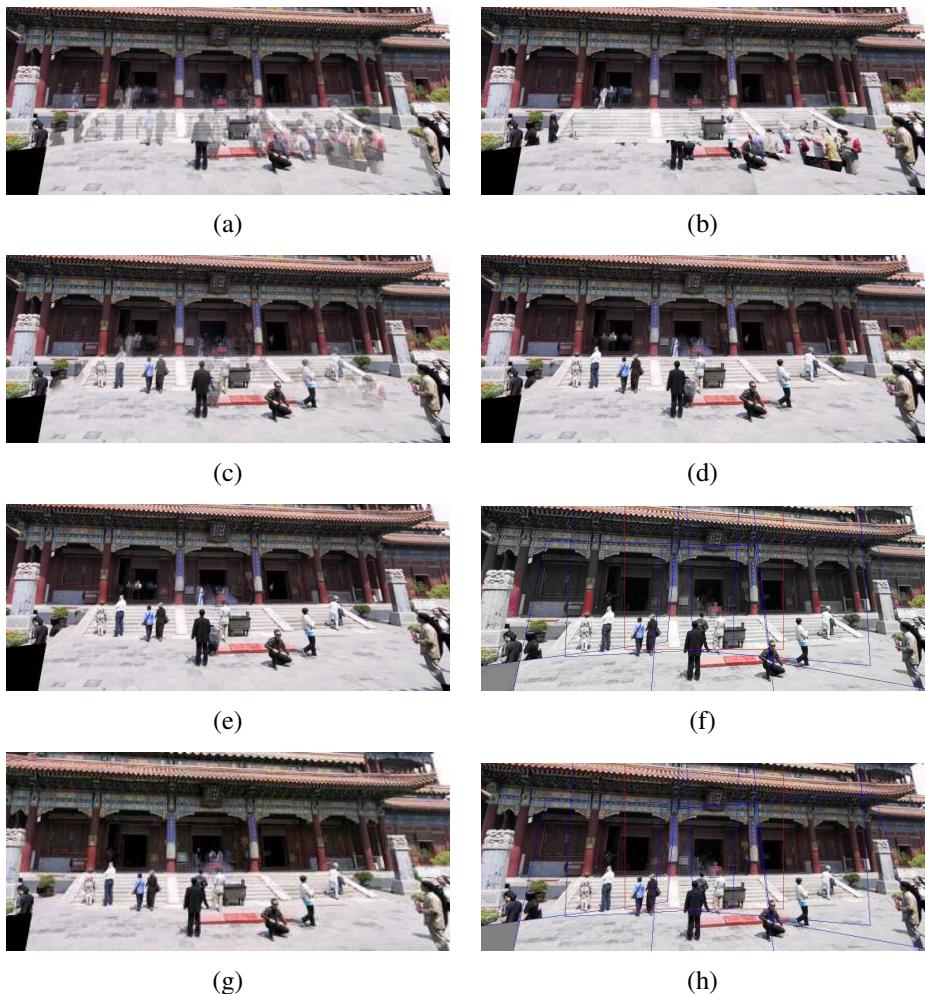


Figure 8.14 Final composites computed by a variety of algorithms (Szeliski 2006a): (a) average, (b) median, (c) feathered average, (d) p-norm $p = 10$, (e) Voronoi, (f) weighted ROD vertex cover with feathering, (g) graph cut seams with Poisson blending, and (h) with pyramid blending.

Feathering and center-weighting. The simplest way to create a final composite is to simply take an *average* value at each pixel,

$$C(\mathbf{x}) = \sum_k w_k(\mathbf{x}) \tilde{I}_k(\mathbf{x}) \Bigg/ \sum_k w_k(\mathbf{x}), \quad (8.70)$$

where $\tilde{I}_k(\mathbf{x})$ are the *warped* (re-sampled) images and $w_k(\mathbf{x})$ is 1 at valid pixels and 0 elsewhere. On computer graphics hardware, this kind of summation can be performed in an *accumulation buffer* (using the A channel as the weight).

Simple averaging usually does not work very well, since exposure differences, misregistrations, and scene movement are all very visible (Figure 8.14a). If rapidly moving objects are the only problem, taking a *median* filter (which is a kind of pixel selection operator) can often be used to remove them (Figure 8.14b) (Irani and Anandan 1998). Conversely, center-weighting (discussed below) and *minimum likelihood* selection (Agarwala, Dontcheva *et al.* 2004) can sometimes be used to retain multiple copies of a moving object (Figure 8.17).

