# Toward Objective Selection of Representative Microscope Images

Mia K. Markey,* Michael V. Boland,# and Robert F. Murphy*#
Center for Light Microscope Imaging and Biotechnology, *Department of Biological Sciences, and #Biomedical Engineering Program, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213 USA

ABSTRACT Scientists wishing to communicate the essential characteristics of a pattern (such as an immunofluorescence distribution) currently must make a subjective choice of one or two images to publish. We therefore developed methods for objectively choosing a typical image from a set, with emphasis on images from cell biology. The methods involve calculation of numerical features to describe each image, calculation of similarity between images as a distance in feature space, and ranking of images by distance from the center of the feature distribution. Two types of features were explored, image texture measures and Zernike polynomial moments, and various distance measures were utilized. Criteria for evaluating methods for assigning typicality were proposed and applied to sets of images containing more than one pattern. The results indicate the importance of using distance measures that are insensitive to the presence of outliers. For collections of images of the distributions of a lysosomal protein, a Golgi protein, and nuclear DNA, the images chosen as most typical were in good agreement with the conventional understanding of organelle morphologies. The methods described here have been implemented in a web server (http://murphylab.web.cmu.edu/services/TypIC).

## INTRODUCTION

It is becoming common to collect many digital images via fluorescence microscopy in a single study, yet publications resulting from such studies can usually include only a few images. Although these images are intended to be representative of the results, there is no standard method for selecting a representative image. A reader of a published paper often has no way of knowing the criteria used to select a published image or whether the image was selected at random. The selection may be influenced (either consciously or unconsciously) by an author's conception of what the results ought to be.

With this background in mind, it was of interest to determine whether prior literature describing methods for selecting a representative image from a set of digital images existed. Searches of the INSPEC, MEDLINE, and ArticleFirst databases yielded no articles that appeared to be concerned with the task of selecting a representative image. Combinations of the terms "typical," "representative," "image," "microscopy," and "selection" were used without success (although articles were found for some combinations of these terms, examination of the abstracts revealed that the articles were not relevant).

One way to select a representative image would be to ask which image is most similar to all of the other images in the set. This question is related to the one asked by the designers of image database query engines where the goal is to find those images in the database that are most similar to a particular query image. The Query by Image Content (QBIC) system (Flickner et al., 1995) is one such tool. The major challenges in querying an image database and selecting a representative image are the same: choosing numerical features so that each image is represented as a point in a multivariate space, and choosing a metric for measuring distance between points in that space. The additional challenge faced in representative image selection is the choice of a definition for the most typical observation in a set of multivariate data (also referred to as the multivariate median). Tools, such as QBIC, that measure pairwise image similarity represent a useful basis for the problem of representative selection, but are not a solution in and of themselves.

In this paper, criteria for evaluating methods for selecting a representative image are described, results from applying these methods to sets of images typical of those generated in studies of cell and molecular biology are reported, and methods for measuring image similarity that perform better than QBIC for these types of images are presented.

## MATERIALS AND METHODS

### Images

A collection of images of Chinese hamster ovary (CHO) cells visualized via indirect immunofluorescence was used (Boland et al., 1998). (The image collection is available at http://murphylab.web.cmu.edu/data.) A brief description of the sample preparation and image acquisition follows. CHO cells were grown on coverslips, fixed with 2% paraformaldehyde, and permeabilized with 0.1% saponin. The coverslips were then incubated with a primary antibody, washed, incubated with a Cy5-conjugated secondary antibody, washed again, and mounted on slides. Primary antibodies directed against the Golgi protein giantin, the lysosomal protein LAMP2, and the cytoskeletal protein tubulin were used. DNA was labeled with Hoechst 33258. Images of Cy5 and Hoechst 33258 fluorescence were acquired using a customized Zeiss epifluorescence microscope (Farkas et al., 1993). The microscope system consisted of a 100×, 1.3 NA oil immersion objective, appropriate filter sets, and a cooled CCD camera with a 512 × 382 array of 23-$\mu$m pixels (resulting in a 0.23-$\mu$m pixel spacing at the object plane). After acquisition, the images were manually cropped

*Received for publication 15 July 1998 and in final form 28 December 1998.*
Address reprint requests to Dr. Robert F. Murphy, Center for Light Microscope Imaging and Biotechnology, Department of Biological Sciences, Carnegie Mellon University, 4400 Fifth Ave., Pittsburgh, PA 15213. Tel.: +1-412-268-3480; Fax: +1-412-268-6571; E-mail: murphy@cmu.edu.

0006-3495/99/04/2230/08 $2.00

to isolate single cells, the background fluorescence was subtracted, and the remaining pixels were thresholded at 1.5 times (DNA images) or 4 times (all others) the value of the background fluorescence. Images in the original data set obtained using a monoclonal antibody against NOP4 were not used in this study because there was an insufficient number of images for some of the methods described here.

Twelve test data sets were constructed from all pairwise combinations of the four image classes where 75% of the images were of one image type and 25% were of another type (see Results). The test sets consisted of 77 giantin images combined with 26 images from one of the other classes, 69 DNA images combined with 23 others, 97 LAMP2 images combined with 33 others, or 51 tubulin images combined with 17 others.

