Kolmogorov-Smirnov Test for Image Comparison

Eugene Demidenko

Dartmouth College, Hanover, NH 03755, USA
eugened@dartmouth.edu
http://www.dartmouth.edu/~eugened

Abstract. We apply the Kolmogorov-Smirnov test to test whether two distributions of 256 gray intensities are the same. Thus, this test may be useful to compare unstructured images, such as microscopic images in medicine. Usually, histogram is used to show the distribution of gray level intensities. We argue that cumulative distribution function (gray distribution) may be more informative when comparing several gray images. The Kolmogorov-Smirnov test is illustrated by hystology images from untreated and treated breast cancer tumors. The test is generalized to ensembles of gray images. Limitations of the Kolmogorov-Smirnov test are discussed.

1 Introduction

An essential task of image analysis is image comparison. Surprisingly, no statistical tests are available to compare images with a fixed type I probability error (the error to reject the true null hypothesis). Under the assumption that images are subject to random noise, we want to test if two images have the same grayscale distribution. Clearly, if two images are the same, up to a small noise, they have close grayscale distributions. The reverse is not true. Thus, grayscale distribution analysis is helpful when images of the same content are compared. This test is especially useful to compare microscopic images of tissue or other *unstructured* images frequently used in biology in medicine.

The image histogram is a frequently used technique of image processing [2]. However, besides the histogram one can compute the distribution, or more specifically the cumulative distribution function as the sum of the probabilities that a pixel takes the grayscale level less than g, where $g = 0, ..., 255$. In fact, when statistical analysis is concerned with the distribution of a random variable the empirical cumulative distribution function is usually used, not the histogram as an estimate of the density [8]. One explanation is that while the estimation of the distribution function is unbiased and straightforward the estimation of the density is not, and moreover leads to an ill-posed problem; see [10] for detail.

Let $\{h_g, g = 0, 1, ..., 255\}$ be the histogram of a gray image as the probability that the gray level takes the intensity value $g = 0, 1, ..., 255$. The empirical cumulative gray level distribution function, or shortly *gray distribution* (gd), is defined as follows

$$F_g = \sum_{g'=0}^{g} h_{g'}.$$

A. Laganà et al. (Eds.): ICCSA 2004, LNCS 3046, pp. 933–939, 2004.

Function F_g is a nondecreasing step-function with the step $1/256$ at $g = 0, 1, ...,$ 255, see the right panel of Figure 2 as a typical gray distribution. An advantage of the gd analysis is that it facilitates visual image comparison by plotting grayscale distribution functions on the same scale. Indeed, it is difficult to plot several histograms on the same scale because they often overlay each other–see the left panel of Figure 2 for an example. However, besides better visualization an advantage is application of nonparametric statistical tests, such as Kolmogorov-Smirnov test. Other nonparametric tests for distribution comparison are available, such as Friedman or Wilcoxon test, as described in the reference book [3]. However, the Kolmogorov-Smirnov test is the most popular, perhaps due to simplicity of computations.

Several authors used Kolmogorov-Smirnov test for image segmentation. See [5], [7], [6], to name a few. We apply this nonparametric test to statistical image comparison, or more precisely to test whether images have the same gray distribution.

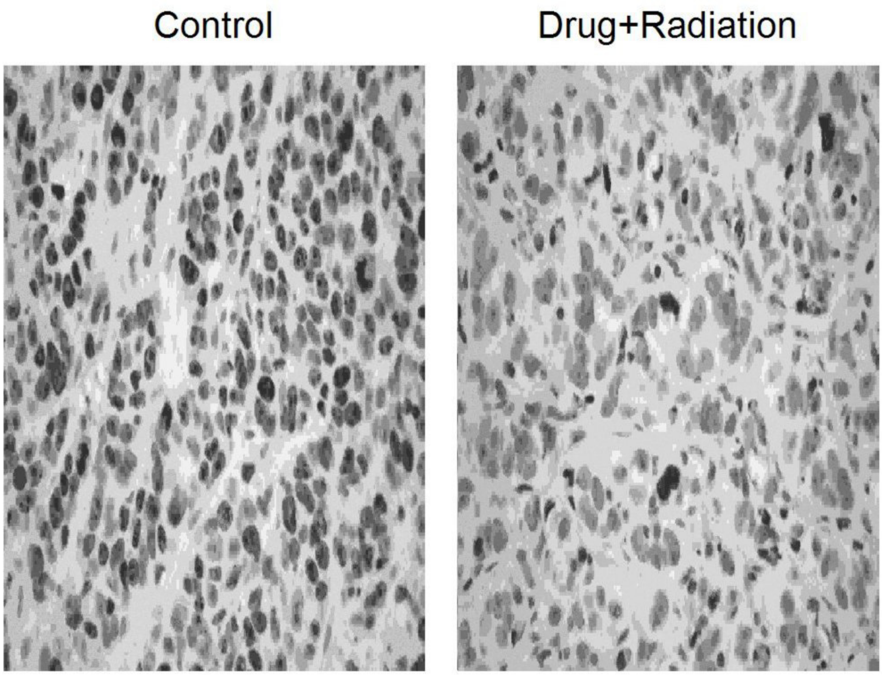

Fig. 1. Histology sections of untreated (control group) and treated tumors. The living cancer cells are the dark spots (blobs). To statistically test that the two images have the same gray distributions the Kolmogorov-Smirnov test is applied.