A better approach to averaging is to weight pixels near the center of the image more heavily and to down-weight pixels near the edges. When an image has some cutout regions, down-weighting pixels near the edges of both cutouts and the image is preferable. This can be done by computing a *distance map* or *grassfire transform*,

$$w_k(\mathbf{x}) = \arg \min_{\mathbf{y}} \{ \|\mathbf{y}\| \mid \tilde{I}_k(\mathbf{x} + \mathbf{y}) \text{ is invalid} \}, \quad (8.71)$$

where each valid pixel is tagged with its Euclidean distance to the nearest invalid pixel (Section 3.3.3). The Euclidean distance map can be efficiently computed using a two-pass raster algorithm (Danielsson 1980; Borgefors 1986).

Weighted averaging with a distance map is often called *feathering* (Szeliski and Shum 1997; Chen and Klette 1999; Uyttendaele, Eden, and Szeliski 2001) and does a reasonable job of blending over exposure differences. However, blurring and ghosting can still be problems (Figure 8.14c). Note that weighted averaging is *not* the same as compositing the individual images with the classic *over* operation (Porter and Duff 1984; Blinn 1994a), even when using the weight values (normalized to sum up to one) as *alpha* (translucency) channels. This is because the over operation attenuates the values from more distant surfaces and, hence, is not equivalent to a direct sum.

One way to improve feathering is to raise the distance map values to some large power, i.e., to use $w_k^p(\mathbf{x})$ in Equation (8.70). The weighted averages then become dominated by the larger values, i.e., they act somewhat like a *p-norm*. The resulting composite can often provide a reasonable tradeoff between visible exposure differences and blur (Figure 8.14d).

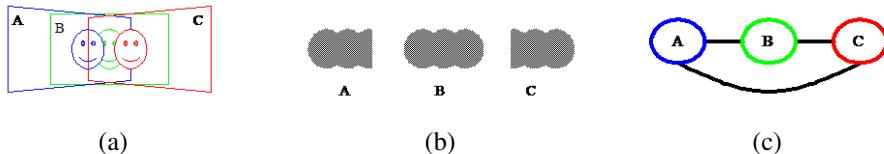


Figure 8.15 Computation of regions of difference (RODs) (Uyttendaele, Eden, and Szeliski 2001) © 2001 IEEE: (a) three overlapping images with a moving face; (b) corresponding RODs; (c) graph of coincident RODs.

In the limit as $p \rightarrow \infty$, only the pixel with the maximum weight is selected. This hard pixel selection process produces a visibility mask-sensitive variant of the familiar *Voronoi diagram*, which assigns each pixel to the nearest image center in the set (Wood, Finkelstein *et al.* 1997; Peleg, Rousso *et al.* 2000). The resulting composite, while useful for artistic guidance and in high-overlap panoramas (*manifold mosaics*) tends to have very hard edges with noticeable seams when the exposures vary (Figure 8.14e).

Xiong and Turkowski (1998) use this Voronoi idea (local maximum of the grassfire transform) to select seams for Laplacian pyramid blending (which is discussed below). However, since the seam selection is performed sequentially as new images are added in, some artifacts can occur.

Optimal seam selection. Computing the Voronoi diagram is one way to select the *seams* between regions where different images contribute to the final composite. However, Voronoi images totally ignore the local image structure underlying the seam. A better approach is to place the seams in regions where the images agree, so that transitions from one source to another are not visible. In this way, the algorithm avoids “cutting through” moving objects where a seam would look unnatural (Davis 1998). For a pair of images, this process can be formulated as a simple dynamic program starting from one edge of the overlap region and ending at the other (Milgram 1975, 1977; Davis 1998; Efros and Freeman 2001).

When multiple images are being composited, the dynamic program idea does not readily generalize. (For square texture tiles being composited sequentially, Efros and Freeman (2001) run a dynamic program along each of the four tile sides.)

