

Image Analysis for Automatic Segmentation of Cytoplasms and Classification of Rac1 Activation

Joakim Lindblad,[1]* Carolina Wählby,[1] Ewert Bengtsson,[1] and Alla Zaltsman[2]

[1]Centre for Image Analysis, Uppsala University, Uppsala, Sweden
[2]Amersham Biosciences, The Maynard Centre, Cardiff, United Kingdom

Received 29 January 2003; Revision Received 2 September 2003; Accepted 30 September 2003

Background: Rac1 is a GTP-binding molecule involved in a wide range of cellular processes. Using digital image analysis, agonist-induced translocation of green fluorescent protein (GFP) Rac1 to the cellular membrane can be estimated quantitatively for individual cells.
Methods: A fully automatic image analysis method for cell segmentation, feature extraction, and classification of cells according to their activation, i.e., GFP-Rac1 translocation and ruffle formation at stimuli, is described. Based on training data produced by visual annotation of four image series, a statistical classifier was created.
Results: The results of the automatic classification were compared with results from visual inspection of the same time sequences. The automatic classification differed from the visual classification at about the same level as visual classifications performed by two different skilled professionals differed from each other. Classification of a second image set, consisting of seven image series with different concentrations of agonist, showed that the classifier could detect an increased proportion of activated cells at increased agonist concentration.
Conclusions: Intracellular activities, such as ruffle formation, can be quantified by fully automatic image analysis, with an accuracy comparable to that achieved by visual inspection. This analysis can be done at a speed of hundreds of cells per second and without the subjectivity introduced by manual judgments. Cytometry Part A 57A: 22–33, 2004. © 2003 Wiley-Liss, Inc.

Key terms: automatic image analysis; cytoplasm segmentation; classification; intracellular translocation; Rac; ruffling

The goal of the present study was to develop an automated image analysis system that could be used to study the behavior of proteins, such as Rac1, that localize to ruffle-like structures at the plasma membrane. Fully automatic image analysis of single cells combined with high-speed, high-resolution imaging over time will allow efficient screening studies, e.g., revealing how drug candidates affect the dynamics of signaling in living cells.

Rac1 is a member of the Ras family of small GTP-binding molecules that play a distinct role in cellular signaling. It is involved in regulation of a wide range of essential cellular processes, including actin reorganization, cell cycle progression, gene transcription, cell adhesion, and migration (1,2). Activation of Rac is associated with its translocation from the cytoplasm to the cellular membrane, resulting in generation of lamellipodia and ruffles, i.e., structural formations and extensions of the cellular membrane that are formed with the reorganization of the actin filaments (3,4). Rac1 is just one example of the many proteins that may localize to ruffles at the plasma membrane.

The use of green fluorescent protein (GFP) as a fluorescent tag and the employment of mammalian cell lines with stably transfected GFP-Rac1 fusion protein allow follow-up and analysis of the activation and translocation of GFP-Rac1 occurring during the cellular response to different agonists. In such events, the ruffle formation appears as bright structures at the membrane edge when imaged in a fluorescence microscope (Fig. 1).

We describe a fully automatic system for image acquisition, image analysis, and quantification of ruffle formation, i.e., formations associated with GFP-Rac1 translocation in individual cells. Images are acquired by using a high-speed confocal system with built-in autofocus. Shading correction and image pre-processing are performed to reduce the effect of imperfections in the imaging step. The individual cells in the images are detected and outlined (segmented). Relevant features describing the biological process are extracted, and the cells are classified according to their level of Rac1 activation.

Contract grant sponsor: Amersham Biosciences, Cardiff, U.K.; Contract grant sponsor: Swedish Foundation for Strategic Research, Visual Information Technology program.
*Correspondence to: Joakim Lindblad, Centre for Image Analysis, Lägerhyddsv. 3, SE-75237 Uppsala, Sweden.
E-mail: joakim@cb.uu.se
Published online in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/cyto.a.10107

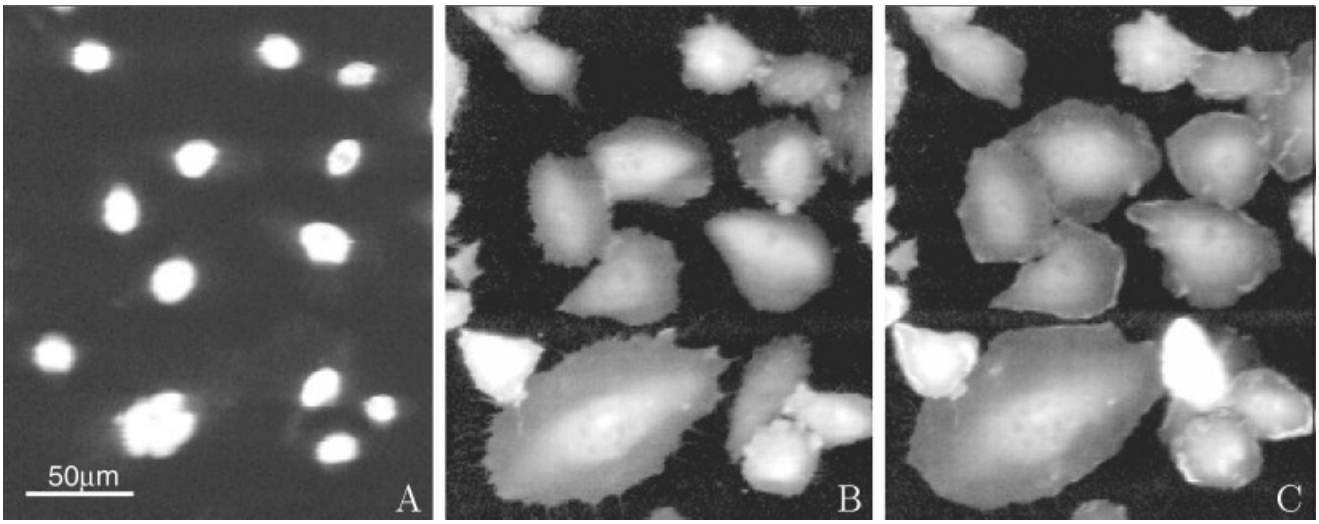

FIG. 1. Chinese hamster ovary hIR cells expressing GFP-Rac1 fusion protein. **A:** Nuclei counterstained with 0.2 µM Hoechst 33258. Cytoplasms (**B**) before and (**C**) 4.3 min after incubation with 100 ng/ml IGF-1. Ruffles appear as bright formations along the edges of the cells. The images of the cytoplasm were preprocessed as described in the text. For improved visibility, images are shown with a logarithmic intensity scale.