## Feature calculation

Two sets of numerical features were used to describe the images, as reported previously (Boland et al., 1998). The first, Haralick's texture features (Haralick, 1979), were calculated using the kharalick function of the cytometry toolbox (Fleming, 1996) for Khoros (version 2.1 Pro; Khoral Research, Albuquerque, NM; http://www.khoral.com). This function calculates the gray-level co-occurrence matrix in four directions (vertical, horizontal, and the two diagonals), and the output is the average of 13 of Haralick's statistics over these four directions. The maximum correlation coefficient was not calculated because of computational instability.

The second set of features was based on Zernike moments (Teague, 1980). Two steps were required to map Cartesian pixel coordinates to a unit circle for calculation of Zernike moments. First, the "center of fluorescence" (center of mass) for each image was calculated and used to redefine the center of the pixel coordinate system. Second, because the Zernike polynomials are defined over a circle of radius 1, the $x$ and $y$ coordinates were divided by 150 (this corresponds to the size of an average cell in the images used here). Only pixels within the unit circle of the resulting normalized image, $f(x,y)$, were used for subsequent calculations. The Zernike moments, $Z_{nl}$, for an image were calculated using

$$Z_{nl} = \frac{n+1}{\pi} \sum_{x} \sum_{y} V_{nl}^{*}(x,y) f(x,y)$$

where $x^2 + y^2 \leq 1$, $0 \leq l \leq n$, $n - l$ is even, and $V_{nl}^{*}(x,y)$ is the complex conjugate of a Zernike polynomial of degree $n$ and angular dependence $l$:

$$V_{nl}(x,y) = \sum_{m=0}^{(n-l)/2} (-1)^{m} \frac{(n-m)!}{m![(n-2m+l)/2]![(n-2m-l)/2]!} (x^2+y^2)^{n/2-m} e^{il\theta}$$

where $0 \leq l \leq n$, $n - l$ is even, $\theta = \tan^{-1}(y/x)$, and $i = (-1)^{1/2}$.

The Zernike moments through degree 12 ($Z_{nl}$ such that $n \leq 12$) were calculated. Because the moments are complex numbers and are sensitive to rotation of the image, the magnitudes of the moments were used as features (i.e., $|Z_{nl}|$) (Khotanzad and Hong, 1990). This provided 49 descriptive features for each image.

## Distance metrics

The Euclidean distance between points in feature space was calculated using

$$\left( \sum_{i=1}^{F} (x_{i1} - x_{i2})^2 \right)^{1/2}$$

where $F$ is the number of features, $x_{i1}$ is the value of feature $i$ for observation (image) 1, and $x_{i2}$ is the value of that same feature for observation 2.

The Mahalanobis distance was calculated using

$$((\mathbf{x} - \bar{\mathbf{x}})' \mathbf{C}^{-1} (\mathbf{x} - \bar{\mathbf{x}}))^{1/2}$$

where $\mathbf{x}$ is a vector of feature values, $\bar{\mathbf{x}}$ is the vector containing the means of the features, and $\mathbf{C}$ is the feature covariance matrix.

Mahalanobis distances were also calculated after robust estimation of population mean and covariance, using multout (version 3.03, available at http://lib.stat.cmu.edu/jasasoftware/rocke). multout generates robust estimates of the mean vector and covariance matrix by detecting and "deleting" outlier events (Rocke and Woodruff, 1996). When using all 49 Zernike features (but not with the 13 Haralick features), multout reported the starting sample covariance matrix as singular (presumably because the number of features was too large for the number of observations available). An iterative procedure was used to reduce the number of Zernike features. The feature with the greatest number of correlations greater than 0.9 or less than −0.9 was identified and removed. In the event of ties, the feature with the largest sum of squared correlations was chosen for deletion. This reduced matrix was used as input to multout, and removal of features continued until the program was able to proceed. This procedure resulted in a reduced feature matrix consisting of 25 Zernike features for giantin, 44 features for LAMP2, 23 for tubulin, and 19 for DNA.

## $\chi^2$ tests

To evaluate the performance of typicality methods on "contaminated" test sets consisting of more than one type of image, the number of minority images assigned a typicality less than or equal to a particular value was tabulated for increments of 0.05 between 0.05 and 1. The comparable ideal cumulative density function (for 25% minority images) was considered to be 0.2, 0.4, 0.6, and 0.8 for typicality values from 0.05 to 0.20 and 1 for all higher typicality values. This was converted to an expected distribution for each data set by multiplying by the number of minority images in that data set. The $\chi^2$ was calculated for each data set to test the hypothesis that the observed and ideal distributions were statistically indistinguishable. The hypothesis was upheld when the $\chi^2$ value for the test data was less than that for 19 degrees of freedom at a 0.99 confidence level ($\chi^2 = 7.633$).

