t-tests, F-tests and Otsu's Methods for Image Thresholding

Jing-Hao Xue* and D. Michael Titterington

***Abstract*—Otsu's binarisation method is one of the most popular image-thresholding methods; Student's t-test is one of the most widely-used statistical tests to compare two groups. This paper aims to stress the equivalence between Otsu's binarisation method and the search for an optimal threshold that provides the largest absolute Student's t-statistic. It is then naturally demonstrated that the extension of Otsu's binarisation method to multi-level thresholding is equivalent to the search for optimal thresholds that provide the largest F-statistic through one-way analysis of variance (ANOVA). Furthermore, general equivalences between some parametric image-thresholding methods and the search for optimal thresholds with the largest likelihood-ratio test statistics is briefly discussed.**

***Index Terms*—Analysis of variance (ANOVA), F-tests, image thresholding, likelihood-ratio tests, Otsu's methods, Student's t-tests.**

EDICS Category: ARS-IIU

I. Introduction

THE purpose of image segmentation is to partition an image into a number of segments, each segment containing some pixels that belong to one of K groups; the number of groups, K, is often predetermined in practice [1], [2].

As the simplest technique for image segmentation, image thresholding usually only needs the information embedded in gray levels of pixels in an image.

For an image with the range of gray levels being $[0, T)$, an image-thresholding method first assumes that the K groups are distinguishable by their gray levels, then determines a vector of $K-1$ thresholds, $\mathbf{t} = (t_1, \ldots, t_{K-1})$ with $t_1 < \ldots < t_{K-1}$, such that K intervals, $[t_0 = 0, t_1), \ldots, [t_{K-1}, t_K = T)$, are constructed, and finally assigns a pixel (denoted by its gray level x) to group k if x lies in the k-th interval (i.e. $x \in [t_{k-1}, t_k)$).

The majority of image-thresholding methods are based on analysis of the gray-level histogram of the image, and were developed for binarisation of the image into two groups (i.e. $K = 2$: one group for the foreground and the other group for the background). Excellent surveys and comparative studies have been reported in [3], [4], [5] and [6], among others.

Among these methods, one of the most popular is Otsu's approach [7]. In the case of image binarisation (i.e. $K = 2$), the basic idea of Otsu's binarisation method to find an optimal threshold t_1^* is as follows.

First, given a candidate threshold t_1, the pixels are divided into two groups. Then a measure of distance between the two groups, defined as a ratio $\lambda(t_1)$ of the between-group variance $\sigma_B^2(t_1)$ to the within-group variance $\sigma_W^2(t_1)$, can be calculated. Finally, the t_1 that provides the highest ratio is selected as t_1^*.

As mentioned in [8], between-group variance and within-group variance are two statistical terms used in analysis of variance (ANOVA). Indeed, the use of the variance ratio of $\sigma_B^2(t_1)$ to $\sigma_W^2(t_1)$ suggests that Otsu's methods should bear some relationship to two-group (or more-commonly-termed two-sample) t-tests or their multi-group counterpart, one-way ANOVA.

In fact, if $K = 2$, we shall show shortly, with simple algebra, that the rule underlying Otsu's binarisation method is equivalent to the search for t_1^* that provides the largest absolute Student's t-statistic, or equivalently the largest F-statistic as defined by the squared t-statistic, from t-tests for two independent normal groups with equal, although unknown, within-group variances. For $K > 2$, we shall show that the rule underlying Otsu's multi-level thresholding method is equivalent to the search for thresholds that provide the largest F-statistic corresponding to one-way ANOVA. The search is over candidate vectors of $K - 1$ thresholds.

Some benefits of stressing such equivalences are as follows.

First, it can enhance understanding of the properties and thresholding performance of Otsu's methods, because t-tests for two groups and F-tests for ANOVA are two of the most established and investigated techniques in statistics. Secondly, it can provide a statistical-hypothesis-testing view of image-thresholding methods and thus facilitate their investigation and development.

II. Methods

In the thresholding procedure for a digital image $\mathcal{X}$ of N pixels (with each pixel represented by its gray level $x_i, i = 1, \ldots, N$), a vector of $K - 1$ gray-level thresholds, $\mathbf{t} = (t_1, \ldots, t_{K-1})$, along with two boundary gray levels, $t_0 = 0$ and $t_K = T$, partitions the image into K groups.

These K groups are denoted by $C_1(\mathbf{t}), \ldots, C_K(\mathbf{t})$ hereafter, such that $C_k(\mathbf{t})$ contains all pixels with gray levels lying in the interval $[t_{k-1}, t_k)$. The value of T is one over the largest possible gray level (i.e. $T = 256$ for an 8-bit gray-level image). As such, $C_K(\mathbf{t})$ represents the background including the brightest pixels, and $C_1(\mathbf{t})$ includes the darkest pixels.