Detection of the individual cells in the image requires automatic outlining of each cell's cytoplasm, i.e., cytoplasm segmentation. Cultured cells often touch and differ in size and shape. This makes cytoplasm segmentation a non-trivial task. In the present study, the cytoplasm segmentation is simplified through the presence of a nuclear counterstain. Nuclei do not touch in the same extent as cells, and the fairly round shape of individual cell nuclei makes it easier to separate clusters. Once the nuclei are segmented, they can be used as seeds to enhance correct segmentation of the cells.

We previously described a direct method, independent of nuclear counterstaining, for cytoplasm segmentation based on watershed segmentation and rule-based merging and splitting of over- and under-segmented objects (5). Watershed segmentation (6,7) enhanced by the use of seeds has been described (8), but in this study the seeds are created from the image that is to be segmented and not from a counterstain. Cytoplasm segmentation by detection of the boundaries of living cells, imaged using modulation contrast and differential interference contrast microscopy, has been described (9), where the cell boundaries were found by image pre-processing and thresholding, followed by morphologic operations removing noise and filling holes in the objects. Segmentation of living cells also has been described (10), where the cells were imaged by brightfield optics. All images in this paper were produced by fluorescence microscopy. Segmentation of fluorescence images of cells has been described (11), but in that case, membrane markers were used to show the borders of the cytoplasm, and seeds were drawn manually within each cytoplasm, making the method semiautomatic.

Using the nuclei as seeds for cytoplasm segmentation requires initial segmentation of the nuclei. Several fully automatic methods for segmentation of cell nuclei in two dimensions (2D) and three dimensions (3D) have been described (6,11–14). Cell nuclei can be stained with good contrast and separated from the image background by intensity thresholding. Clustered nuclei thereafter can be separated based on shape. For example, indentations on the contours of the cluster can be paired (12), or the thresholded image can be transformed into a distance image, where the intensity of each foreground pixel corresponds to the distance to the nearest background pixel (15). Watershed segmentation can then be applied directly to the distance image to separate round objects (16), or the maxima in the distance transform can be used as markers for subsequent watershed segmentation of the original image of the cluster (13). Watershed segmentation also can be applied directly to the intensity image of the nuclei, but this often results in over-segmentation due to intensity variations within the nuclei. Over-segmentation however can be reduced by rule-based merging of the fragmented objects (14).

Once each cell has been outlined, several features describing the biological process have to be defined and extracted from the image data. The features are used for evaluating the level of ruffling, i.e., classifying the cells as showing no, moderate, or high activation. Many descriptive features for cell image analysis have been described; see Rodenacker and Bengtsson (17) for an overview. Together with general purpose features, a limited number of problem-specific features also was designed and used to better capture the property of interest, i.e., the ruffle formation at the cell border.

Linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) (18) are well-known and well-tested statistical methods for performing automatic classification. Both methods are derived from Bayes decision theory and are designed to minimize the error rate, i.e., the percentage of incorrectly classified instances in the

data set. Both methods were tested on the described problem.

With a classifier and a set of features, it is tempting to use the complete set of features in the classification, as every included feature should, in the ideal case, contribute to reducing the probability of error. Unfortunately, it has frequently been observed that, beyond a certain point, the inclusion of additional features leads to worse rather than to better performance (18). This can be seen as a type of over-training, where the classifier "learns" the training set well but does not manage to generalize to new samples. If the amount of training data is limited, there exists a "peaking" phenomenon (19), where the performance of the classifier first increases with the inclusion of additional features, attains a maximum, and then begins to decrease. In other words, for a given situation, there exists an optimal set of features that maximizes the performance of the classifier. In addition, reducing the number of features needed for classification leads to an overall speed improvement of the algorithm.

To find the optimal set of features to use, several feature selection methods have been suggested in the literature (20). Feature selection algorithms can be characterized by a search rule, a selection criterion, and a stopping rule (21).

The optimal search rule is to perform an exhaustive search over all possible combinations of the features available. As the number of possible combinations increases exponentially with the number of features, it is not feasible to use this method for more than a very limited number of features. Instead, the standard procedure is to apply some suboptimal search rule to traverse the space of feature combinations. The search rule that we have applied is the so-called sequential floating backward selection procedure (22,23). This search rule has proven to be a good tradeoff between speed and performance (24–26).

The selection criterion decides which features to include in the classifier, and the most natural selection criterion is to include those features that increase the performance of the classifier (27). The direct use of the error rate as a performance measure of the classifier, however, may not be optimal. For example, if 90% of the samples belong to class 1 and 10% belong to class 2, a classifier that classifies everything as belonging to class 1 will be 90% correct. If considering only error rate, this looks like a fairly good classifier. Nevertheless, the classifier is useless. An alternative performance measure, which takes into account the proportions of the different classes in the sample, is Cohen's weighted κ (28). Cohen's weighted κ was used as a performance measure throughout this paper for the feature selection procedure and to evaluate the performance of the overall result.

The stopping rule for the feature selection procedure states how many features to use (29). In this study, when using the classifier performance as a selection criterion, the obvious stopping rule was to pick the set of features leading to the best performing classifier according to our performance measure.