# RESULTS

## Images and features

The choice of the particular numeric features used to describe an image or set of images is crucial to the success of an image analysis task. We have previously demonstrated that either of two feature sets, Haralick texture features (Haralick, 1979) or Zernike polynomial moment features of degree 12 (Khotanzad and Hong, 1990), are adequate to correctly classify nearly 90% of the images in a collection containing five different subcellular localization patterns in Chinese hamster ovary (CHO) cells (Boland et al., 1998). The collection includes images of the subcellular location of giantin (which is found in the Golgi apparatus), LAMP2 (which is found predominantly in lysosomes), tubulin (which forms part of the cytoskeleton), and DNA (which is, of course, primarily found in the nucleus). The results suggest that each of the feature sets contains enough information to summarize each image class in a biologically meaningful manner (i.e., that the features describe characteristics that are important for recognizing the image from

among other, similar images). This knowledge provides an important starting point for choosing representative images based on these features.

The high level of compression achievable using these features should be noted. By using 49 Zernike features instead of $150^2\ \pi$ pixel intensities, we reduce the number of values used to represent the image by a factor greater than 1400. When 13 Haralick features are used, the compression is ~5400:1. While much information is lost in this reduction, it is expected that these feature sets still capture the essential characteristics of the images (as discussed above). For computational reasons (see Materials and Methods), the maximum number of features that could be used to describe 50–100 images was found in all cases to be less than half the number of observations.

## Creation of test sets

In data sets collected via fluorescence microscopy, it is quite common to have some unusual images, both through error in collection and from biological variation. Any method for selecting a representative image must be able to tolerate modest numbers of unusual images, that is, to have the choice of representative image not be significantly affected by the presence or absence of atypical images. To evaluate the extent to which various methods met this criterion, "contaminated" test sets in which 75% of the images were of one class (e.g., giantin) and 25% were of another class (e.g., LAMP2) were constructed from the images in our collection. The choice of 25% contamination was made because it is within the range that can be detected by methods for identifying outliers, given the number of features and observations used here. In addition, this percentage of contamination is expected to be higher than seen in most data sets obtained via fluorescence microscopy.

Sets of images were constructed with each of the four classes (giantin, LAMP2, tubulin, DNA) as majority class and each of the other three classes as the minority class. This approach to contamination was chosen because it was believed that the presence of a second, fairly homogeneous population of images, when added to the main population, would do the most to skew any estimates of image typicality. This is in contrast to combining several images from each of the other classes to constitute the contamination, where one might expect the minority class to be widely and uniformly distributed and potentially have little effect on the assignment of typicality.

It should be noted that biologists do not perceive these image classes as being equally similar to each other. For instance, the pattern of giantin is more like that of LAMP2 than that of tubulin. By separately using each class as a contaminant, the performance of each typicality method could be evaluated with the qualitative perceptions of biologists in mind.

## Criteria for evaluating typicality methods

Two criteria for evaluating the performance of a typicality assignment method on test data sets containing known contaminant images were developed. The first is based on the straightforward concept that for such a contaminated data set, a good typicality method would assign high values to images of the majority class and low values to contaminant images. For a data set containing a fraction of contamination images given by $\epsilon$, an ideal method would always assign all contaminant images typicality values less than $\epsilon$. This is easily evaluated by measuring the cumulative fraction of contaminant images with a typicality less than $t$ as a function of $t$. For an ideal typicality method, this cumulative distribution would increase from 0 at $t = 0$ to 1 at $t = \epsilon$ and then remain at 1 for $\epsilon < t \leq 1$. This is illustrated in Fig. 1. For comparison, the cumulative distribution for a method that is unable to distinguish majority from minority images would increase linearly from 0 at $t = 0$ to 1 at $t = 1$ (Fig. 1). To measure the extent to which the observed distribution for a given typicality method matches the ideal, expected distribution, a $\chi^2$ goodness-of-fit test was used (see Materials and Methods).

The second, simpler criterion is that an ideal typicality method will always pick a member of the majority class as the most typical image (i.e., never assign a rank of "1" to a contaminant image).

## Comparison of approaches using texture features

The contaminated data sets were analyzed using four methods. Each method ranked the images in a set according to

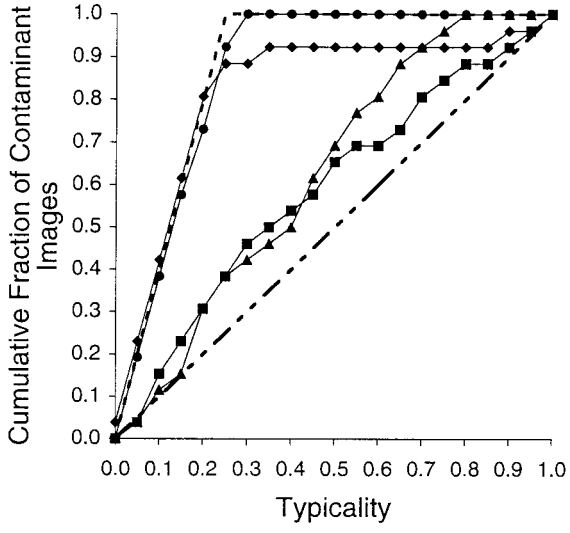


FIGURE 1 The cumulative distribution of contaminant images from each of the methods described in the text for determining image typicality, using the contaminated set consisting of a majority of giantin and a minority of tubulin. ●, HTFR; ◆, QQBE; ▲, HTFM; ■, HTFE. The performance of an ideal system (i.e., one that assigns all contaminant images the lowest typicality possible) is indicated for reference (– – –), as is the performance expected from a system that randomly assigns typicality values to images (– ·· –).