The core of an image-thresholding method is its rule or algorithm for determining an optimal $\mathbf{t}^*$. The majority of existing methods determine $\mathbf{t}^*$ by analysis of the gray-level histogram of the image $\mathcal{X}$.

The histogram of $\mathcal{X}$ can be simply constructed by first counting the frequencies of gray levels and then dividing them by N. The histogram is an empirical probability mass function (PMF), or probability density function (PDF), of the gray level. In the histogram, the proportion for a gray level x is denoted by $h(x)$ hereafter; it follows that $\sum_{x=0}^{T-1} h(x) = 1$, although x is often assumed to be a continuous random variable with Gaussian group-conditional PDF's.

A. Otsu's binarisation method and Student's t-tests

In the case $K = 2$, the threshold vector $\mathbf{t}$ has only one element, t_1; pixels are grouped into $C_1(t_1)$ with a group proportion $\pi_1(t_1) = \sum_{x=0}^{t_1-1} h(x)$, and into $C_2(t_1)$ with a proportion $\pi_2(t_1) = \sum_{x=t_1}^{T-1} h(x)$. The two group means (also called population means) can be estimated by their sample versions, denoted by $\mu_1(t_1)$ and $\mu_2(t_1)$ hereafter respectively. Correspondingly, the group variances are estimated by (biased) sample estimators $\sigma_1^2(t_1)$ and $\sigma_2^2(t_1)$.

J.-H. Xue is with the Department of Statistical Science, University College London, London WC1E 6BT, UK (Tel.: +44-20-7679-1863; Fax: +44-20-3108-3105; e-mail: jinghao@stats.ucl.ac.uk). D.M. Titterington is with the School of Mathematics and Statistics, University of Glasgow, Glasgow G12 8QQ, UK (e-mail: michael.titterington@gla.ac.uk).

This work was partly supported by funding to J.-H.X. from the Internal Visiting Programme, under the EU-funded PASCAL2 Network of Excellence. Thanks to the referees for their constructive and thought-provoking comments.

It follows that the between-group variance and within-group variance can be written as

$$\begin{aligned}\sigma_B^2(t_1) &= \sum_{k=1}^{2}\pi_k(t_1)\{\mu_k(t_1)-\mu_T\}^2 \\ &= \pi_1(t_1)\pi_2(t_1)\{\mu_1(t_1)-\mu_2(t_1)\}^2 , \quad (1)\\ \sigma_W^2(t_1) &= \pi_1(t_1)\sigma_1^2(t_1)+\pi_2(t_1)\sigma_2^2(t_1) , \quad (2)\end{aligned}$$

respectively, where $\mu_T = \sum_{x=0}^{T-1} xh(x)$ is the grand mean of gray levels (also called intensities) of all pixels in the image.

For the selection of an optimal threshold t_1^*, reference [7] suggests optimising either of the following three equivalent measures: $\lambda(t_1) = \sigma_B^2(t_1)/\sigma_W^2(t_1)$, $\kappa(t_1) = \sigma_T^2/\sigma_W^2(t_1)$ and $\eta(t_1) = \sigma_B^2(t_1)/\sigma_T^2$. They are equivalent because the total variance $\sigma_T^2 = \sigma_B^2(t_1) + \sigma_W^2(t_1)$ is a constant with respect to t_1,

The use of the ratio $\lambda(t_1) = \sigma_B^2(t_1)/\sigma_W^2(t_1)$ can be traced back to Fisher's linear discriminant analysis, in which σ_B^2 and σ_W^2 are the variances along a direction $\mathbf{w}$, and λ is maximised with respect to $\mathbf{w}$ for the selection of the normal to the best plane separating the two given groups. Here let us give $\lambda(t_1)$ another interpretation by rewriting it as

$$\begin{aligned}\lambda(t_1) &= \frac{\sigma_B^2(t_1)}{\sigma_W^2(t_1)} = \frac{\pi_1(t_1)\pi_2(t_1)\{\mu_1(t_1)-\mu_2(t_1)\}^2}{\pi_1(t_1)\sigma_1^2(t_1)+\pi_2(t_1)\sigma_2^2(t_1)} \\ &= \frac{\{\mu_1(t_1)-\mu_2(t_1)\}^2}{\frac{\sigma_1^2(t_1)}{\pi_2(t_1)}+\frac{\sigma_2^2(t_1)}{\pi_1(t_1)}} . \quad (3)\end{aligned}$$

It can be recognised that the expression of $(N-2)\lambda(t_1)$ is the classical F-statistic, or equivalently the square of a Student's t-statistic, for two independent normal groups with equal group variances. In addition, the scaling constant, $N-2$, is the degrees of freedom of the Student's t distribution of the t-statistic; it is also the ratio between the two degrees of freedom of the $F_{1,N-2}$ distribution of the F-statistic, if the null hypothesis is that the locations (or more precisely the means) of the two groups are the same.