For classifier training and evaluation and for evaluation of the over all methodology, a "true answer," or gold standard, was needed. In other words, data with known classes were needed. For single-cell analysis, the only available gold standard was the classification of cells by visual inspection. However, visual inspection is biased by the observer, and visual inspection by different persons may produce different answers (interobserver variability). If the same visual inspection is performed twice by the same person, there also may be intraobserver variability. In the present study, classification of cells by visual inspection was not trivial, leading to considerable inter- and intraobserver variabilities. Problems with visual classification were not due to poor image quality but to the task itself being difficult. The ruffle formation is a rather subtle property to visually detect and define. The final classification results achieved by the automatic system were compared with the inter- and intraobserver variabilities of the manual classification.

Once the fully automatic system for segmentation of cells and detection and for classification of ruffle formation had been defined and evaluated, it was tested on a completely new set of images. This set of images also showed cells stably transfected with the GFP-Rac1 construct, but a different agonist was used. Also, the concentration of the agonist was varied between different images. Thus, the percentage of activated cells was expected to increase with higher concentration.

MATERIALS AND METHODS

Cell Culture and Preparation

Chinese hamster ovary human insulin receptor (hIR) cells, stably transfected with GFP-Rac1 reporter protein, were seeded in a 96-well plate at $0.6–1 \times 10^4$ cells/well in 200 μl of growth medium. After 24 h of incubation at 37°C, cells were washed with fresh growth medium and serum starved for 2 h. The transfected cells constitutively express the hIR and therefore can be stimulated with insulin or insulin-like growth factor 1 (IGF-1). For the first set of images, IGF-1 was used as the agonist and added at a concentration of 100 ng/ml during the course of the experiment. For the second set of images, insulin was used as the agonist. Six different concentrations of insulin were used: 0, 0.1, 1, 3, 10, and 30 nM. To both image sets, nuclear stain Hoechst 33258, 0.2 μM, was added to each well 30 min before imaging, and cells were incubated at 37°C.

Image Acquisition

Images of the cells were acquired with an IN Cell Analyzer 3000 (Amersham Biosciences, Cardiff, UK). The IN Cell Analyzer 3000 is an automated confocal high-speed fluorescent system allowing imaging and analysis of real-time cellular events. The system employs an automated laser-diode–based autofocus system to identify the bottom of each well before image acquisition by locating the plastic/liquid interface. The autofocus system also uses a tracking feature that monitors and maintains focus

at a user-specified distance from the plastic/liquid interface throughout the scan, checking for the plastic/liquid interface approximately every 30 lines.

Sequential excitation at 488 and 364 nm was used, and green and blue emissions were collected on two CCD cameras. Emission filters used were 535–45 and 450–65 nm. Each image of 1,280 × 1,280 pixels represents a 0.75- × 0.75-mm field of view. The measured intensity was quantized to 4,096 gray levels (12 bit).

In the first experiment, images of four wells, showing a total of 688 cells, were taken every 43 s over a 7-min period, resulting in 10 images/well. Cells were stimulated with IGF-1 after acquisition of the third image. The second data set, consisting of images from 22 wells, acquired just before and 3.5 min after stimulation with insulin, contained 3,966 cells.

Visual Classification

Visual classification was performed on images of cells from four replicate wells, before and 3 min after the addition of IGF-1. Images obtained by the IN Cell Analyzer were annotated in Image-Pro Plus 4.x (Media Cybernetics, Silver Spring, MD) image analysis software, where each cell was manually tagged and classified.

The visual classification was based on comparison of each cell image before and after stimulation. If the appearance of ruffles in a cell after stimulation was noted and differed from the image of the same cell before stimulation, the intensity and size of the ruffles were assessed in relation to intensity of the corresponding cytoplasm and the size of the cell. Cells were classified as belonging to one of five classes:

Class −1: very low or no GFP-Rac1 expression
Class 0: no ruffle formation in response to IGF-1, or cell had ruffles before stimulation and the ruffles did not change in size and/or intensity after stimulation
Class 1: a low to medium response to the addition of IGF-1, with a moderate degree of ruffle appearance or intensity translocation
Class 2: a high response to the addition of IGF-1, i.e., high degree of ruffling; large ruffle size relative to the cell size, and significant intensity translocation to cellular edge as compared with overall cytoplasm intensity
Class 3: the cellular response was difficult to assess visually, e.g., the response of the cell may be obscured, or the cell has had membrane ruffling before stimulation and the extent of its change was difficult to judge, etc.

The time point after stimulation, 3 min, was chosen because it clearly showed ruffle formation but did not necessarily coincide with the peak of the cellular response.

Visual inspections and manual classifications of the first data set were performed by two different skilled individuals, A and B. Person A did the assessment twice, 9 months apart, leading to a total of three classifications, A1, A2, and B. Part of the second data set (1,142 cells from seven wells) was also classified by person A to enable further verification of the performance of the method.

Image Preprocessing

No imaging system is perfect, and uneven illumination and a striped pattern in the scanning direction of the images had to be reduced before image analysis. Dark current and flat field corrections were applied to the images. However, when trying to extract a good global threshold, we observed that additional correction was needed. The striped pattern that appears in some of the images is due to dust on the confocal slit and can be corrected during servicing. Because we wanted to proceed with the analysis of the available images, we chose to apply a data-driven approach for reduction of the striped pattern and other remaining intensity non-uniformities.

For each vertical pixel column in the image, the 20% percentile was calculated. This was considered to represent the background of the image of that column. Each column was then divided with this background level, so that the resulting background intensity was constant for each vertical column in the image. Thereafter we applied a background correction algorithm (30,31). The algorithm iteratively produces a better and better estimate of the intensity variations of the image background by fitting a cubic B-spline surface (32) to the image. One of the nice features of cubic B-splines is that they are continuous and smooth, and their flexibility is controlled by the number of control points used. The distance between the spline surface and the image is minimized by least squares regression. To get a first estimate of the background, the spline surface is initially fitted to the entire image. This first estimate will give a too-bright estimate of the background, because it takes the entire image, including the brighter objects, into consideration at regression. All image pixels that deviate more than a constant number of standard deviation from the background estimate are considered to belong to the foreground and are masked away.

