CSE 252C: Computer Vision III

Lecturer: Serge Belongie
Scribes: Andrew Rabinovich and Vincent Rabaud
Edited by: Catherine Wah

LECTURE 3
Distributions and Histograms-of-X

3.1. Histograms

Used fairly frequently in computer vision in the late 1990s, histograms are a convenient way of packaging a distribution. A *histogram* (Fig. 1) is a non-parametric estimate of a density, in contrast to a parametric approach such as fitting a Gaussian to a sample of data, which is completely defined by two parameters.

Histograms, on the other hand, are defined by their bin sizes. Similar to choosing k or σ, bin size selection is a hard problem. A related idea is the kernel density estimate, which puts a kernel (*e.g.* a Gaussian) around each data point. Instead of binning, we must choose an appropriate kernel width. Such model selection problems are often solved in practice using cross validation and/or information theoretic methods.

In a nutshell, histograms are relevant to object recognition in that they allow us to do *recognition without feature correspondence* (Schiele and Crowley, 2000). At first glance, this seems surprising, since histograms (viz. global

[1] Department of Computer Science and Engineering, University of California, San Diego.

August 10, 2009

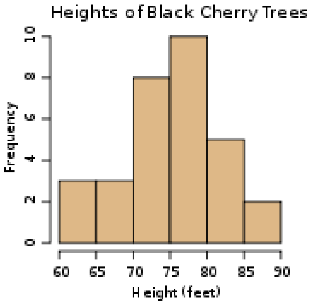

Figure 1. An example histogram of the heights of 31 black cherry trees. (http://en.wikipedia.org/wiki/Histogram)

histograms) discard so much information. Nevertheless, this is significant because correspondence (*i.e.* between image features in template and model) is hard. In principle, recognition with correspondence is more precise, for example, with faces and facial features.

Indeed, histogram (or distribution) based matching is weaker than correspondence based matching, but having proper training data for the latter is quite onerous in practice. Perhaps because histograms are so user friendly, computer vision researchers have pushed hard to apply them in innovative ways that go beyond simple things like color.

3.1.1. Types

Let's take a look at some examples of histograms.

Definition 3.1. A *Color histogram* can represent distributions in one of many possible color spaces.

Definition 3.2. In a *marginal histogram*, each axis of multi-dimensional data is considered independently. For example, in an RGB image, three histograms are built, one for each of the channels.

Definition 3.3. In a *joint histogram*, one multi-dimensional histogram is built, but it is subject to the curse of dimensionality. While more informative, the bins are more likely to be impoverished (*i.e.* empty) in this case.

The use of color histograms had a big impact in the early days of CBIR; one of these systems was Blobworld (Carson et al., 1999).

3.2. Comparing Histograms

In order to leverage (color) histograms for these purposes, we need a means of comparing them. How do we do this? This is the problem of computing the distance, dissimilarity, or divergence between two distributions. There are many options; which one to use depends on certain characteristics of the distributions. Broadly, these dissimilarity measures fall into four categories:

(a) Heuristic histogram distances
(b) Nonparametric test statistics
(c) Information theoretic divergences
(d) Ground distances

In our discussion of dissimilarity measures, we denote $\boldsymbol{h}$ as a normalized histogram with nonnegative values and k bins, where $\sum_{i=1}^{k} h_i = 1$; $\boldsymbol{g}$ is a second such histogram. These represent histogram density estimates of some random variables G and H.

3.2.1. Heuristic Histogram Distances

Minkowski or L_p distance.

$$D(\boldsymbol{g},\boldsymbol{h}) = \left(\sum_i |g_i - h_i|^p\right)^{\frac{1}{p}}, 1 \leq p \leq \infty \tag{3.4}$$

Some examples:

- $p = 1$: for color histograms (Swain and Ballard, *IJCV*'91)
- $p = 2$: used for SIFT descriptors (Lowe, *ICCV*'99)
- $p = \infty$: used for texture filter responses (Voorhees and Poggio, *Nature*'88)

This category, which emerged from CBIR literature, does not “respect” that $\boldsymbol{h}$ and $\boldsymbol{g}$ are distributions, but it often works well in practice.

Weighted Mean Variance (WMV). (Manjunath and Ma, *PAMI*'96)

$$D(H,G) = \frac{|\mu(H) - \mu(G)|}{|\sigma(\mu_{all})|} + \frac{|\sigma(H) - \sigma(G)|}{|\sigma(\sigma_{all})|} \tag{3.5}$$

The WMV does not use histograms, using mean and standard deviation instead, *i.e.* the first and second moments of the distribution, in a weighted combination. Denominators are standardized for μ and σ over some large database. If there is some benefit in the higher moments, however, then WMV will miss out.

3.2.2. Nonparametric Test Statistics

Kolmogorov-Smirnow (KS) distance. Based on c.d.f.s, the KS distance (Fig. 2) can only be used on marginal distributions, but when applicable, it is very elegant and effective (using $\boldsymbol{G}$ and $\boldsymbol{H}$ for c.d.f.s):

$$D(\boldsymbol{G},\boldsymbol{H}) = \max_i |G_i - H_i| = \|\boldsymbol{G} - \boldsymbol{H}\|_\infty. \tag{3.6}$$

This comparison is binning independent.