To see this, rewrite the classical Student's t-statistic as

$$\mathcal{T}(t_1) = \frac{\mu_1(t_1)-\mu_2(t_1)}{s_p(t_1)\sqrt{\frac{1}{N_1(t_1)}+\frac{1}{N_2(t_1)}}} , \quad (4)$$

where, respectively for the two groups determined by t_1, $N_1(t_1)$ and $N_2(t_1) = N - N_1(t_1)$ are the group sizes, and $s_p^2(t_1)$ is the pooled estimator of variance:

$$\begin{aligned}s_p^2(t_1) &= \frac{\{N_1(t_1)-1\}s_1^2(t_1)+\{N_2(t_1)-1\}s_2^2(t_1)}{N-2} \\ &= \frac{N_1(t_1)\sigma_1^2(t_1)+N_2(t_1)\sigma_2^2(t_1)}{N-2} , \quad (5)\end{aligned}$$

in which $s_1^2(t_1)$ and $s_2^2(t_1)$ are the unbiased estimators of the group variances.

It follows that, with simple algebra, if $\mu_1(t_1) > \mu_2(t_1)$,

$$\begin{aligned}\mathcal{T}(t_1) &= \frac{\mu_1(t_1)-\mu_2(t_1)}{\sqrt{\frac{N_1(t_1)\sigma_1^2(t_1)+N_2(t_1)\sigma_2^2(t_1)}{N-2}\left\{\frac{1}{N_1(t_1)}+\frac{1}{N_2(t_1)}\right\}}} \\ &= \frac{\mu_1(t_1)-\mu_2(t_1)}{\sqrt{\frac{1}{(N-2)}}\sqrt{\frac{\sigma_1^2(t_1)}{\pi_2(t_1)}+\frac{\sigma_2^2(t_1)}{\pi_1(t_1)}}} = \sqrt{(N-2)\lambda(t_1)} , \quad (6)\end{aligned}$$

and $\mathcal{T}(t_1) = -\sqrt{(N-2)\lambda(t_1)}$ otherwise.

Equation (6) suggests that, for image binarisation, the optimal t_1^* determined by Otsu's binarisation method is the same as that which can be obtained by searching over t_1 for the value that provides the largest absolute Student's t-statistic, or the value that provides the lowest p-value. In principle, using p-values or absolute Student's t-statistics will give the same optimal t_1^*, because for different t_1 the degrees of freedom, $N-2$, are the same. For image-thresholding practice, however, using p-values is not a good strategy, because Student's t-statistics are often large enough to make all p-values very close to zero.

Equation (6) also suggests that the equivalence between Otsu's binarisation method and the comparison of F-statistics from a set of F-tests also holds, and in fact becomes more clearly in the case of multi-level thresholding, as shown below in section II-B.

The existence of such an equivalence may imply that we can base a measure of the thresholding performance of Otsu's method on established properties of Student's t-test. We discuss some examples briefly as follows.

First, Student's t-test is based on the assumption of two normally distributed groups with equal within-group variances. Therefore, Otsu's binarisation method is expected to work well when that assumption is satisfied. However, the normality of the two group (sample) means is more important. If the data are not far away from being normally distributed, Student's t-test still performs well, because the approximate normality of two group (sample) means can be asserted by the central limit theorem for sufficiently large groups. Therefore, Otsu's binarisation method is expected to be robust for an image in which the within-group gray levels only roughly follow normal distributions.

Secondly, as just mentioned, Student's t-test also assumes that the two groups share a common, although unknown, within-group variance. In spite of this, the test is in general insensitive to the presence of unequal variances across the two groups, when the two groups are of roughly equal sizes [9]. However, when the two group sizes are fairly unequal, Student's t-test is not so robust to the assumption of unequal within-group variances; therefore, Otsu's binarisation method may perform poorly for an image in such cases. We shall use simple numerical examples to demonstrate these in section II-C.

B. Otsu's multi-level thresholding method and F-tests for one-way ANOVA

Reference [7] proposes a straightforward extension of Otsu's binarisation method for multi-level thresholding. In this extension, an optimal vector of $K-1$ thresholds, $\mathbf{t}^*$, is determined by use of the following rule:

$$\mathbf{t}^* = \underset{\mathbf{t}}{\operatorname{argmax}}\ \sigma_B^2(\mathbf{t}) = \underset{\mathbf{t}}{\operatorname{argmax}}\ \sum_{k=1}^{K}\pi_k(\mathbf{t})\{\mu_k(\mathbf{t})-\mu_T\}^2 , \quad (7)$$

where $\pi_k(\mathbf{t})$ and $\mu_k(\mathbf{t})$ are the proportion and the sample mean of the k-th group, respectively, and μ_T is, as before, the grand mean.