The second iteration starts again with the original image, but this time, the spline surface is fitted only to the pixels that were not already masked away. All image pixels that deviate from the background estimate are found and masked away. This iterative procedure continues until the average change in pixel value between two successively calculated backgrounds is less than half the original quantization step of the image. Convergence is fast and the stop criterion is usually reached after 4–10 iterations. The last foreground/background threshold estimated by the algorithm is used later for the cytoplasm segmentation (see Segmentation of Cells). The images showing the nuclear stain did not need additional correction, because they were not used for measurements sensitive to small background variations. Before any further processing, the images were scaled down (subsampled) by a factor 2 in size to speed up the data processing, because the resolution was found to be high enough not to affect the further processing after scaling. Figure 1 shows a representative region of cells after the preprocessing.

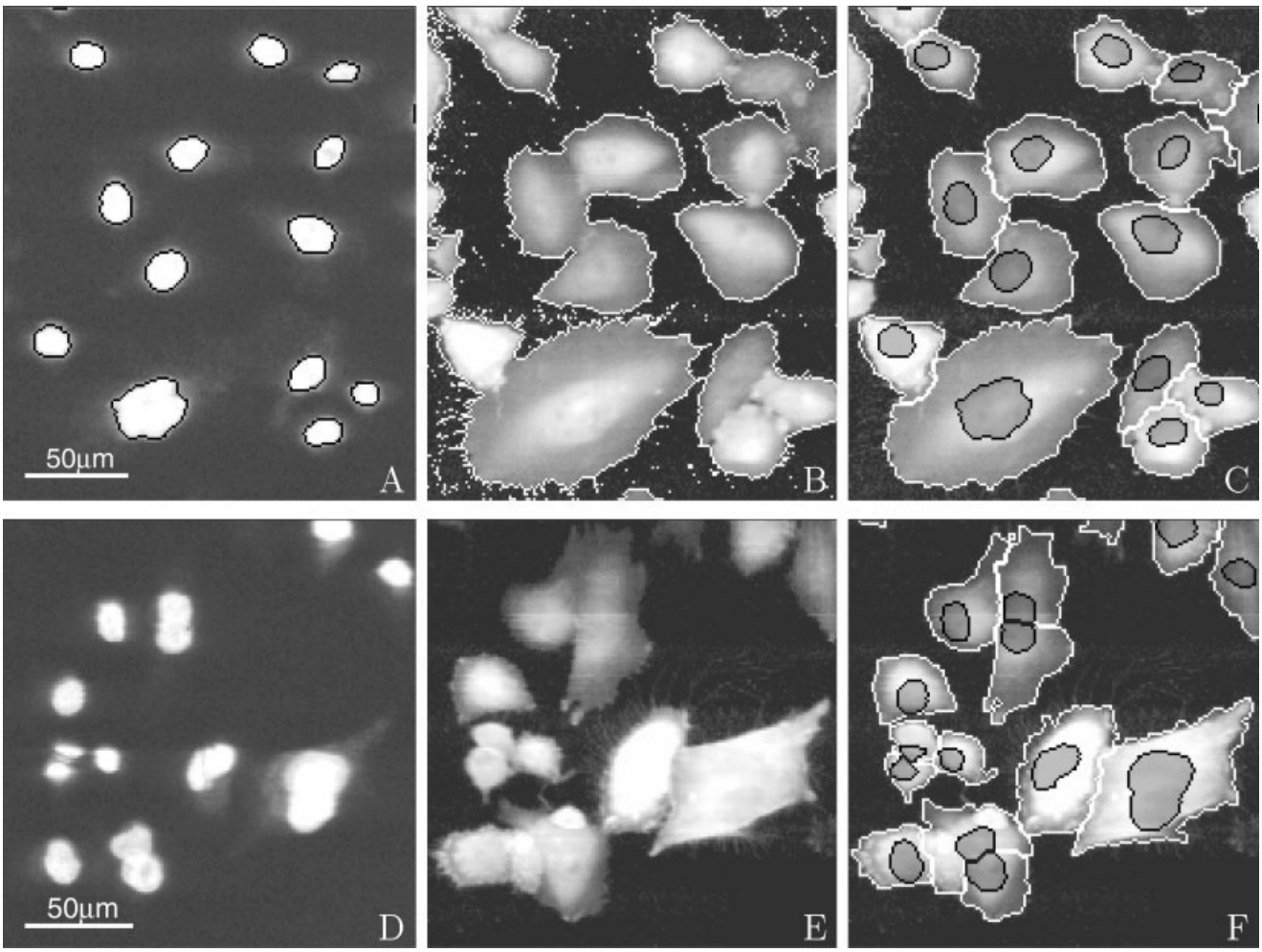

FIG. 2. Segmentation results. **A:** Segmentation of the nuclei is seen as black borders outlining the nuclei. **B:** Thresholding of the cytoplasm image (white borders) does not separate individual cells. **C:** If the nuclei (darker regions with black borders) are used as seeds for watershed segmentation of the cells, all cells are separated (white borders). **D–F:** A difficult region, showing clustered nuclei (D) and overlapping cytoplasms (E). The resulting segmentation is shown in F.

Image Segmentation

Before any information can be extracted from the individual cells in the image, the cells have to be identified by segmenting the image into individual objects and background. Each cell contains one nucleus, and by first segmenting the nuclei and using them as seeds in watershed segmentation of the cells, cytoplasm segmentation is simplified.

Segmentation of nuclei. One image of the cell nuclei was acquired for each time point of the experiment. A single image of the nuclei is sufficient for cytoplasm segmentation at all time points; but instead of just choosing one of the images, a minimum intensity projection (mIP) of all the time point images was created. An mIP consists of the darkest pixel over time for each x–y position in the output image. Bright pixels that were not present at all time points are not present in the mIP; as a consequence, nuclei that are lost during the experiment and added debris will not be present in the mIP image.