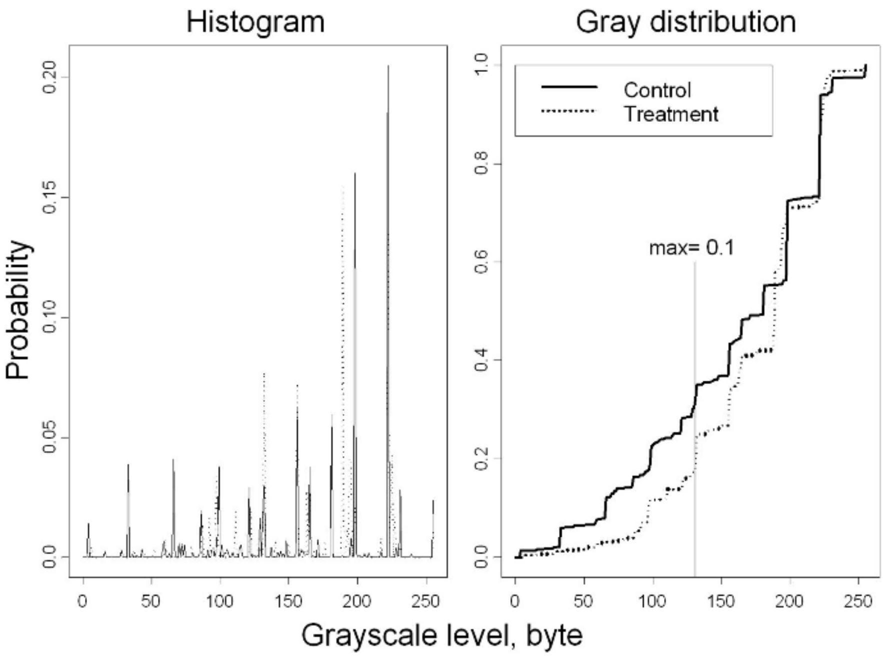

Fig. 2. Histogram and gray distributions for two histology images. The distribution function of the treated tumor is less (almost everywhere) than control (maximum difference is 0.1) which means that the control image is darker. The Kolmogorov-Smirnov distance is used to test the statistical significance.

2 Kolmogorov-Smirnov Test for Image Comparison

Let $F^{(1)} = \{F_g^{(1)}, g = 0, ..., 255\}$ and $F^{(2)} = \{F_g^{(2)}, g = 0, ..., 255\}$ be two gray distributions for $P_1 \times Q_1$ and $P_2 \times Q_2$ images M_1 and M_2. We compute the maximum,

$$\widehat{D} = \max_g \left| F_g^{(1)} - F_g^{(2)} \right|,$$

the distance of one empirical distribution from the other. Kolmogorov [4] and Smirnov [9] proved that the probability that $\widehat{D} > D$, i.e. the observed distance is greater than the threshold, is

$$Q_{KS}(\lambda) = 2 \sum_{j=1}^{\infty} (-1)^{j-1} \exp\left(-2j^2\lambda^2\right),$$

where $\lambda_{KS} = D[\sqrt{J} + 0.11/\sqrt{J} + 0.12]$ and

$$J = \frac{P_1 Q_1 P_2 Q_2}{P_1 Q_1 + P_2 Q_2}.$$

Thus, $Q_{KS}(\lambda)$ with D replaced with $\widehat{D}$ may be treated as the p-value of the test. We notice, the greater the distance between the two distributions the greater the value of $\widehat{\lambda}_{KS}$ and lesser is the probability $Q_{KS}(\widehat{\lambda}_{KS})$. For example, if two images yield distance $\widehat{D}$ and the computed probability $Q_{KS}(\widehat{\lambda}_{KS}) < .05$, we reject the hypothesis that the two images are the same with a 5% error. We can find λ_{KS} such that $Q_{KS}(\lambda_{KS}) = 0.05$ which yields the threshold $\lambda_{KS} = 1.358$.

As a word of caution, all nonparametric tests, including Kolmogorov-Smirnov, have the alternative $H_A : F_1(x) \neq F_2(x)$ for at least one x. Therefore, this test may be conservative.