Here, as with binarisation, let us look at the rule in equation (7) in the context of F-tests.

Since, with $\sigma_W^2(\mathbf{t}) = \sum_{k=1}^{K}\pi_k(\mathbf{t})\sigma_k^2(\mathbf{t})$, the total variance $\sigma_T^2 = \sigma_B^2(\mathbf{t}) + \sigma_W^2(\mathbf{t})$ is a constant with respect to $\mathbf{t}$, the rule in equation (7) is equivalent to the following rule:

$$\mathbf{t}^* = \underset{\mathbf{t}}{\operatorname{argmax}}\ \sigma_B^2(\mathbf{t})/\sigma_W^2(\mathbf{t}) , \quad (8)$$

where the ratio $\lambda(\mathbf{t})$ is defined as

$$\lambda(\mathbf{t}) = \frac{\sum_{k=1}^{K}\pi_k(\mathbf{t})\{\mu_k(\mathbf{t})-\mu_T\}^2}{\sum_{k=1}^{K}\pi_k(\mathbf{t})\sigma_k^2(\mathbf{t})} = \frac{\sum_{k=1}^{K}N_k(\mathbf{t})\{\mu_k(\mathbf{t})-\mu_T\}^2}{\sum_{k=1}^{K}N_k(\mathbf{t})\sigma_k^2(\mathbf{t})} , \quad (9)$$

in which $N_k(\mathbf{t})$ and $\sigma_k^2(\mathbf{t})$ are the sample size and the sample variance of the k-th group.

Comparing the ratio $\lambda(\mathbf{t})$ with the classical F-statistic, $F(\mathbf{t})$, that is adopted in one-way ANOVA for K groups, we can find that

$$F(\mathbf{t}) = \frac{\frac{1}{K-1}\sum_{k=1}^{K} N_k(\mathbf{t})\{\mu_k(\mathbf{t}) - \mu_T\}^2}{\frac{1}{N-K}\sum_{k=1}^{K} N_k(\mathbf{t})\sigma_k^2(\mathbf{t})} = \frac{N-K}{K-1}\lambda(\mathbf{t}) \ ; \quad (10)$$

that is, the multi-level version of Otsu's method is equivalent to the selection of $\mathbf{t}^*$ by searching for the thresholds $\mathbf{t}$ that provide the largest F-statistic, or equivalently the one that provides the lowest p-value using the $F_{K-1,N-K}$ distribution. This is because the scaling constant, $\frac{N-K}{K-1}$, is the ratio of the two degrees of freedom of the $F_{K-1,N-K}$ distribution, and this ratio has no influence on the selection of $\mathbf{t}^*$. In addition, when $K = 2$, equation (10) becomes equation (6).

Therefore, some advantages and disadvantages of using Otsu's multi-level thresholding method can be implied from the established properties of one-way ANOVA, similarly to those of using Otsu's binarisation method described in section II-A.

C. Numerical validation of properties of Otsu's binarisation method

In this section, we shall numerically validate the properties of Otsu's binarisation method that have been discussed in the end of section II-A based on some established properties of Student's t-statistics. For illustrative purposes, we shall use gray-level histograms constructed from four simulated data sets, which can be viewed as four types of image. As with [10], [11] and [4], we use histograms of simulated Gaussian-mixture data.

Here four simulated data sets, each with 10,000 members corresponding to a virtual image $\mathcal{X}_j$ of 100×100 pixels with $j = 1, \ldots, 4$, were constructed by randomly sampling from four mixtures, respectively. Each mixture is composed of two normal distributions, which are $\mathrm{N}(\mu_1, \sigma_1^2)$ and $\mathrm{N}(\mu_2, \sigma_2^2)$ for the two groups, respectively; the group proportions are π_1 and $\pi_2 = 1 - \pi_1$.

As with [4], we let $T = 256$, $\mu_1 = 100$ and $\mu_2 = 151$. The data are discretised into the range of $[0, T)$. The differences among the four data sets lie in the group proportions and within-group variances, as shown in Table I.

	π_1	π_2	σ_1	σ_2
$\mathcal{X}_1$	0.50	0.50	10	10
$\mathcal{X}_2$	0.50	0.50	15	5
$\mathcal{X}_3$	0.95	0.05	15	5
$\mathcal{X}_4$	0.95	0.05	5	15

TABLE I
PARAMETERS OF TWO-COMPONENT GAUSSIAN MIXTURES FOR FOUR SIMULATED DATA SETS.