The nuclei of the mIP image were separated from the background by an image-specific intensity threshold. The threshold was automatically set so that the contrast between border pixels belonging to object and to background was maximized. After thresholding, clustered nuclei had to be separated. A distance image was created by applying the Euclidean distance transform to the binary objects (15). Touching circular nuclei were thereafter separated by applying watershed segmentation to the distance image (7,16). The result of the nuclear segmentation can be seen in Figure 2A, 2C, and 2F.

Segmentation of cells. The entire cytoplasm of each cell has to be delineated to know what pixels (ruffles, intensities, etc.) are associated with each cell. The cells are separated from the image background by using the foreground/background intensity threshold found by the preprocessing step (see Image Preprocessing). However, touching cells are not separated by simple thresholding (Fig. 2B). Instead, the cells can be separated using water-

shed segmentation (6,7). If the intensity of the image is thought of as height of a landscape, watershed segmentation can be described as submerging the image landscape in water and allowing water to rises from each minimum in the landscape. Thus each minimum will give rise to a catchment basin, and when the water rising from two different catchment basins meet, a watershed, or border, is built into the image landscape. All pixels associated with the same catchment basin are assigned the same label.

The images used in this paper show bright objects on a dark background, and the negative of the image therefore must be used when applying watershed segmentation. As every image minimum gives rise to a catchment basin, applying watershed segmentation directly to the cell images will lead to over-segmentation. However, we know that each cell should contain one nucleus, and using the segmented nuclei as seeds significantly improves the segmentation result.

In seeded watershed segmentation, water can rise only from pixels marked as seeds. New local minima found in the image are flooded just as in standard watershed segmentation, but watersheds are built only between catchment basins associated with different seeds. Thus, water rises from the seeds until water rising from a neighboring seed or the foreground/background threshold is reached. Any object not containing a nucleus and not touching any other object is discarded as debris or noise.

Jagged edges and thin structures in the segmentation result are removed by morphologic opening by using a 3×3 structuring element. The result of the cell segmentation can be seen in Figure 2C and 2F, where the nuclei are outlined in black and the cells are outlined in white. Note that all cells are nicely separated by the algorithm.

Watershed segmentation separates touching cells where the border between them is the darkest (i.e., brightest in the negative image). This is a reasonable assumption at time 0, when the GFP-Rac1 is evenly spread in the cytoplasm. After addition of agonist, the signal is less smooth, and the borders between the cytoplasms tend to be brighter due to the GFP-Rac1 translocation. Applying watershed segmentation to these images therefore may not position the cell borders correctly. Robust analysis of GFP-Rac1 translocation also requires well-defined cell borders after addition of agonist. One approach is to use the segmentation result from time 0 as a mask, but cell motion and shape changes resulted in an unreliable segmentation. The changes in shape and position were compensated for by first thresholding the images in the same manner as for time 0. The segmentation result from time 0 was thereafter allowed to grow within the Voronoi region (33) associated with each cell, i.e., the region that is not closer to any other cell. The Voronoi neighborhoods were constructed by dilating the segmentation result from time 0 six times using a 3×3 structuring element. Dilations were restricted to the foreground pixels found by the thresholding, and the different regions were restricted from overlapping each other. Bright regions located farther away from the original cell border were discarded as debris or noise.

Feature Extraction

Once each cell has been outlined, several features have to be defined and extracted from the image data. The combination of the extracted feature values for a cell is the basis for classifying that cell as showing no, moderate, or high activation and should describe the accumulation of GFP-Rac1 at the cytoplasmic membrane. There exist thousands of different features in the literature, and the choice of features to include in the classification process is not trivial. More features, as previously mentioned, are not always better (34). We therefore restricted the search for good features to those that intuitively felt related to the biological process we wanted to describe. We also created a number of problem-specific features specifically designed to capture the ruffling formation.

Selected general features. The Rac1 activation is essentially a translocation of signal from the cytoplasm as a whole to the cytoplasmic membrane, i.e., the peripheral parts of the cytoplasm. By making a physical interpretation here, thinking of the intensity as material density and the activation as a transport of matter, one concludes that the responding cell should experience an increase in the moment of inertia. The moment of inertia, however, is scale dependent, i.e., if the radius of the cell increases as a whole, the moment of inertia will also increase. A measure that is normalized with respect to scale is *kurtosis*. A distribution with low kurtosis has more intensity near the edges than does a distribution with high kurtosis. This fits well with visual observations of the Rac1 activation. We used the multivariate definition of kurtosis for density functions proposed by Mardia (35).

In addition to using the moment of inertia and the multivariate kurtosis, we included the moment of inertia calculated on a binary image. This provides an intensity-independent reference for the classifier to relate to the gray-scale moment of inertia. There exists a multitude of other moments (36), but we feel that the ones mentioned above are the ones that are most closely related to the property of interest.

We included the total area of the cell, the integrated intensity of the cell, and the perimeter of the cell (37,38). We also included two convex hull features. By using the Quickhull algorithm (39), the area of the 2D convex hull and the volume of the 3D convex hull of the individual cells were calculated. The 3D convex hull was calculated on the volume between the intensity surface of the cytoplasm and the threshold plane. These features allow the classifier to produce convex deficiency-like measures.

Problem-specific features. In addition to the set of general features described above, a set of problem-specific features was defined. The first set of problem-specific features tries to define the regions of increased border intensity of the cell. Imagine that the cell is an island, where the bright inner part is the central massif and the background surrounding the cell is the ocean. If we have