some measure of typicality such that the image with a rank of 1 was most typical. In all cases, the ranks were converted to a "normalized typicality" between 0 (least typical) and 1 (most typical), using

$$\frac{\text{Number of images} - \text{Rank}}{\text{Number of images} - 1}$$

The first method used the version of QBIC available over the Internet from IBM (http://wwwqbic.almaden.ibm.com). QBIC ranks images according to their similarity to a query image, using three texture features based on those of Tamura et al. (1978) and a weighted Euclidean distance measure. QBIC was loaded with each test data set and then queried using each of the images in the set in turn. For each query, QBIC returned a ranked list of the images in order of similarity to that query. Each image was given a score equal to its rank, and the scores were summed for each image across all queries. A final ranking was created from the summed scores and referred to as the QQBE (QBIC query by each) typicality (the image with the lowest sum of scores was defined as the most typical of the set). The two evaluation criteria described above were applied to these typicality assignments. The method passed the $\chi^2$ test for only two of the 12 test sets (Table 1); the cumulative distribution for one of these cases is shown in Fig. 1. Furthermore, the method chose a contaminant image as most typical for one of the sets (majority DNA, minority LAMP2). The inadequate performance of the QQBE method is not surprising, given the low number of features describing each image and the relatively simple distance metric used to define similarity.

In the second method we tested, Haralick's texture features (Haralick, 1979) were calculated to describe each image. A vector of the feature means was calculated, and the Euclidean distance from an image to the mean vector was determined. The images were ranked according to this distance. This was referred to as the HTFE (Haralick texture features, Euclidean distance) typicality. The performance of this method was again less than adequate, not passing the $\chi^2$ test for any of the 12 sets (Table 1), and choosing a representative image from the contaminant class for one set (majority tubulin, minority LAMP2). The example cumulative distribution shown in Fig. 1 shows how poorly this method works for contaminated sets, because the line is much closer to that expected for a random chooser than to that expected for an ideal method. Despite the increase in the number of features describing each image (over QBIC), the Euclidean distance metric was still unable to generate typicalities that met the specified performance criteria.

**TABLE 1 Typicality methods showing desired behavior on test sets containing mixtures of two types of images**

| | Majority image | | | |
|---|---|---|---|---|
| Minority image | Giantin | LAMP2 | Tubulin | DNA |
| Giantin | | | HTFM,HTFR | HTFM,HTFR |
| LAMP2 | | | HTFR | HTFR |
| Tubulin | QQBE,HTFR | QQBE | | |
| DNA | HTFR | | | |

The QQBE, HTFE, HTFM, and HTFR methods described in the text were used to calculate typicality scores for each image in 12 test sets consisting of 75% of one type of image and 25% of another type. The cumulative percentage of minority images with a typicality less than or equal to a given value was tabulated as a function of that value (as in Fig. 1). Entries in the table show methods whose performance, for a given test set, could not be statistically distinguished from ideal performance using the $\chi^2$ criterion (see Materials and Methods).

Using a Euclidean distance measure weights all features equally and therefore does not attempt to compensate for possible differences in the magnitudes of features (by Euclidean distance, two points with values of the first feature of 10 and 9.9 are considered to be the same distance apart as two points with values of the second feature of 0.1 and 0.2). It also does not attempt to compensate for correlations between features (a distance of 0.1 along a diagonal for two highly correlated features is considered equal to the same distance perpendicular to that diagonal, even though the latter is far more significant statistically). Use of the Mahalanobis distance (see Materials and Methods) addresses both of these limitations, using the covariance matrix of the data to adjust for scaling differences and correlations between features.

Therefore, the third method for measuring typicality we evaluated again used Haralick features to describe each image, but used the Mahalanobis distance in place of Euclidean distance. This was referred to as the HTFM (Haralick texture features, Mahalanobis distance) typicality. The change in distance metric resulted in better performance than the HTFE method, meeting the $\chi^2$ criterion for two sets (Table 1) and in no case choosing a contaminant image as most typical.

The problem with the three methods just described is that if the data set contains images that are outliers (extreme values or contaminants), then the selection of representative images may be skewed toward the outliers. This is because calculation of a distance requires knowledge of the population mean vector and, in the case of the Mahalanobis distance, the covariance matrix. The estimation of these population parameters (especially the covariance matrix) can be very sensitive to small numbers of "unusual" observations. Much work has been done to develop methods for estimating population mean and covariance that are robust (Rousseeuw and Leroy, 1987), by which it is meant that the estimates will be accurate for the majority population, even if the sample contains a small number of outliers. The fourth method we tested, then, was identical to the HTFM method, except that a robust estimate of the mean vector and covariance matrix (Rocke and Woodruff, 1996) was used to improve the Mahalanobis distance calculation. The result was referred to as the HTFR (Haralick texture features, robust estimate of Mahalanobis distance) typicality. This method satisfied the $\chi^2$ test for six of the 12 sets (Table 1). The cumulative distribution for one of these cases shown in Fig. 1 can be seen to fall directly along the distribution expected for an ideal method. As for the HTFM method, the HTFR

method did not chose a contaminant image as most representative in any case. When considered along with the results above, this outcome indicates that a more sophisticated estimate of the size and shape of the majority population produces a better (according to the two criteria used here) assignment of image typicality.