To overcome this problem, Uyttendaele, Eden, and Szeliski (2001) observed that, for well-registered images, moving objects produce the most visible artifacts, namely translucent looking *ghosts*. Their system therefore decides which objects to keep and which ones to erase. First, the algorithm compares all overlapping input image pairs to determine *regions of difference* (RODs) where the images disagree. Next, a graph is constructed with the



Figure 8.16 *Photomontage (Agarwala, Dontcheva et al. 2004) © 2004 ACM.* From a set of five source images (of which four are shown on the left), Photomontage quickly creates a composite family portrait in which everyone is smiling and looking at the camera (right). Users simply flip through the stack and coarsely draw strokes using the designated source image objective over the people they wish to add to the composite. The user-applied strokes and computed regions (middle) are color-coded by the borders of the source images on the left.

RODs as vertices and edges representing ROD pairs that overlap in the final composite (Figure 8.15). Since the presence of an edge indicates an area of disagreement, vertices (regions) must be removed from the final composite until no edge spans a pair of remaining vertices. The smallest such set can be computed using a *vertex cover* algorithm. Since several such covers may exist, a *weighted vertex cover* is used instead, where the vertex weights are computed by summing the feather weights in the ROD (Uyttendaele, Eden, and Szeliski 2001). The algorithm therefore prefers removing regions that are near the edge of the image, which reduces the likelihood that partially visible objects will appear in the final composite. (It is also possible to infer which object in a region of difference is the foreground object by the “edginess” (pixel differences) across the ROD boundary, which should be higher when an object is present (Herley 2005).) Once the desired excess regions of difference have been removed, the final composite can be created by feathering (Figure 8.14f).

A different approach to pixel selection and seam placement is described by Agarwala, Dontcheva *et al.* (2004). Their system computes the label assignment that optimizes the sum of two objective functions. The first is a per-pixel *image objective* that determines which pixels are likely to produce good composites,

$$E_D = \sum_{\mathbf{x}} D(\mathbf{x}, l(\mathbf{x})), \quad (8.72)$$

where $D(\mathbf{x}, l)$ is the *data penalty* associated with choosing image l at pixel \mathbf{x} . In their system, users can select which pixels to use by “painting” over an image with the desired object or



Figure 8.17 Set of five photos tracking a snowboarder’s jump stitched together into a seamless composite. Because the algorithm prefers pixels near the center of the image, multiple copies of the boarder are retained.

appearance, which sets $D(\mathbf{x}, l)$ to a large value for all labels l other than the one selected by the user (Figure 8.16). Alternatively, automated selection criteria can be used, such as *maximum likelihood*, which prefers pixels that occur repeatedly in the background (for object removal), or *minimum likelihood* for objects that occur infrequently, i.e., for moving object retention. Using a more traditional center-weighted data term tends to favor objects that are centered in the input images (Figure 8.17).

The second term is a *seam objective* that penalizes differences in labels between adjacent images,

$$E_S = \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{N}} S(\mathbf{x}, \mathbf{y}, l(\mathbf{x}), l(\mathbf{y})), \quad (8.73)$$

where $S(\mathbf{x}, \mathbf{y}, l_x, l_y)$ is the image-dependent *interaction penalty* or *seam cost* of placing a seam between pixels \mathbf{x} and \mathbf{y} , and \mathcal{N} is the set of \mathcal{N}_4 neighboring pixels. For example, the simple color-based seam penalty used in Kwatra, Schödl *et al.* (2003) and Agarwala, Dontcheva *et al.* (2004) can be written as

$$S(\mathbf{x}, \mathbf{y}, l_x, l_y) = \|\tilde{I}_{l_x}(\mathbf{x}) - \tilde{I}_{l_y}(\mathbf{x})\| + \|\tilde{I}_{l_x}(\mathbf{y}) - \tilde{I}_{l_y}(\mathbf{y})\|. \quad (8.74)$$

More sophisticated seam penalties can also look at image gradients or the presence of image edges (Agarwala, Dontcheva *et al.* 2004). Seam penalties are widely used in other computer vision applications such as stereo matching (Boykov, Veksler, and Zabih 2001) to give the labeling function its *coherence* or *smoothness*. An alternative approach, which places seams along strong consistent edges in overlapping images using a watershed computation is described by Soille (2006).

The sum of these two objective functions gives rise to a *Markov random field* (MRF), for which good optimization algorithms are described in Sections 4.3 and 4.3.2 and Appendix B.5. For label computations of this kind, the α -expansion algorithm developed by Boykov, Veksler, and Zabih (2001) works particularly well (Szeliski, Zabih *et al.* 2008).