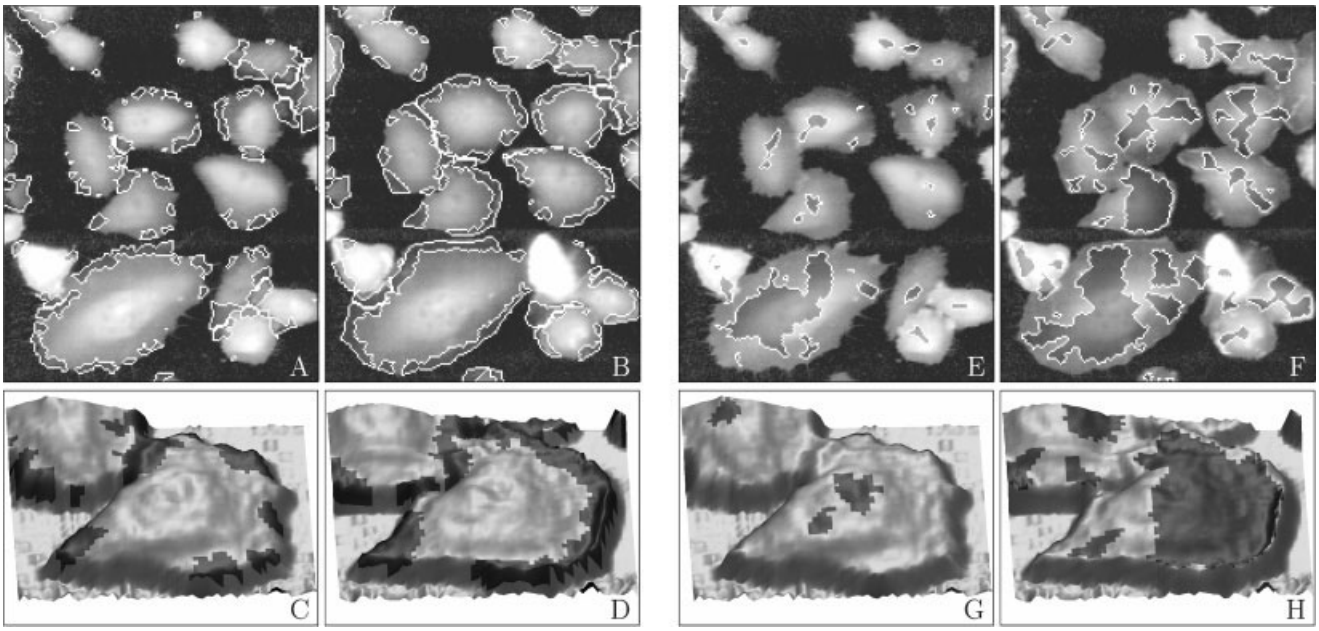

FIG. 3. **A–D:** The ruffle regions (darker shade) before and after stimulation with IGF-1. **C, D:** 3D visualization, where intensity is interpreted as height in a landscape, of ruffle regions before and after stimulation. In the lower right of B, debris has complicated the situation. **E–H:** The internal drainage regions (darker shade) before and after stimulation with IGF-1. G, H: 3D visualization of internal drainage regions before and after stimulation.

a cell showing Rac1 activity, this island will have coastal mountains, i.e., ruffles.

We define the "ruffle regions" as the regions of the cytoplasm that are outside the central massif. The central massif is found by first eroding the cell mask six times with a 3×3 structuring element. However, the erosion is not allowed to go into the nuclear mask. This gives us a region that is well inside the cell and not covering the regions where we are looking for ruffles. The central massif is thereafter outlined by letting the eroded mask grow back out only as long as it is growing downhill in the intensity landscape. The resulting ruffle regions can be seen in Figure 3A–D.

The ruffle regions provide the following set of problem-specific features:

Ruffle area: the total area of all the ruffle regions of the cell

Ruffle pixsum: the integrated intensity of the ruffle regions

Ruffle volume: the total volume between the intensity surface of each individual ruffle region and a horizontal plane positioned at the highest point of the border toward the central massif

The second set of problem-specific features is based on what we call the *drainage regions* of the cell island. Sometimes the cells exhibit bright regions near the cytoplasmic border also at time 0. These regions cause problems for the classifier. However, observations have shown that these bright border regions tend not to be aligned with the cell boundary in the same manner as the ruffles that are associated with the Rac1 translocation.

Once again, imagine that the cell is an island with coastal mountains (ruffles). Behind these coastal mountains, there will be small lakes. If we let it rain over the cell-island, a raindrop falling on the central massif probably will drain down into one of the small lakes, whereas a raindrop falling outside the rim of the coastal mountains will drain into the ocean. Similarly for a cell with no ruffles or only ruffles not aligned with the cell boundary, a raindrop falling on the central massif will drain directly down into the surrounding ocean. We define the "internal drainage regions" of the cell as the area of the cytoplasm that is drained by the internal local minima. The internal drainage regions are found by applying the watershed algorithm to the image, using the background as seed. The resulting internal drainage regions can be seen in Figure 3E–H.

The internal drainage regions provide the following set of problem-specific features:

Internal drainage area: the total area of the internal drainage regions for the cell

Internal drainage pixsum: the integrated intensity of the internal drainage regions

The described features were extracted from all cells at each time point of the experiment. The differences of feature measures before and after stimulation also were used as input for the feature selection described below. Data from two, three, and four time points from before and after stimulation were tested.

Creation of a Classifier

A classifier that, based on the extracted feature values, can classify each cell as belonging to one of the first four

classes described under Visual Classification had to be created. We previously tested two types of statistical classifiers, LDA and QDA (18). Both classifiers use a set of discriminant functions, $g_i(\boldsymbol{x})$, $i = 1,2,\ldots c$, and assign a feature vector, $\boldsymbol{x}$, to class ω_i if $g_i(\boldsymbol{x}) > g_j(\boldsymbol{x})$ for all $i \neq j$. General Bayes minimum error rate classification can be achieved by the use of the discriminant functions

$g_i(\boldsymbol{x}) = \log p(\boldsymbol{x}|\omega_i) + \log P(\omega_i)$, where $P(\omega_i)$ is the a priori probability of class i. By evaluating this expression for a d-dimensional multivariate normal distribution, we arrive at the following discriminant functions:

$$g_i(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)^t \sum_i^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i) - \frac{d}{2}\log 2\pi - \frac{1}{2}\log\left|\sum_i\right| + \log P(\omega_i)\,. \quad (1)$$

The LDA classifier assumes that the covariance matrices for all classes are identical, $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$. By using equal a priori probabilities for the different classes and substituting the mean and covariance matrix with the sample mean $\boldsymbol{m}_i$ and sample covariance matrix $\boldsymbol{S}$, the minimum error rate classifier discriminant functions reduce to:

$$g_i(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{m}_i)^t \boldsymbol{S}^{-1}(\boldsymbol{x} - \boldsymbol{m}_i) + \log P(\omega_i)\,, \quad (2)$$

which gives linear decision boundaries, hence the name. Assuming equal a priori probabilities, the LDA classifier described by equation (2) can be described as a minimum Mahalanobis distance classifier, thus classifying the sample $\boldsymbol{x}$ as belonging to the class ω_i with the shortest Mahalanobis distance to $\boldsymbol{x}$.