It should be noted that some of the minority images may in fact be similar to some of the majority images (e.g., cells with dispersed giantin staining may appear similar to punctate LAMP2 staining). For the CHO images used here, it was observed previously that ~10% of the images in some of the classes cannot be readily distinguished (at least with the features used). Thus the observation that a particular method does not yield ideal results does not imply that that method is not useful, especially if it consistently assigns the highest typicality values to majority images. However, the observation that the HTFR method can perform ideally in cases where the other methods cannot does indicate that this method is preferable.

By creating test sets with the same majority type and different minority types, it was possible to compare the performance of the method on sets subjectively considered to be of unequal difficulty. The HTFR method, for example, did well on mixtures of patterns that are easy for biologists to distinguish. The mixtures that are more difficult for biologists to distinguish (such as giantin and LAMP2) were not ideally resolved by any of the methods.

## Selection of representative images from uncontaminated data sets

Given the performance of the HTFR method, we proceeded to apply it to the uncontaminated data sets. In addition, we explored the impact of the feature set used on the choice of representative image. Because the results above demonstrated the importance of using robust estimation of mean and covariance, this approach was applied to distance estimation with Zernike features in place of the Haralick texture features (this combination is referred to as ZMFR typicality, for Zernike moment features, robust estimation of Mahalanobis distance).

The Zernike features and Haralick features are designed to capture different information about an image, so it was of interest to determine the degree of correlation between the typicality scores obtained with each. Typicality scores were therefore obtained (for a data set consisting only of giantin images) using the HTFR and ZMFR methods separately, and the two values were plotted against each other (Fig. 2). The lack of visible correlation ($R = 0.154$) between Zernike and Haralick typicality scores demonstrates that the two feature sets capture different information. Images that are most typical overall (i.e., images that are most typical by both feature sets) are found in the upper right corner of the plot in Fig. 2. As a means of combining the two methods of calculating typicality, the HZRC (Haralick and Zernike features, robust estimation of Mahalanobis distance, combined) typicality was defined as the square root of the sum of the squares of the HTFR and ZMFR typicalities.

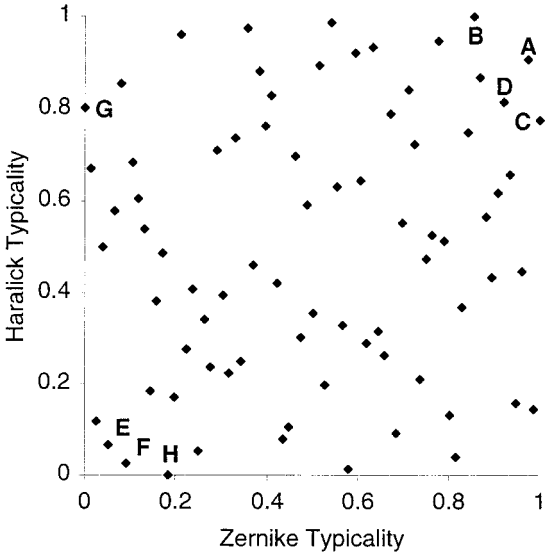


FIGURE 2 Comparison of typicality using the HTFR method and typicality using the ZMFR method for the entire, uncontaminated giantin image set. The letters adjacent to various points reflect the panels in Fig. 3, where the corresponding image is shown.

Because these results were for a data set that did not contain known "contaminant" images, the only method available to assess the utility of the three methods in this case was subjective inspection of images with various typicality scores. To this end, the most and least typical images by each method, along with additional images with high and low HZRC values, are displayed in Fig. 3 (the images are identified by letter in the scatter plot in Fig. 2). The difference between the most and least typical images (with one exception) is clear. Those images that depict vesiculated, dispersed Golgi are found to be least typical, and those that depict compact Golgi are picked as the most typical. This is as expected, because vesiculated Golgi will be found only in those cells preparing for mitosis or in those cells that are not functioning normally. It is important to note that we did not have to introduce biological knowledge or bias (other than that the features used are able to distinguish common biological patterns) into the methods for them to be able to find that a compact Golgi apparatus is more typical than a dispersed one. The exception is the image shown in Fig. 3 *G*, which appears similar in many respects to the images in Fig. 3 *A–D*. Although that image has the lowest ZMFR typicality, it has a high HTFR typicality (giving it a medium composite score). The results indicate that the HTFR typicality may be more reliable for images of structures such as the Golgi apparatus and confirm that the composite (HZRC) method performs well (because the four images with the highest HZRC typicalities all match expectations for a Golgi protein pattern).