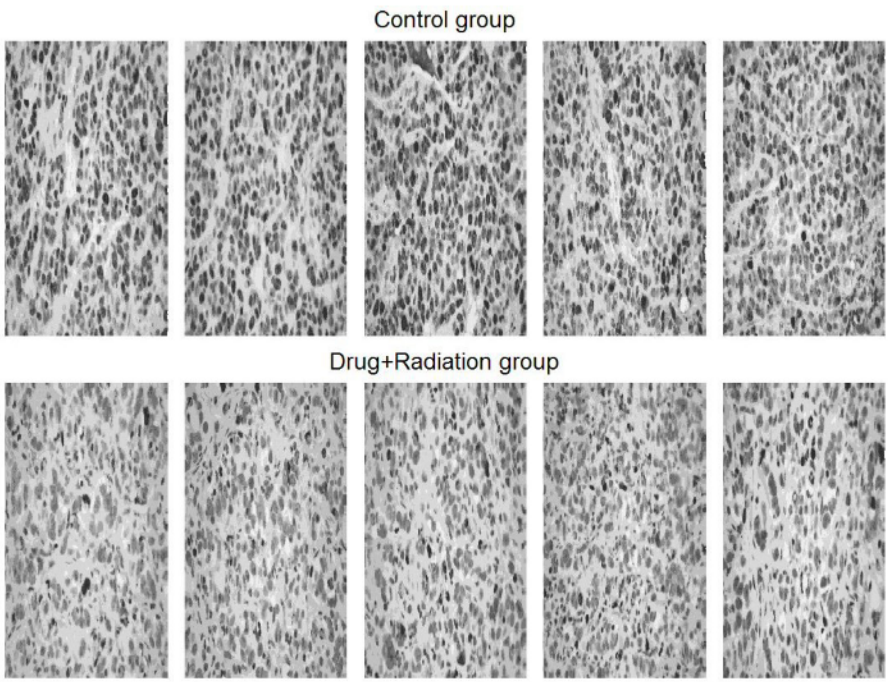

Fig. 3. Two ensembles of histology images taken from 10 mouse tumors in two groups. To determine the cancer kill effect these images are compared by the Kolmogorov-Smirnov test. The darker the image more living cancer cells.

3 Example: Histological Analysis of Cancer Treatment

We illustrate the Kolmogorov-Smirnov test by a histological analysis of breast cancer treatment as described in [11]. Two 2048×1536 images of proliferative activity tumor tissue sections are shown in Figure 1. Dark blobs are cancer cells. In the control tumor (left panel) no treatment was given. In the treated tumor (right panel) a combination of drug, EB 1089 plus radiation, seems to reduce the number of living cancer cells. We want to confirm this reduction statistically using the Kolmogorov-Smirnov test by computing the p-value.

The grayscale histogram and the distribution functions for these images are shown in Figure 2. Clearly, it is difficult to judge the difference in the images by histogram. To the contrary, the distribution functions reveal the difference with the absolute maximum 1/10 at gray level $g = 131$. We notice that the treatment distribution function is below control (for most gray levels) which means that the right image is lighter. For these images $P_1Q_1 = P_2Q_2 = 2048 \times 1536 = 3.1457 \times 10^6$, yielding $\widehat{\lambda}_{KS} = 186.96$ and $Q(\widehat{\lambda}_{KS}) < 0.0001$, near zero. Since the p-value is very small, we infer that the null hypothesis that two images are the same should be rejected. This means that the kill effect of combination of drug and radiation is significant.