In the general multivariate case, one can not assume the covariance matrices of the different classes to be identical. This leads to the QDA classifier, which has the following discriminant functions:

$$g_i(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{m}_i)^t \boldsymbol{S}_i^{-1}(\boldsymbol{x} - \boldsymbol{m}_i) - \frac{1}{2}\log|\boldsymbol{S}_i| + \log P(\omega_i)\,, \quad (3)$$

giving hyper-quadric decision boundaries.

The estimates of the class mean values, covariance matrices, and a priori probabilities have to be derived from a "training set" of known data, or gold standard. As described below, the classifier had to be tested on a known "test set" and finally evaluated on a known "evaluation set." In other words, data with known classes were needed. For single-cell analysis, the only available gold standard was the classification of cells by visual inspection. Visual classifications, as described under Visual Classification, were performed by two different individuals, A and B, where person A did the assessment twice, leading to the three classifications, A1, A2, and B. All three visual classifications were used together as a gold standard. Visual classification was not trivial, and A1, A2, and B did not always agree. By including all three classifications into the training set, a stronger weight is given to the clear cases in which A1, A2, and B agree.

Feature selection. For good classifier performance, features relevant for the classification have to be selected. Selecting a limited number of features reduces the tendency for over-training and improves the classifier performance on new, unseen, data sets (18,19). Feature selection algorithms can be characterized by a search rule, a selection criterion, and a stopping rule (21).

The search rule that we have applied is the sequential floating backward-selection procedure (22,23). Because the initial feature set was fairly small, the backward version was used, and the problem of stopping too early was avoided (25). First, all features described under Feature Extraction were extracted from the training set and included in the classifier. Second, one feature at a time was temporarily removed, and the performance of the classifier was tested on a test set. Third, the feature that contributed the least to the classification performance was removed. This was done over and over again until only one feature was left. Fourth, to be sure not to accidentally remove the best feature from the beginning, before removing another feature, it was always checked if the inclusion of one of the previously removed features gave an improved performance. In other words, features were removed and put back, but when a feature was put back, we always made sure that the performance improved. In this way, the process did not go on forever, and only one feature was left.

The selection criterion tells us which features to include or not include in the classifier. As mentioned above, we use increase in classifier performance as a selection criterion (27). Alternative selection criteria include the use of F-statistics or Wilks' Λ (which are not optimal for the task) (27,40) or the more advanced Bhattacharyya distance measure (41). The main motivation to apply any of these alternative criteria is preventing the direct measure of the classification performance from being too time consuming. However, this was not the case here.

The performance measure that we have used throughout this paper for the feature selection procedure and to evaluate the performance of the overall result is Cohen's weighted Kappa (κ_W) (28). Cohen's weighted Kappa provides a measure of the degree to which two classifiers agree in their respective sortings of N samples into k classes. This gives us a useful performance measure during feature selection and training of the classifier against our gold standard and when comparing the classifications made by two different individuals (42).

Assume that N samples are distributed into k classes by one classifier and, independently, into the same k classes by a second classifier, resulting in a matrix with k^2 cells. Let p_{ij} be the proportion of objects placed in the i,jth cell. Let p_i be the proportion of objects in the ith row and let p_j be the proportion of objects in the jth column. Then κ_W is defined as

$$\kappa_W = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}\,p_{ij} - \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij} p_i \cdot p_j}{1 - \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}\,p_i \cdot p_j} \quad (4)$$

The weights w_{ij} of the diagonal (correct) classifications were set to 1, and the off-by-one misclassifications of assigning a non-activated cell as a moderately activated cell or a moderately activated cell as a highly activated cell, or vice versa, were set to 0.5. All other weights where set to 0.

As a stopping rule, we picked the set of features leading to the best performing classifier. During the feature selection we keep track of the features included in the best performing classifier for every possible number of features. We can then backtrack this list and pick the feature set resulting in the best classifier, i.e., giving the highest κ value on the test set.

Training and evaluation. For training and classifier evaluation, four images of cells that had been visually classified were available. The limited amount of training data was due mainly to the tedious visual classification. The visually classified cells had to be used as training and test sets at feature selection and for the final evaluation of classifier performance. Because evaluation of the result in an unbiased way is most important (34,43), great care was taken to never use the same data for both training and evaluation.

A four-fold cross-validation scheme (44) was applied to maximize the use of the data. Three images at a time were used for feature selection and to create a classifier, where the fourth image was saved for evaluation. This was repeated four times, so that each of the four images of the first data set was classified without being involved in the training. The resulting classifications were then compared with the gold standard using Cohen's weighted κ.

The feature selection procedure also used a cross-validation scheme internally to be able to better estimate the true performance of the classifier when selecting features. Of the three images used for feature selection, training of the classifier was done on two of them, and the third image was used for testing the performance of the classifier. In each step of the feature selection procedure, all combinations of the three images were used before the decision of adding or removing a feature was taken. When the feature selection procedure had finished and a reduced set of features had been selected, all three images could be used to train the classifier when evaluating the performance on the fourth "untouched" image.

Complete mixing of the data, such as the leave-one-out method, was avoided, because it most probably would introduce a bias due to cells from the same image being correlated with each other.

The classification results were compared with the inter- and intraobserver variabilities of the visual classifications. The created classifier was then tested on a second data set consisting of images of cells created under the same conditions as those used for the first data set, except that a different agonist (insulin) was used. The concentration of agonist was changed within this data set, and because ruffling activity is expected to be related to concentration of agonist (in a limited range), one could expect to see an increase in cell activation at an increased concentration of agonist. In other words, this data set contained a priori information, independent of visual classifications. In addition to the a priori information based on agonist concentration, some of the images in the second data set had been visually inspected and cells classified.