The data set for $\mathcal{X}_2$ is similar to that used by [4]. Similarly to [11] and [4], we chose these simply-structured data sets to illustrate the performance of Otsu's binarisation method, here in particular for the following three patterns (as expected in section II-A).

First, for $\mathcal{X}_1$, Otsu's method should perform well in threshold selection, because the assumptions of normality and equal within-group variances are satisfied. Secondly, for $\mathcal{X}_2$, although the within-group variances are unequal, Otsu's method should perform acceptably as the two groups are of the same size. Thirdly, for $\mathcal{X}_3$, Otsu's method is expected to perform poorly because of extremely-unbalanced group sizes and unequal within-group variances.

These characteristics are clearly apparent in the histograms, superimposed with the selected thresholds, in Fig. 1: Otsu's method performs the best for $\mathcal{X}_1$, worse but acceptably for $\mathcal{X}_2$ with a slight bias, and the worst for $\mathcal{X}_3$ with an unacceptable bias towards equal group sizes. This can also be observed from the boxplots (in Fig. 2) for the misclassification error rates obtained from applying Otsu's method to 100 replicates of each simulated data set.

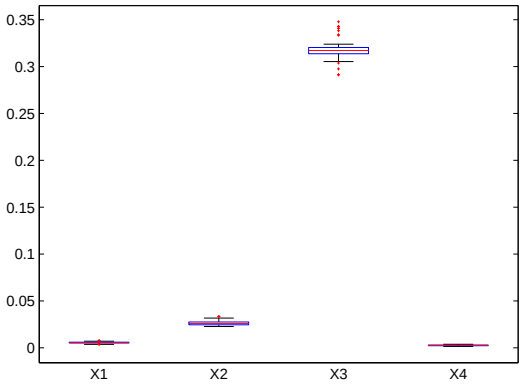

Fig. 2. Boxplots for the misclassification error rates obtained from applying Otsu's method to 100 replicates of each of $\mathcal{X}_1$, $\mathcal{X}_2$, $\mathcal{X}_3$ and $\mathcal{X}_4$.

Usually, as with $\mathcal{X}_3$, in an image with two groups of extremely-unbalanced sizes, the majority group has a larger variance than the minority group. Nevertheless, a case that the minority group has a larger variance may happen, as with $\mathcal{X}_4$. In this case, as shown in Fig. 1, the performance of Otsu's method is in general acceptable; however, we should be cautious if the cost of misclassification of the minority group is much higher than that of the majority group.

The simulation of two-component Gaussian mixtures, Otsu's binarisation method and Student's t-tests are readily implemented in the software MATLAB (The MathWorks, Inc. 2010).

III. DISCUSSION

We now discuss some issues related to the work presented in this paper.

A. A hypothesis-testing view of image thresholding

ANOVA has been employed in many areas of image processing [12], but not yet for histogram-based multi-level thresholding as far as we know. The link between image thresholding and two-group t-tests and one-way ANOVA is stressed in this paper, with the expectation of providing a novel view of image thresholding from the perspective of statistical hypothesis testing, in addition to those from cluster analysis and probabilistic distances.

Otsu's binarisation method is not only an original, simple and elegant approach to image thresholding, but also has a solid foundation in statistics: its rule for selecting t_1^* is based on Fisher's linear discriminant analysis, and involves only means and variances (i.e. up to the second moments of the underlying within-group distributions). As a result, the rule for selecting t_1^* can also be derived from the point of view of normal-distributions-based maximum likelihood, as follows.

For pixels grouped by using t_1, a maximum log-likelihood can be obtained based on the conditional distribution $p(x|y;t_1)$ of x given the group (indexed by y, with $y = 1, 2$ for the two groups), under the assumption that $p(x|y;t_1)$ is a normal distribution with a common variance shared by the two groups. In the end, t_1^* is determined as the t_1 that provides the largest of the maximum log-likelihoods [11]. Meanwhile, as mentioned in [11], the extension of Otsu's binarisation method to multi-level thresholding can also be derived in terms of