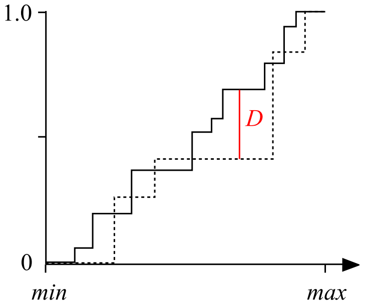

Figure 2. An illustration of a KS distance determination, where D is the maximum distance.

Cramer/vonMises (CvM). While also based on c.d.f.s, it uses the L_2 norm instead:

$$D(\boldsymbol{G},\boldsymbol{H}) = \|\boldsymbol{G}-\boldsymbol{H}\|_2^2 = \sum_i (G_i - H_i)^2. \tag{3.7}$$

χ^2 *statistic.* Employed by (Puzicha et al., *CVPR*'97), this statistic is very important, widely used, and highly effective. It can also be used on joint histograms:

$$D(\boldsymbol{g},\boldsymbol{h}) = \frac{1}{2}\sum_i \frac{(h_i-g_i)^2}{h_i+g_i} \in [0,1]\,. \tag{3.8}$$

The χ^2 statistic can be thought of as a re-weighted version of the L_2 norm, where a given bin difference is counted less if the average bin count is large.

The symmetric version of the more familiar χ^2 test between an empirical histogram and a theoretical distribution, with samples at t_i, is:

$$D(\boldsymbol{g},\boldsymbol{t}) = \frac{1}{2}\sum_i \frac{(h_i-t_i)^2}{t_i}. \tag{3.9}$$

If these per-bin errors are normally distributed, then this distance will have a χ^2 distribution with $K-C$ degrees of freedom, where C depends on the number of parameters estimated. We can use this for formal hypothesis testing.

We observe that this is a special case of the gamma distribution. Given K i.i.d. random variables $X_i \sim \mathcal{N}(0,1)$, the r.v. $Q = \sum_i \chi_i^2$ is a χ^2 distribution with K degrees of freedom.

For $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, we form

$$\sum_{i=1}^{K} \left(\frac{X_i-\mu_i}{\sigma_i}\right)^2, \tag{3.10}$$

an example of "z-scoring."

3.2.3. Information Theoretic Divergences

Kullback Leibler (KL) Divergence.

$$D(\boldsymbol{h}, \boldsymbol{g}) = \sum_i h_i \log \frac{h_i}{g_i} \tag{3.11}$$

KL measures how inefficient on average it would be to code a histogram ($\boldsymbol{h}$) using the other ($\boldsymbol{g}$) as the true distribution for the coding. One of its problems is that it is asymmetric.

Jeffrey or Jensen-Shannon. On the other hand, we have a symmetrized version of KL:

$$D(\boldsymbol{h}, \boldsymbol{g}) = \sum_i h_i \log \frac{h_i}{h_i + g_i} + g_i \log \frac{g_i}{h_i + g_i}. \tag{3.12}$$

Both of these are generally not used because empirical studies have shown no benefit here over χ^2.

3.2.4. Ground Distances

Quadratic Form. (Hafner et al., 1995)

$$D^2(\boldsymbol{h}, \boldsymbol{g}) = (\boldsymbol{h} - \boldsymbol{g})^\top A(\boldsymbol{h} - \boldsymbol{g}), \tag{3.13}$$

where $A_{ij} \in [0, 1]$ encodes the cross-bin similarity, *e.g.* the perceptual similarity between color bins, which decreases sensitivity to binning. Note that since $\boldsymbol{h}$ and $\boldsymbol{g}$ are normalized histograms, D^2 will be positive semidefinite even if A is indefinite (normally, we would need A to have all positive eigenvalues, etc.).

Earth Mover's Distance (EMD). The EMD is very interesting since it can operate on a pair of distributions with different numbers of bins, often adapted to the distributions. It is based on a linear optimization problem called the "transportation problem." We'll skip over the formulation, but the intuition is to make one distribution into hills of dirt and the other into holes. The EMD is the cost of optimally pushing the dirt into the holes. While comparatively expensive, it performs well on very compact representations of the distributions. For a comparative study, refer to (Rubner et al., 2001).

3.3. Histograms-of-X

Now that we've seen a wide variety of methods for comparing histograms, we note that, while it is not the best in all cases, χ^2 does a good job in practice.

Let us return to the histograms themselves and what image characteristics people have chosen to represent in this way:

- color - there are many color spaces

- texture (requires filtering) - filter responses, arranged into channels
- shape - using edges, gradient angles

For each of these characteristics, there are many instances of both marginal and joint histograms. In the latter case, in order to avoid the curse of dimensionality, adaptive binning is often used, *e.g.* computed via k-means (colorons, textons, shapemes, motons, etc.). In such cases, the histogram bins are not neatly ordered along a line, and KS is not applicable. We will defer discussion on image filtering to later in the course.

The upshot is that histograms are powerful descriptors, benefitting from not needing correspondences, but they are limited with respect to conceptual resolution: the equivalence class of images in existence with less than a given histogram distance may be unacceptably cluttered with irrelevant images. Eventually, some geometry will be needed.