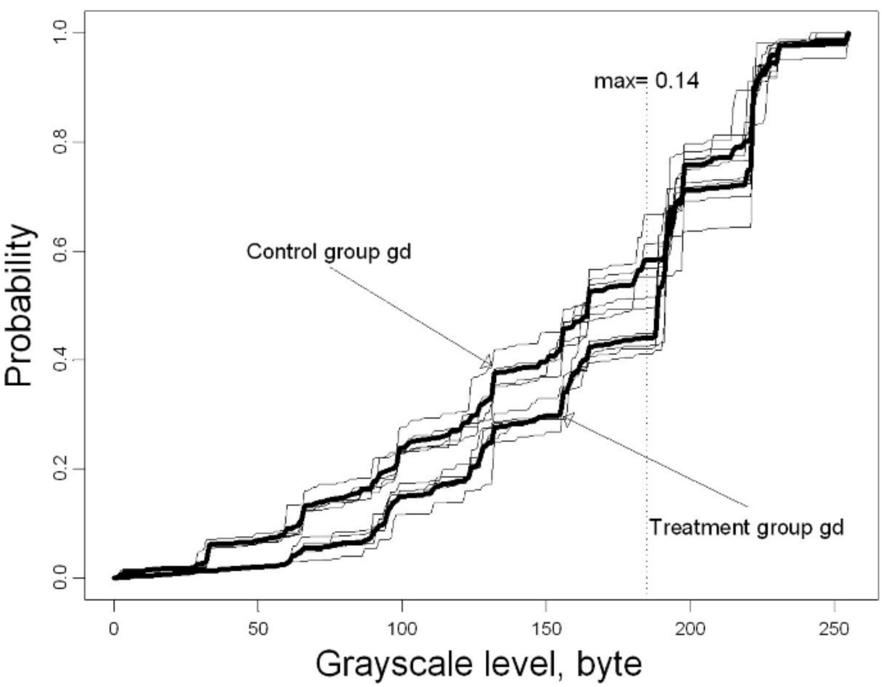

Fig. 4. Gray distributions for 10 histology images in two groups. The bold line shows the groups mean gd. The distance 0.14 between group gds is attained at $g = 185$.

4 Ensemble of Images

In many instances we deal with repeated images or ensemble of images. For example, to determine the kill effect histology images may be taken at different tumor sites, from different animals, etc. In Figure 3, we show histology from 10 different animals, 5 in each group. While human eye may detect difference between two images it becomes difficult to judge several images. Then multivariate statistical testing becomes essential. If two ensembles of images are compared, we may assume that the members within one ensemble have identical distribution. Then each ensemble can be treated as a pooled sample of gray levels. If images are of the same size, it is elementary to show that the group gd is the arithmetic mean of individual gds. Therefore, the group comparison reduces to a comparison of two gray distribution.

In Figure 4 we show 10 individual gds and two group gds (bold). The difference between two gds is 0.14 and is attained at $g = 185$. The p-value of the Kolmogorov-Smirnov test is less than 0.0001. This confirms that the treatment kills statistically significant number of cancer cells.

5 Discussion

We have demonstrated that the Kolmogorov-Smirnov test can be used for image comparison. This test may be used to compare two images or ensemble of images with the null hypothesis that two images (or two samples of images) have the same distribution of 256 gray level intensities. This test is useful to compare content-free or unstructured images when human eye is not able to detect the difference. In particular, the Kolmogorov-Smirnov test can be applied to compare treatment groups in biology and medicine when several dozens or even hundreds of microscopic images, such as histology images, are compared. Showing several gray distributions on one plot is more feasible than histograms and as such it is more convenient for visual comparison.

It is worthwhile to remember that the Kolmogorov-Smirnov test is a two-sided test so we cannot test the hypothesis that one image is darker than another. Also, the Kolmogorov-Smirnov test is a stringent test because the hypothesis is rejected if at least one out of 256 gray level intensities is different.

A limitation of the Kolmogorov-Smirnov test applied to two ensembles of images is the assumption that members from one groups have identical gray distributions and there is no room for site or animal heterogeneity. To account for individual image variation more advanced methods of mixed models should be employed. The interested reader is referred to a recent book [1], where this methodology is applied to image analysis and comparison.

References

1. Demidenko, E. (2004). *Mixed Models: Theory and Applications*. New York: Wiley.
2. Gonzalez, R.C. and Woods, R.E. (2002). *Digital Image Processing*. Second edition. Upper Saddle River, NJ: Prentice Hall.
3. Hollander M. and Wolfe, D.A. (1999). *Nonparametric Statistical Methods*. New York: Wiley.
4. Kolmogorov, A.N. (1941). Confidence limits for an unknown distribution function. *Annals of Mathematical Statistics* **12**, 461-483.

5. Koster, K. and Spann, M. (2000). MIR: An approach to robust clustering - Application to range image segmentation. *IEEE Transactions in Pattern Analysis* **22**: 430-444.
6. Lozano, M.A. and Escolano, F. (2003). Two new scale-adapted texture descriptors for image segmentation. *Lecture Notes in Computer Science* 2905: 137-144.
7. Pauwels, E.J. and Frederix, G. (2000). Image segmentation by nonparametric clustering based on the Kolmogorov-Smirnov distance. *CiteSeer*: http://citeseer.nj.nec.com/pauwels00image.html.
8. Sheskin, D.J. (2004). *Handbook of Parametric and Nonparametric Statistical Procedures*. Boca Raton: Chapman and Hall.
9. Smirnov, N.V. (1948). Table for estimating the goodness of fit of empirical distribution. *Annals of Mathematical Statistics* **19**, 279-281.
10. Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.
11. Sundaram, S., Sea, A., Feldman, S., Strawbridge, R., Hoopes, P.J., Demidenko, E., Binderup, L., and Gewirtz, D.A. (2003). The combination of a potent vitamin D_3 analog, EB 1089, with ionizing radiation reduces tumor growth and induces apoptosis of MCF-7 breast tumor xenografts in nude mice. *Clinical Cancer Research* **9**, 2350-2356.