The three robust methods were also applied to the other uncontaminated image sets. For DNA images, a moderate

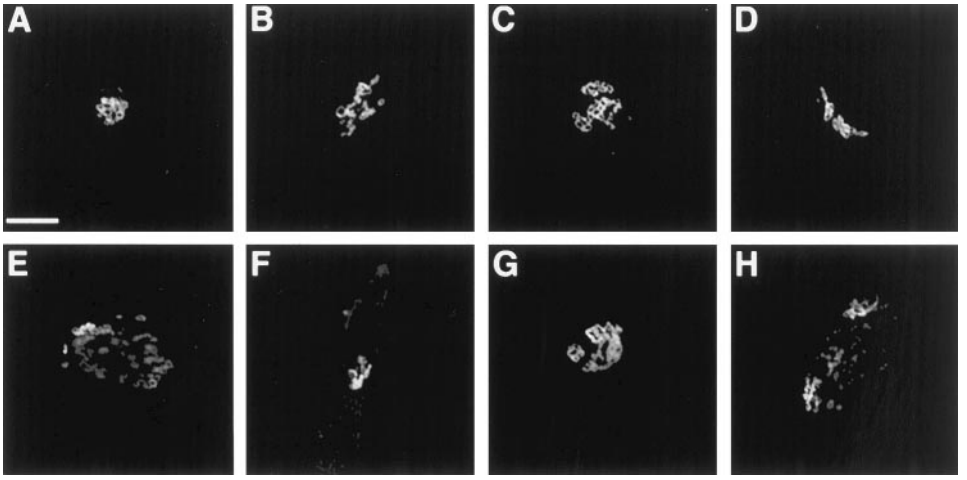


FIGURE 3 The most (*A–D*) and least (*E–H*) typical giantin images, as determined using various methods. (*A*) Highest HZRC. (*B*) Highest HTFR, second highest HZRC. (*C*) Highest ZMFR, third highest HZRC. (*D*) Fourth highest HZRC. (*E*) Lowest HZRC. (*F*) Second lowest HZRC. (*G*) Lowest ZMFR. (*H*) Lowest HTFR, fourth lowest HZRC. Scale bar = 10 $\mu$m.

degree of correlation ($R = 0.41$) was observed between the HTFR and ZMFR typicalities (data not shown). This may be expected, given the simplicity of the nuclear DNA pattern. The images chosen as most typical (Fig. 4, *A* and *B*) show a sharp boundary between nucleus and cytoplasm, whereas those chosen as least typical (Fig. 4, *C* and *D*) show weak, punctate staining extending from the nucleus (perhaps due to poor fixation or cell death before fixation). The agreement with expectation is excellent for all methods.

As with giantin, the correlations between typicality scores by the HTFR and ZMFR methods were low for LAMP2 and tubulin ($R = 0.19$ and 0.11, respectively; data not shown). The most and least typical images chosen by each method are shown in Figs. 5 and 6. The most typical images for LAMP2 (Fig. 5, *A* and *B*) are reasonable choices, showing some lysosomes concentrated in the perinuclear region and some more peripherally located. The image selected as least typical by the HTFR and HZRC methods is clearly abnormal. However, the least typical image by the ZMFR method does not appear grossly abnormal, although the cell represented is larger than usual and the lysosomes appear somewhat larger than normal. With respect to the tubulin images (Fig. 6), it is difficult to conclude that any of them are that much different from any of the others. Because Fig. 6 *B* appears more typical of the cytoskeleton than Fig. 6 *E* (whereas Fig. 6, *A* and *D*, seems equally typical), the HTFR method may be more valuable for tubulin images than the ZMFR method (as was the case for giantin).

## Number of images required for typicality determinations

The method used for robust estimation of means and covariances requires sufficient samples to enable removal of outliers while maintaining a statistically useful number of observations in the main population. The theoretical minimum number of samples (in this case, images) that are needed to reliably detect a certain percentage of outliers at a given confidence level can be determined for a particular number of variables (Rocke and Woodruff, 1996). The numbers obtained by this method for typical confidence are much larger than the number of images in the sets we have used here. It is important to note that because the goal pursued in this paper is to find the most typical images, it is not essential that all outliers be removed. Thus biologically significant results were obtained with only 77 giantin images (Fig. 3). To explore the practical minimum number of images required for finding typical images using the HTFR method, subsets containing specified numbers of randomly chosen giantin images were created. To evaluate how well the method performed for these smaller subsets, the rank of each image in a subset was compared to its rank in the entire set. For five subsets of 40 images each, the overall correlation coefficient relating the two ranks was 0.900, and acceptable results were also obtained for sets of 35 images (correlation coefficient of 0.873). This is near the minimum number necessary for the HTFR method, because the multout program ran indefinitely for subsets of 30 images (and

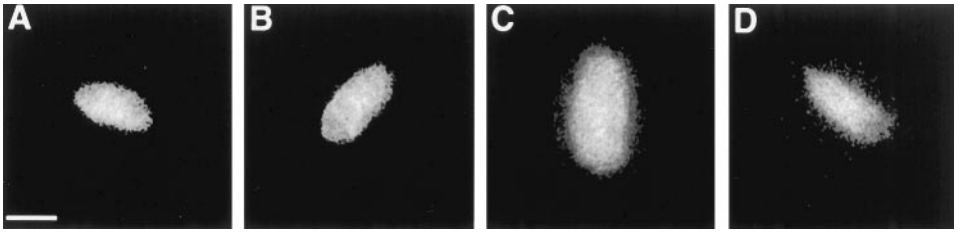


FIGURE 4 The most (*A, B*) and least (*C, D*) typical DNA images, as determined using various methods. (*A*) Highest ZMFR and HZRC. (*B*) Highest HTFR. (*C*) Lowest ZMFR. (*D*) Lowest HTFR and HZRC. Scale bar = 10 $\mu$m.