For the result shown in Figure 8.14g, Agarwala, Dontcheva *et al.* (2004) use a large data penalty for invalid pixels and 0 for valid pixels. Notice how the seam placement algorithm avoids regions of difference, including those that border the image and that might result in objects being cut off. Graph cuts (Agarwala, Dontcheva *et al.* 2004) and vertex cover (Uyttendaele, Eden, and Szeliski 2001) often produce similar looking results, although the former is significantly slower since it optimizes over all pixels, while the latter is more sensitive to the thresholds used to determine regions of difference. More recent approaches to seam selection include SEAGULL (Lin, Jiang *et al.* 2016), which jointly optimizes local alignment and seam selection, and object-centered image stitching (Herrmann, Wang *et al.* 2018a), which uses an off-the-shelf object detector to avoid cutting through objects.

8.4.3 Application: Photomontage

While image stitching is normally used to composite partially overlapping photographs, it can also be used to composite repeated shots of a scene taken with the aim of obtaining the best possible composition and appearance of each element.

Figure 8.16 shows the *Photomontage* system developed by Agarwala, Dontcheva *et al.* (2004), where users draw strokes over a set of pre-aligned images to indicate which regions they wish to keep from each image. Once the system solves the resulting multi-label graph cut (8.72–8.73), the various pieces taken from each source photo are blended together using a variant of Poisson image blending (8.75–8.77). Their system can also be used to automatically composite an all-focus image from a series of bracketed focus images (Hasinoff, Kutulakos *et al.* 2009) or to remove wires and other unwanted elements from sets of photographs. Exercise 8.14 has you implement this system and try out some of its variants.

8.4.4 Blending

Once the seams between images have been determined and unwanted objects removed, we still need to blend the images to compensate for exposure differences and other misalignments. The spatially varying weighting (feathering) previously discussed can often be used to accomplish this. However, it is difficult in practice to achieve a pleasing balance between smoothing out low-frequency exposure variations and retaining sharp enough transitions to prevent blurring (although using a high exponent in feathering can help).

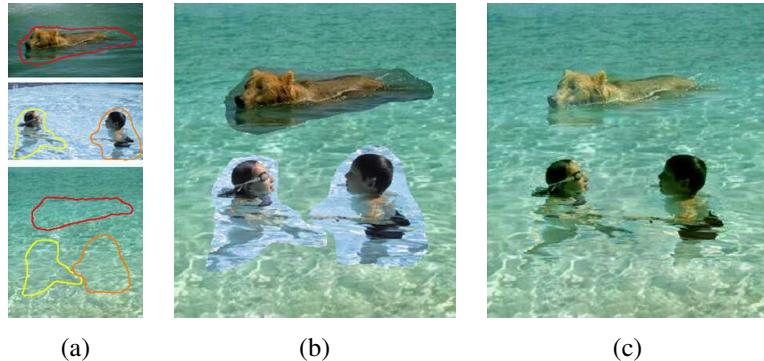


Figure 8.18 *Poisson image editing* (Pérez, Gangnet, and Blake 2003) © 2003 ACM: (a) The dog and the two children are chosen as source images to be pasted into the destination swimming pool. (b) Simple pasting fails to match the colors at the boundaries, whereas (c) Poisson image blending masks these differences.

Laplacian pyramid blending. An attractive solution to this problem is the Laplacian pyramid blending technique developed by Burt and Adelson (1983b), which we discussed in Section 3.5.5. Instead of using a single transition width, a frequency-adaptive width is used by creating a band-pass (Laplacian) pyramid and making the transition widths within each level a function of the level, i.e., the same width in pixels. In practice, a small number of levels, i.e., as few as two (Brown and Lowe 2007), may be adequate to compensate for differences in exposure. The result of applying this pyramid blending is shown in Figure 8.14h.

Gradient domain blending. An alternative approach to multi-band image blending is to perform the operations in the *gradient domain*. Reconstructing images from their gradient fields has a long history in computer vision (Horn 1986), starting originally with work in brightness constancy (Horn 1974), shape from shading (Horn and Brooks 1989), and photometric stereo (Woodham 1981). Related ideas have also been used for reconstructing images from their edges (Elder and Goldberg 2001), removing shadows from images (Weiss 2001), separating reflections from a single image (Levin, Zomet, and Weiss 2004; Levin and Weiss 2007), and *tone mapping* high dynamic range images by reducing the magnitude of image edges (gradients) (Fattal, Lischinski, and Werman 2002).