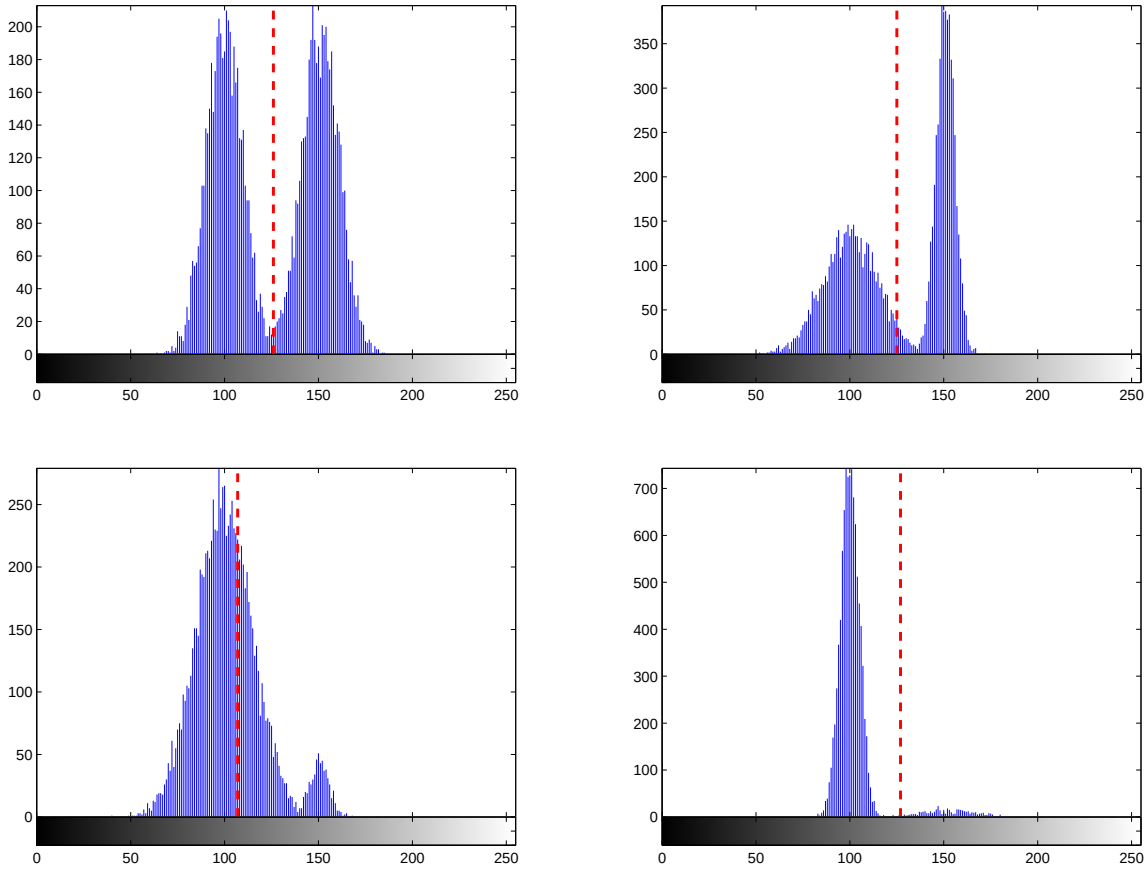

Fig. 1. From left to right and from top to bottom: histograms from simulated data sets, $\mathcal{X}_1$, $\mathcal{X}_2$, $\mathcal{X}_3$ and $\mathcal{X}_4$, with the corresponding thresholds ($t_1^* = 126, 125, 107$ and 127) selected by Otsu's binarisation method and indicated by dashed lines.

selecting $\mathbf{t}^*$ as the $\mathbf{t}$ that provides the largest maximum log-likelihood based on normal $p(x|y;\mathbf{t})$ with equal variances across groups.

If we look at such derivation from a hypothesis-testing point of view, it essentially suggests the equivalence of Otsu's method to the search for the largest test statistic from a set of likelihood-ratio tests. Here the tests in the set for comparison are indexed by the candidate thresholds t_1, and the null model in the tests is the same for different t_1 (e.g. by assuming that two groups follow a common distribution independent of t_1), while the alternative model varies with t_1 (e.g. by assuming that, for each value of t_1, two groups follow two different distributions). Such an equivalence, as well as the derivation originated in [11], is not surprising, because, in our cases of two independent normal groups and one-way ANOVA, the F-test is equivalent to a likelihood-ratio test.

Furthermore, such equivalence can be extended to offering a general (log-)likelihood-ratio view of other parametric image-thresholding methods, such as minimum error thresholding [10], [11] and its variants based on approximating the histogram by a finite mixture of distributions other than normal distributions, such as Poisson distributions [13], generalised Gaussian distributions [14], [15] and certain distributions derived from Rayleigh [16], Nakagami-Gamma, Weibull and log-normal distributions [17].

B. Variants based on variance decomposition/combination

The total variance σ_T^2 can be decomposed into a sum of $\sigma_B^2(t_1)$ and $\sigma_W^2(t_1)$. Otsu's binarisation method maximises, over t_1, the ratio of the between-group variance $\sigma_B^2(t_1)$ and the within-group variance $\sigma_W^2(t_1)$. Since the sum σ_T^2 is invariant to t_1, the rule is equivalent to maximising $\sigma_B^2(t_1)$ only or to minimising $\sigma_W^2(t_1)$ only [7], [11].