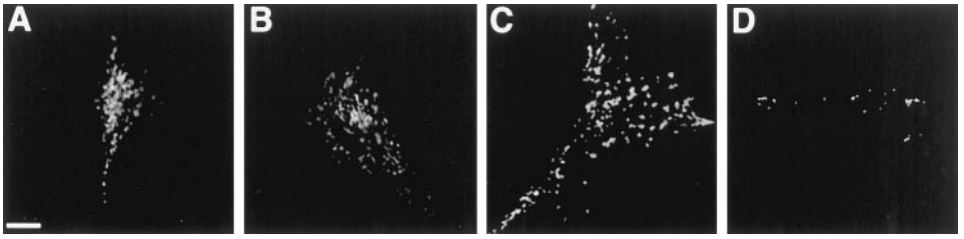


FIGURE 5 The most (*A, B*) and least (*C, D*) typical LAMP2 images, as determined using various methods. (*A*) Highest ZMFR. (*B*) Highest HTFR and HZRC. (*C*) Lowest ZMFR. (*D*) Lowest HTFR and HZRC. Scale bar = 10 $\mu$m.

subsets of 20 images always yielded singular matrices). We conclude that the HTFR method is applicable to image collections of sizes that can be readily collected with digital microscopes. It should be noted that in cases where insufficient images are available for the HTFR method, the HTFM method (which requires at least 14 images) may be used (with awareness of its potential sensitivity to outliers). For image collections smaller than this, the meaning of a representative image may be called into question.

## DISCUSSION

The selection of an image to represent the entirety of the data from which it was derived is currently a subjective, nonrepeatable step in the scientific process. This problem is particularly acute in cell biology and is being compounded by the large numbers of digital images that are now routinely acquired. To overcome this problem, we have developed and tested quantitative methods for selecting representative images from a set. The two steps in this process are 1) extraction of features descriptive of the images in the set and 2) calculation of a distance metric that, in some way, defines how close each image is to the center of the population.

As with any choice of a representative from a set, the selection of a representative image is biased by the numeric values used to describe the images. It is not possible, in fact, to choose a representative from a set without first selecting the criteria by which we will define typicality. As an analogy, consider all criteria that could be used to select a typical person from a population: height, hair color, IQ, etc. Clearly a representative chosen using each of these measurements by themselves need not be typical in all of them, but will be a person of average height, average IQ, or with the most common hair color. Does this fact remove the utility of selecting a representative? The answer is clearly no, but the analogy illustrates the importance of recognizing the criteria used in any system (including the human visual system) when evaluating the representative that is returned. An advantage of automated methods is that the selection bias is explicitly defined by the feature set used. This is in contrast to the subjective determinations made by the human visual system, where the criteria used for selecting a representative are often unknown. Given the vast number of possible feature sets, we do not claim that those tested here are the only ones suitable for fluorescence images; nor do we claim that they are the best. Instead, we chose these features because we believed them to be general-purpose descriptors of the images (i.e., they were not designed with specific protein distributions in mind; nor, for that matter, were they originally intended for use with fluorescence images at all).

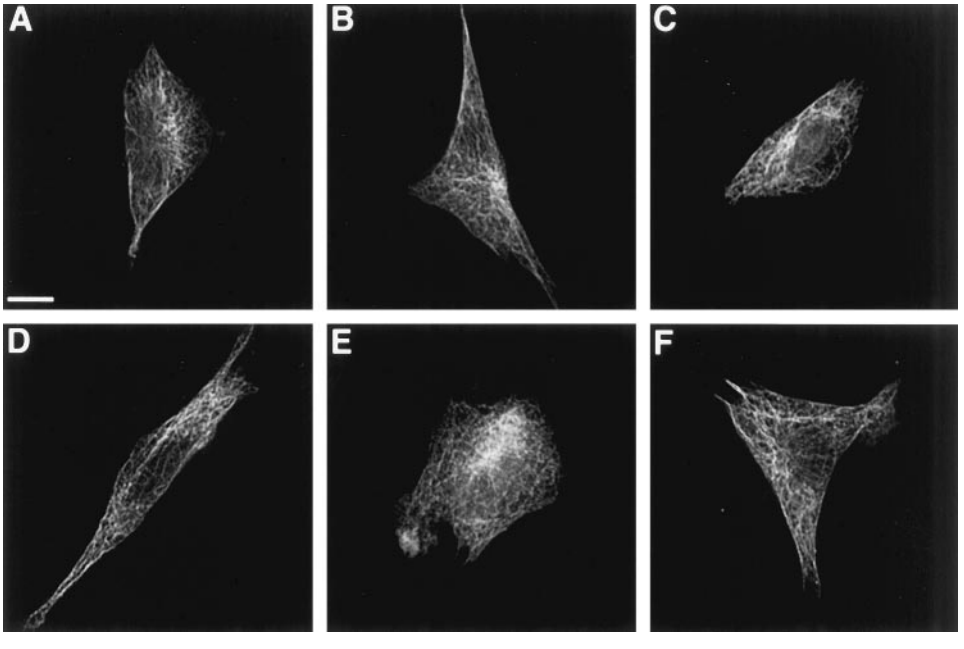


FIGURE 6 The most (*A–C*) and least (*E–F*) typical tubulin images, as determined using the ZMFR (*A, D*), HTFR (*B, E*), and HZRC (*C, F*) methods. Scale bar = 10 $\mu$m.