Pérez, Gangnet, and Blake (2003) show how gradient domain reconstruction can be used to do seamless object insertion in image editing applications (Figure 8.18). Rather than copying pixels, the *gradients* of the new image fragment are copied instead. The actual pixel values for the copied area are then computed by solving a *Poisson equation* that locally matches the

gradients while obeying the fixed *Dirichlet* (exact matching) conditions at the seam boundary. Pérez, Gangnet, and Blake (2003) show that this is equivalent to computing an additive *membrane* interpolant of the mismatch between the source and destination images along the boundary.²⁵ In earlier work, Peleg (1981) also proposed adding a smooth function to enforce consistency along the seam curve.

Agarwala, Dontcheva *et al.* (2004) extended this idea to a multi-source formulation, where it no longer makes sense to talk of a destination image whose exact pixel values must be matched at the seam. Instead, *each* source image contributes its own gradient field and the Poisson equation is solved using *Neumann* boundary conditions, i.e., dropping any equations that involve pixels outside the boundary of the image.

Rather than solving the Poisson partial differential equations, Agarwala, Dontcheva *et al.* (2004) directly minimize a *variational problem*,

$$\min_{C(\mathbf{x})} \|\nabla C(\mathbf{x}) - \nabla \tilde{I}_{l(\mathbf{x})}(\mathbf{x})\|^2. \quad (8.75)$$

The discretized form of this equation is a set of gradient constraint equations

$$C(\mathbf{x} + \hat{i}) - C(\mathbf{x}) = \tilde{I}_{l(\mathbf{x})}(\mathbf{x} + \hat{i}) - \tilde{I}_{l(\mathbf{x})}(\mathbf{x}) \quad \text{and} \quad (8.76)$$

$$C(\mathbf{x} + \hat{j}) - C(\mathbf{x}) = \tilde{I}_{l(\mathbf{x})}(\mathbf{x} + \hat{j}) - \tilde{I}_{l(\mathbf{x})}(\mathbf{x}), \quad (8.77)$$

where $\hat{i} = (1, 0)$ and $\hat{j} = (0, 1)$ are unit vectors in the x and y directions.²⁶ They then solve the associated sparse least squares problem. Since this system of equations is only defined up to an additive constraint, Agarwala, Dontcheva *et al.* (2004) ask the user to select the value of one pixel. In practice, a better choice might be to weakly bias the solution towards reproducing the original color values.

In order to accelerate the solution of this sparse linear system, Fattal, Lischinski, and Werman (2002) use multigrid, whereas Agarwala, Dontcheva *et al.* (2004) use hierarchical basis preconditioned conjugate gradient descent (Szeliski 1990b, 2006b; Krishnan and Szeliski 2011; Krishnan, Fattal, and Szeliski 2013) (Appendix A.5). In subsequent work, Agarwala (2007) shows how using a quadtree representation for the solution can further accelerate the computation with minimal loss in accuracy, while Szeliski, Uyttendaele, and Steedly (2008) show how representing the per-image offset fields using coarser splines is even faster. This latter work also argues that blending in the log domain, i.e., using multiplicative rather than additive offsets, is preferable, as it more closely matches texture contrasts across seam boundaries. The resulting seam blending works very well in practice (Figure 8.14h), although care

²⁵The membrane interpolant is known to have nicer interpolation properties for arbitrary-shaped constraints than frequency-domain interpolants (Nielsen 1993).

²⁶At seam locations, the right-hand side is replaced by the average of the gradients in the two source images.

must be taken when copying large gradient values near seams so that a “double edge” is not introduced.