There can be other way of combining these two variances to select an optimal threshold t_1^* for image thresholding. Two interesting proposals recently reported are as follows: reference [8] minimises a variant of $\sigma_W^2(t_1)$ by ensuring that the two groups are of the same size when $\sigma_W^2(t_1)$ is calculated; reference [18] minimises an weighted average of the within-group standard deviation $\sigma_W(t_1)$ and the negative distance between two group means, the latter of which can be regarded as a negatively-scaled version of between-group standard deviation $\sigma_B(t_1)$. These proposals may not be as well-founded as Otsu's binarisation methods from a statistical point of view, but this does not necessarily cast a shadow on their encouraging performance in optimal threshold selection for synthetic and real images, as well demonstrated in those papers.

C. Variants based on t-tests

1) Rank-based non-parametric tests: As mentioned in the end of section II-A, when group sizes are sufficiently large (which is often the case in image thresholding), Student's t-test is fairly insensitive to the violation of normality of within-group data. In statistical practice, when the normality assumption is violated for small groups, non-parametric tests such as the Wilcoxon rank-sum test (also termed the Wilcoxon-Mann-Whitney test) are robust alternatives to Student's t-test. However, such rank-based tests are not appropriate for image thresholding, as the two groups determined by the candidate threshold, t_1, are always perfectly separated, whatever the value of t_1 is.

Therefore, these rank-based tests do not offer appropriate variants of, or extensions to, Otsu's binarisation method.

2) Welch's t-tests: When the equal-variance assumption is violated and when the sizes of the two groups are clearly different, inference based on Student's t-statistic may be misleading. In this case, the use of Student's t-statistics (or equivalently Otsu's binarisation method) for threshold selection is hindered; from hypothesis-testing point of view, this explains a characteristic mentioned in [10], [11] and [4] and in section II-C of this paper, among others, namely that Otsu's binarisation method gives a biased threshold, t_1^*, when the variances and sizes of the two groups are distinctly different from each other.

In statistical hypothesis testing, Welch's t-test is a variant of Student's t-test that is used when the two group variances are assumed unequal [19]. Therefore, based on Welch's t-statistics, a variant of Otsu's binarisation method can be proposed as follows.

First, in our case, Welch's t-statistic can be written as

$$\mathcal{T}_W(t_1) = \frac{\mu_1(t_1) - \mu_2(t_1)}{\sqrt{\frac{s_1^2(t_1)}{N_1(t_1)} + \frac{s_2^2(t_1)}{N_2(t_1)}}} = \sqrt{N}\frac{\mu_1(t_1) - \mu_2(t_1)}{\sqrt{\frac{s_1^2(t_1)}{\pi_1(t_1)} + \frac{s_2^2(t_1)}{\pi_2(t_1)}}} \,. \quad (11)$$

Then, it is possible simply to select as t_1^* the t_1 that provides the largest absolute $\mathcal{T}_W(t_1)$, or the smallest p-value.

Such a threshold-selection method based on Welch's t-statistics might be expected to be less sensitive to the presence of two unequal within-group variances than Otsu's original binarisation method, which is based on Student's t-tests. However, this may not be the case in image-thresholding practice, for various reasons including the following.

First, the degrees of freedom for the approximated Student's t distribution of $\mathcal{T}_W(t_1)$ depends on the values of $\sigma_k^2(t_1)$ and $\pi_k(t_1)$, $k = 1, 2$. That is, for different t_1, the degrees of freedom are not the same in our case. Therefore, it may not be reasonable to compare $\mathcal{T}_W(t_1)$ without performing calibration.

Secondly, although we may compare p-values instead of the absolute values of $\mathcal{T}_W(t_1)$, the p-values for different t_1 are often all very close to zero, as with Otsu's binarisation method in section II-A, and the degrees of freedom for the calculation of a p-value are estimated by using, for example, the Welch-Satterthwaite approximation for each t_1 [19], [20].

Thirdly, when t_1 is at either end of the gray-level range, a small group with a small variance is often obtained. This leads to a denominator (with squaring) in equation (11) that is much smaller than that in equation (3), i.e., there might be spikes in the absolute value of $\mathcal{T}_W(t_1)$ at both ends of the gray-level range. In addition, the absolute value of $\mathcal{T}(t_1)$ is not as much unimodal over t_1 as is the case with that of $\mathcal{T}_W(t_1)$.

Therefore, Welch's t-test may not be suggested as an appropriate variant of, or extension to, Otsu's binarisation method to mitigate its sensitivity to the presence of two unequal within-group variances.

D. Unifying image-thresholding methods

Last but not the least, the investigation of the unification of various image-thresholding methods has attracted research effort for a long time, not just shown in those nice surveys mentioned in section I, but also by individual reports such as [11] and [21].