Based on the results described above, it is clear that the choice of a distance metric is also important for assigning typicality to the members of a set, regardless of the underlying features used to describe them. As demonstrated using the contaminated data sets, a distance metric that takes into account the statistical properties of the features (Mahalanobis distance) is superior to a Euclidean distance with the same features. It is also clear that the detection of outliers and their removal from estimates of population parameters are important steps. Both of these results are as expected.

The approaches we have described for assessing typicality methods using intentionally "contaminated" data sets should have broad application, because the only way to assess the performance of a typicality method applied to a homogeneous (i.e., single class) data set is to subjectively rate the typicality of the representative image. This analysis via human perception was applied to the representative images obtained from the noncontaminated data sets, and it was determined that any of the representative images generated by three methods (HTFR, ZMFR, HZFC) could be considered acceptable from a biological perspective. If all we had were these subjective determinations regarding the methods, we would be back where we started, namely at a point where we are relying on the ambiguities of our own unstated criteria for choosing the images we deem most typical. Instead, we have the statistical and objective results from the contaminated data sets to assure us that the methods employed on the uncontaminated data are the best available, and we are reassured that the representative images look as we expect them to look based on our experience. The important advance is not so much that the system gives us the answer we expected, but rather that it does so in an objective, repeatable manner.

As this area is further investigated and developed, we believe it will have significant impact. First of all, there is the obvious application to areas where it is necessary to represent a set of images with a single member of that set. The first such task that comes to mind is the selection of a single image or of a few images to represent an entire experiment in the scientific literature. This might include research areas as diverse as microscopy, astronomy, spaced-based imaging of the earth, and medical imaging. Another rapidly advancing field is that of image databases. As currently implemented, most of these databases are configured to rank the constituent images in relation to a query image. Selection of representative images, however, might allow databases to be summarized with a handful of images that in some way describe the contents of the database. Furthermore, typicality methods might be able to play a role in data-mining efforts by choosing representatives from among a large collection of poorly understood data. Such goals are beyond the immediate reach of the methods developed here, but it is our belief that this work has defined a novel area of research that has been neglected to this point but that should have broad appeal to investigators from a variety of fields in the future.

To facilitate use and further development of the approaches described here, a web-based typical image chooser (TypIC) (located at http://murphylab.web.cmu.edu/services/TypIC) has been made available. The service will accept a collection of images and rank them according to their typicality.

We thank Christos Faloutsos and Larry Wasserman for helpful discussions.

This work was supported in part by research grant RPG-95-099-03-MGO from the American Cancer Society (RFM), by grants from the Carnegie Mellon Undergraduate Research Initiative and the Howard Hughes Medical Institute Undergraduate Education Program (MKM), by National Science Foundation (NSF) grant BIR-9217091, and by NSF Science and Technology Center grant MCB-8920118. MVB was supported by National Institutes of Health training grant T32 GM08208 and by NSF training grant BIR-9256343.

## REFERENCES

Boland, M. V., M. K. Markey, and R. F. Murphy. 1998. Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images. *Cytometry.* 33:366–375.

Farkas, D. L., G. Baxter, R. L. DeBiasio, A. Gough, M. A. Nederlof, D. Pane, D. R. Patek, K. W. Ryan, and D. L. Taylor. 1993. Multimode light microscopy and the dynamics of molecules, cells, and tissues. *Annu. Rev. Physiol.* 55:785–817.

Fleming, M. G. 1996. Design of a high resolution image cytometer with open software architecture. *Anal. Cell. Pathol.* 10:1–11.

Flickner, M., H. Sawhney, W. Niblack, J. Ashley, H. Qian, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. 1995. Query by image and video content: the QBIC system. *Computer*. 28: 23–32.

Haralick, R. M. 1979. Statistical and structural approaches to texture. *Proc. IEEE.* 67:786–804.

Khotanzad, A., and Y. H. Hong. 1990. Rotation invariant image recognition using features selected via a systematic method. *Pattern Recognit.* 23:1089–1101.

Rocke, D. M., and D. L. Woodruff. 1996. Identification of outliers in multivariate data. *J. Am. Stat. Assoc.* 91:1047–1061.

Rousseeuw, P. J., and A. M. Leroy. 1987. Robust Regression and Outlier Detection. Wiley, New York.

Tamura, H., S. Mori, and T. Yamawaki. 1978. Textural features corresponding to visual perception. *IEEE Trans. Syst. Man Cybernet.* SMC-8:460–473.

Teague, M. R. 1980. Image analysis via the general theory of moments. *J. Opt. Soc. Am.* 70:920–930.