Copying gradients directly from the source images after seam placement is just one approach to gradient domain blending. The paper by Levin, Zomet *et al.* (2004) examines several different variants of this approach, which they call *Gradient-domain Image STitching* (GIST). The techniques they examine include feathering (blending) the gradients from the source images, as well as using an L_1 norm in performing the reconstruction of the image from the gradient field, rather than using an L_2 norm as in Equation (8.75). Their preferred technique is the L_1 optimization of a feathered (blended) cost function on the original image gradients (which they call GIST1- l_1). Since L_1 optimization using linear programming can be slow, they develop a faster iterative median-based algorithm in a multigrid framework. Visual comparisons between their preferred approach and what they call *optimal seam on the gradients* (which is equivalent to the approach of Agarwala, Dontcheva *et al.* (2004)) show similar results, while significantly improving on pyramid blending and feathering algorithms.

Exposure compensation. Pyramid and gradient domain blending can do a good job of compensating for moderate amounts of exposure differences between images. However, when the exposure differences become large, alternative approaches may be necessary.

Uyttendaele, Eden, and Szeliski (2001) iteratively estimate a local correction between each source image and a blended composite. First, a block-based quadratic transfer function is fit between each source image and an initial feathered composite. Next, transfer functions are averaged with their neighbors to get a smoother mapping and per-pixel transfer functions are computed by *splining* (interpolating) between neighboring block values. Once each source image has been smoothly adjusted, a new feathered composite is computed and the process is repeated (typically three times). The results shown by Uyttendaele, Eden, and Szeliski (2001) demonstrate that this does a better job of exposure compensation than simple feathering and can handle local variations in exposure due to effects such as lens vignetting.

Ultimately, however, the most principled way to deal with exposure differences is to stitch images in the radiance domain, i.e., to convert each image into a radiance image using its exposure value and then create a stitched, high dynamic range image, as discussed in Section 10.2 and Eden, Uyttendaele, and Szeliski (2006).

8.5 Additional reading

Hartley and Zisserman (2004) provide a wonderful introduction to the topics of feature-based alignment and optimal motion estimation. Techniques for robust estimation are discussed

in more detail in Appendix B.3 and in monographs and review articles on this topic (Huber 1981; Hampel, Ronchetti *et al.* 1986; Rousseeuw and Leroy 1987; Black and Rangarajan 1996; Stewart 1999). The most commonly used robust initialization technique in computer vision is RANdom SAmple Consensus (RANSAC) (Fischler and Bolles 1981), which has spawned a series of more efficient variants (Torr and Zisserman 2000; Nistér 2003; Chum and Matas 2005; Raguram, Chum *et al.* 2012; Brachmann, Krull *et al.* 2017; Barath and Matas 2018; Barath, Matas, and Noskova 2019; Brachmann and Rother 2019). The MAGSAC++ paper by Barath, Noskova *et al.* (2020) compares many of these variants.

The literature on image stitching dates back to work in the photogrammetry community in the 1970s (Milgram 1975, 1977; Slama 1980). In computer vision, papers started appearing in the early 1980s (Peleg 1981), while the development of fully automated techniques came about a decade later (Mann and Picard 1994; Chen 1995; Szeliski 1996; Szeliski and Shum 1997; Sawhney and Kumar 1999; Shum and Szeliski 2000). Those techniques used direct pixel-based alignment but feature-based approaches are now the norm (Zoghlaei, Faugeras, and Deriche 1997; Capel and Zisserman 1998; Cham and Cipolla 1998; Badra, Qumsieh, and Dudek 1998; McLauchlan and Jaenicke 2002; Brown and Lowe 2007). A collection of some of these papers can be found in the book by Benosman and Kang (2001). Szeliski (2006a) provides a comprehensive survey of image stitching, on which the material in this chapter is based. More recent publications include Zaragoza, Chin *et al.* (2013), Zhang and Liu (2014), Lin, Pankanti *et al.* (2015), Lin, Jiang *et al.* (2016), Herrmann, Wang *et al.* (2018b), Lee and Sim (2020), and Zhuang and Tran (2020).

High-quality techniques for optimal seam finding and blending are another important component of image stitching systems. Important developments in this field include work by Milgram (1977), Burt and Adelson (1983b), Davis (1998), Uyttendaele, Eden, and Szeliski (2001), Pérez, Gangnet, and Blake (2003), Levin, Zomet *et al.* (2004), Agarwala, Dontcheva *et al.* (2004), Eden, Uyttendaele, and Szeliski (2006), Kopf, Uyttendaele *et al.* (2007), Lin, Jiang *et al.* (2016), and Herrmann, Wang *et al.* (2018a).