E. Robust statistics for image thresholding

Our intention in this paper was to demonstrate the link between Otsu's method and some statistical tests. Nevertheless, it merits a mention that some estimators developed in robust statistics [22], in particular those for the estimation of variances, can be employed to improve the robustness of Otsu's method. This was highlighted by a referee of this paper and coincidentally explored in one of our recent pieces of work [23].

IV. Conclusions

In this paper, the equivalences of Otsu's binarisation method to the search for an optimal threshold that provides the largest absolute Student's t-statistic, and of Otsu's multi-level thresholding method to the search for optimal thresholds that provide the largest F-statistic from one-way ANOVA, have been stressed. Moreover, general equivalences of some parametric image-thresholding methods to the search for optimal thresholds with the largest likelihood-ratio test statistics have also been discussed.

References

[1] N. R. Pal and S. K. Pal, "A review on image segmentation techniques," *Pattern Recognition*, vol. 26, no. 9, pp. 1277–1294, 1993.

[2] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 3rd ed. Upper Saddle River, New Jersey: Pearson Prentice Hall, 2008.

[3] P. K. Sahoo, S. Soltani, A. K. C. Wong, and Y. C. Chen, "A survey of thresholding techniques," *Computer Vision, Graphics, and Image Processing*, vol. 41, no. 2, pp. 233–260, 1988.

[4] C. A. Glasbey, "An analysis of histogram-based thresholding algorithms," *CVGIP: Graphical Models and Image Processing*, vol. 55, no. 6, pp. 532–537, 1993.

[5] Ø. D. Trier and A. K. Jain, "Goal-directed evaluation of binarization methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 12, pp. 1191–1201, 1995.

[6] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *Journal of Electronic Imaging*, vol. 13, no. 1, pp. 146–165, 2004.

[7] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-9, pp. 62–66, 1979.

[8] Z. Hou, Q. Hu, and W. L. Nowinski, "On minimum variance thresholding," *Pattern Recognition Letters*, vol. 27, no. 14, pp. 1732–1743, 2006.

[9] C. A. Markowski and E. P. Markowski, "Conditions for the effectiveness of a preliminary test of variance," *The American Statistician*, vol. 44, no. 4, pp. 322–326, 1990.

[10] J. Kittler and J. Illingworth, "Minimum error thresholding," *Pattern Recognition*, vol. 19, no. 1, pp. 41–47, 1986.

[11] T. Kurita, N. Otsu, and N. Abdelmalek, "Maximum likelihood thresholding based on population mixture models," *Pattern Recognition*, vol. 25, no. 10, pp. 1231–1240, 1992.

[12] L. Kurz and M. H. Benteftifa, *Analysis of variance in statistical image processing*. Cambridge: Cambridge University Press, 1997.

[13] N. R. Pal and D. Bhandari, "Image thresholding: some new techniques," *Signal Processing*, vol. 33, no. 2, pp. 139–158, 1993.

[14] Y. Bazi, L. Bruzzone, and F. Melgani, "Image thresholding based on the EM algorithm and the generalized Gaussian distribution," *Pattern Recognition*, vol. 40, no. 2, pp. 619–634, 2007.

[15] S.-K. S. Fan, Y. Lin, and C.-C. Wu, "Image thresholding using a novel estimation method in generalized Gaussian distribution mixture modeling," *Neurocomputing*, vol. 72, no. 1-3, pp. 500–512, 2008.

[16] J.-H. Xue, Y. J. Zhang, and X. G. Lin, "Rayleigh-distribution based minimum error thresholding for SAR images," *Journal of Electronics (China)*, vol. 16, no. 4, pp. 336–342, 1999.

[17] G. Moser and S. B. Serpico, "Generalized minimum-error thresholding for unsupervised change detection from SAR amplitude imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 10, pp. 2972–2982, 2006.

[18] Y. Qiao, Q. Hu, G. Qian, S. Luo, and W. L. Nowinski, "Thresholding based on variance and intensity contrast," *Pattern Recognition*, vol. 40, no. 2, pp. 596–608, 2007.

[19] B. L. Welch, "The generalization of 'Student's' problem when several different population variances are involved," *Biometrika*, vol. 34, no. 1-2, pp. 28–35, 1947.

[20] F. E. Satterthwaite, "An approximate distribution of estimates of variance components," *Biometrics Bulletin*, vol. 2, no. 6, pp. 110–114, 1946.

[21] H. Yan, "Unified formulation of a class of image thresholding techniques," *Pattern Recognition*, vol. 29, no. 12, pp. 2025–2032, 1996.

[22] P. J. Rousseeuw and A. M. Leroy, *Robust regression and outlier detection*. New York: Wiley, 1987.

[23] J.-H. Xue and D. M. Titterington, "Median-based image thresholding," manuscript, 2010.