In addition to the merging of multiple overlapping photographs taken for aerial or terrestrial panoramic image creation, stitching techniques can be used for automated white-board scanning (He and Zhang 2005; Zhang and He 2007), scanning with a mouse (Nakao, Kashitani, and Kaneyoshi 1998), and retinal image mosaics (Can, Stewart *et al.* 2002). They can also be applied to video sequences (Teodosio and Bender 1993; Irani, Hsu, and Anandan 1995; Kumar, Anandan *et al.* 1995; Sawhney and Ayer 1996; Massey and Bender 1996; Irani and Anandan 1998; Sawhney, Arpa *et al.* 2002; Agarwala, Zheng *et al.* 2005; Rav-Acha, Pritch *et al.* 2005; Steedly, Pal, and Szeliski 2005; Baudisch, Tan *et al.* 2006) and can even be used for video compression (Lee, Chen *et al.* 1997).

8.6 Exercises

Ex 8.1: Feature-based image alignment for flip-book animations. Take a set of photos of an action scene or portrait (preferably in burst shooting mode) and align them to make a composite or flip-book animation.

1. Extract features and feature descriptors using some of the techniques described in Sections 7.1.1–7.1.2.
2. Match your features using nearest neighbor matching with a nearest neighbor distance ratio test (7.18).
3. Compute an optimal 2D translation and rotation between the first image and all subsequent images, using least squares (Section 8.1.1) with optional RANSAC for robustness (Section 8.1.4).
4. Resample all of the images onto the first image’s coordinate frame (Section 3.6.1) using either bilinear or bicubic resampling and optionally crop them to their common area.
5. Convert the resulting images into an animated GIF (using software available from the web) or optionally implement cross-dissolves to turn them into a “slo-mo” video.
6. (Optional) Combine this technique with feature-based (Exercise 3.25) morphing.

Ex 8.2: Panography. Create the kind of panograph discussed in Section 8.1.2 and commonly found on the web.

1. Take a series of interesting overlapping photos.
2. Use the feature detector, descriptor, and matcher developed in Exercises 7.1–7.4 (or existing software) to match features among the images.
3. Turn each connected component of matching features into a *track*, i.e., assign a unique index i to each track, discarding any tracks that are inconsistent (contain two different features in the same image).
4. Compute a global translation for each image using Equation (8.12).
5. Since your matches probably contain errors, turn the above least square metric into a robust metric (8.25) and re-solve your system using iteratively reweighted least squares.
6. Compute the size of the resulting composite canvas and resample each image into its final position on the canvas. (Keeping track of bounding boxes will make this more efficient.)

7. Average all of the images, or choose some kind of ordering and implement translucent *over* compositing (3.8).
8. (Optional) Extend your parametric motion model to include rotations and scale, i.e., the similarity transform given in Table 8.1. Discuss how you could handle the case of translations and rotations only (no scale).
9. (Optional) Write a simple tool to let the user adjust the ordering and opacity, and add or remove images.
10. (Optional) Write down a different least squares problem that involves pairwise matching of images. Discuss why this might be better or worse than the global matching formula given in (8.12).

Ex 8.3: 2D rigid/Euclidean matching. Several alternative approaches are given in Section 8.1.3 for estimating a 2D rigid (Euclidean) alignment.

1. Implement the various alternatives and compare their accuracy on synthetic data, i.e., random 2D point clouds with noisy feature positions.
2. One approach is to estimate the translations from the centroids and then estimate rotation in polar coordinates. Do you need to weight the angles obtained from a polar decomposition in some way to get the statistically correct estimate?
3. How can you modify your techniques to take into account either scalar (8.10) or full two-dimensional point covariance weightings (8.11)? Do all of the previously developed “shortcuts” still work or does full weighting require iterative optimization?

Ex 8.4: 2D match move/augmented reality. Replace a picture in a magazine or a book with a different image or video.

1. Take a picture of a magazine or book page.
2. Outline a figure or picture on the page with a rectangle, i.e., draw over the four sides as they appear in the image.
3. Match features in this area with each new image frame.
4. Replace the original image with an “advertising” insert, warping the new image with the appropriate homography.
5. Try your approach on a clip from a sporting event (e.g., indoor or outdoor soccer) to implement a billboard replacement.

Ex 8.5: Direct pixel-based alignment. Take a pair of images, compute a coarse-to-fine affine alignment (Exercise 9.2) and then blend them using either averaging (Exercise 8.2) or a Laplacian pyramid (Exercise 3.18). Extend your motion model from affine to perspective (homography) to better deal with rotational mosaics and planar surfaces seen under arbitrary motion.

Ex 8.6: Featured-based stitching. Extend your feature-based alignment technique from Exercise 8.2 to use a full perspective model and then blend the resulting mosaic using either averaging or more sophisticated distance-based feathering (Exercise 8.13).

Ex 8.7: Cylindrical strip panoramas. To generate cylindrical or spherical panoramas from a horizontally panning (rotating) camera, it is best to use a tripod. Set your camera up to take a series of 50% overlapped photos and then use the following steps to create your panorama:

1. Estimate the amount of radial distortion by taking some pictures with lots of long straight lines near the edges of the image and then using the plumb-line method from Exercise 11.5.
2. Compute the focal length either by using a ruler and paper (Debevec, Wenger *et al.* 2002) or by rotating your camera on the tripod, overlapping the images by exactly 0% and counting the number of images it takes to make a 360° panorama.
3. Convert each of your images to cylindrical coordinates using (8.45–8.49).
4. Line up the images with a translational motion model using either a direct pixel-based technique, such as coarse-to-fine incremental or an FFT, or a feature-based technique.
5. (Optional) If doing a complete 360° panorama, align the first and last images. Compute the amount of accumulated vertical misregistration and re-distribute this among the images.
6. Blend the resulting images using feathering or some other technique.

Ex 8.8: Coarse alignment. Use FFT or phase correlation (Section 9.1.2) to estimate the initial alignment between successive images. How well does this work? Over what range of overlaps? If it does not work, does aligning sub-sections (e.g., quarters) do better?

Ex 8.9: Automated mosaicing. Use feature-based alignment with four-point RANSAC for homographies (Section 8.1.3, Equations (8.19–8.23)) or three-point RANSAC for rotational motions (Brown, Hartley, and Nistér 2007) to match up all pairs of overlapping images.

Merge these pairwise estimates together by finding a spanning tree of pairwise relations. Visualize the resulting global alignment, e.g., by displaying a blend of each image with all other images that overlap it.

For greater robustness, try multiple spanning trees (perhaps randomly sampled based on the confidence in pairwise alignments) to see if you can recover from bad pairwise matches (Zach, Klopschitz, and Pollefeys 2010). As a measure of fitness, count how many pairwise estimates are consistent with the global alignment.

Ex 8.10: Global optimization. Use the initialization from the previous algorithm to perform a full bundle adjustment over all of the camera rotations and focal lengths, as described in Section 11.4.2 and by Shum and Szeliski (2000). Optionally, estimate radial distortion parameters as well or support fisheye lenses (Section 2.1.5).

As in the previous exercise, visualize the quality of your registration by creating composites of each input image with its neighbors, optionally blinking between the original image and the composite to better see misalignment artifacts.

Ex 8.11: Deghosting. Use the results of the previous bundle adjustment to predict the location of each feature in a consensus geometry. Use the difference between the predicted and actual feature locations to correct for small misregistrations, as described in Section 8.3.2 (Shum and Szeliski 2000).

Ex 8.12: Compositing surface. Choose a compositing surface (Section 8.4.1), e.g., a single reference image extended to a larger plane, a sphere represented using cylindrical or spherical coordinates, a stereographic “little planet” projection, or a cube map.

Project all of your images onto this surface and blend them with equal weighting, for now (just to see where the original image seams are).

Ex 8.13: Feathering and blending. Compute a feather (distance) map for each warped source image and use these maps to blend the warped images.

Alternatively, use Laplacian pyramid blending (Exercise 3.18) or gradient domain blending.

Ex 8.14: Photomontage and object removal. Implement a “Photomontage” system in which users can indicate desired or unwanted regions in pre-registered images using strokes or other primitives (such as bounding boxes).

(Optional) Devise an automatic moving objects remover (or “keeper”) by analyzing which inconsistent regions are more or less typical given some consensus (e.g., median filtering) of the aligned images. Figure 8.17 shows an example where the moving object was kept. Try

to make this work for sequences with large amounts of overlaps and consider averaging the images to make the moving object look more ghosted.