Implementation

All the algorithms described above were implemented in MATLAB (The MathWorks, Natick, MA). Because the ultimate goal of this study was to use the described algorithms as a fully automatic tool for real-time image analysis, as a part of the IN Cell Analyzer 3000, care was taken to use fast algorithms and to avoid those that are intrinsically slow, e.g., active contour models or diffusion models. To verify that the goal of analyzing roughly 200 cells in less than 2 s was not unrealistic, the more time critical parts (seeded watershed segmentation, etc.) of the algorithms were also been implemented in C++. Preliminary studies showed that, on a modern PC, this goal is most realistic.

RESULTS

LDA and QDA were tested on the described problem. The more flexible QDA was the one that best managed to capture the variations of the cells. It was interesting to see that the feature selection procedure selected a larger number of features when using the less flexible LDA, possibly to compensate for lack of descriptive power.

The feature selection and the overall performance indicated that using the feature measures directly provides better results than supplying the classifier with only "after minus before" differences.

Using three images (before and 3 and 4.5 min after addition of agonist) produced good results, and adding more images from other time points did not significantly improve the classification. For the second data set, only two time points (before and after) were available, so when using the first data set to train a classifier for the second data set, only two images could be used. This many partly explain the slightly worse result for the second data set.

When trained on the first data set, the feature selection picked out the following 14 features (counting a feature extracted before and after addition of the agonist as two different features):

Moment of inertia: before and after addition of agonist
Kurtosis: before addition of agonist
Cell area: before and after addition of agonist
Cell pixsum: before addition of agonist
Cell perimeter: before addition of agonist
Ruffle area: before and after addition of agonist
Ruffle volume: before and after addition of agonist
Internal drainage area: after addition of agonist
Internal drainage pixsum: before and after addition of agonist

Table 1
Comparing Automatic and Visual Classifiers

	All classes		Classes 1 + 2 grouped	
Classifier	% Correct	Cohen's κ	% Correct	Cohen's κ
C vs. W	46.2	0.241	74.0	0.313
C vs. A1	43.8	0.251	68.3	0.304
C vs. A2	44.2	0.209	72.2	0.249
C vs. B	47.0	0.237	76.4	0.295
A1 vs. B	44.9	0.243	80.4	0.510
A1 vs. A2	70.1	0.485	82.3	0.592
C vs. most similar	65.9	0.494	79.5	0.475
C vs. A for second data set	33.8	0.163	60.3	0.279

The results of the automatic classifications using the described method were compared with three visual classifications (see Visual Classification) and with a weighted summary of all the visual classifications, summarized in Table 1. The computer classification is denoted classifier C, and the weighted summary of A1, A2, and B is denoted W. Comparing A1 (or A2) with B gives an estimate of the interobserver variability, and comparing A1 with A2 gives an estimate of the intraobserver variability.

To better capture uncertainty in the visual classification, the computer classifications were also compared with the most similar of the three visual classifications, "C versus most similar" in Table 1. For example, if the three visual classifications of a cell are 1, 1, and 2, the weighted summary W will say class 1. If the computer classification is 2, this would normally be considered an error; but if the computer classification is compared with the most similar visual result, here class 2, it will not be an error. This can also be seen in Figure 4, where the visual and computer-calculated classes are plotted on top of each cell. In Figure 4, 70% of the cells are correctly classified when compared with W, but 100% are correct when compared with the most similar of the visual classes.

The results presented in Table 1 were achieved by using the classifier created from the first data set to classify the second data set. Only part of this data set was visually classified, and the visual classification was performed by one person (A) at one time. Therefore, no information on inter- or intraobserver variability was available. However, one can assume a precision similar to that of the first data set.

For the second data set, an increased proportion of activated cells (classes 1 and 2) as compared with inactive cells (class 0) was, a priori, expected for higher insulin concentrations. Figure 5 shows the proportion of the different classes at increasing concentrations of agonist. The classifier clearly detects a larger proportion of activated cells (two shades of light gray) at higher insulin concentrations.

DISCUSSION

The goal of this study was to develop an automated image analysis system that could be used to study the behavior of proteins, such as Rac1, that localize to ruffle-like structures at the plasma membrane.

Two different data sets were created by stably transfecting cells with GFP-Rac1. The cells were imaged over time, before and after addition of a Rac1 activating agonist. The first data set was used for training, testing, and evaluating a classifier. The only gold standard available on which to base the training, testing, and evaluation was visual classification of the cells. This is commonly the case in single-cell–based image cytometry, because few other methods can evaluate time-dependent events in single cells. Despite high image quality, visual classification of Rac-1 activation was not trivial. Two professionals classified the cells, and one of them performed a second visual classification of the same cells 9 months after the first classification.

The difference between Rac1 classification made by the fully automatic image analysis procedure described in this paper and the gold standard was roughly as large as the difference between two visual classifications of the same material. This indicates that the image analysis procedure is almost as reliable as the manual classification. It should be noted that visual classification of the complex Rac-1 activation is very subjective, and the distinction between "moderate" and "high" responding cells was not easy to make by eye. Visual classification is also very time consuming as compared with fully automatic classification, making the automatic system attractive for analysis of large data sets. Because the difficulty in classifying the cells does not lie in the delineation of the cells (which is performed in a satisfactory way by the automatic system) but rather in the quantification of the ruffling, methods for semiautomatic classification based on user interaction are difficult to define.

It is difficult to improve the performance of the automatic procedure without having a better, more stable, gold standard for classifier training. However, if the gold standard improves, it may well be possible that the automatic classifier will outperform the visual classification made by one individual.

A second data set, consisting of cells exposed to different concentrations of agonist, made it possible to test the automatic classifier independent of visual classifications. Because Rac1 activation is expected to be related to concentration of agonist (in a limited range), one expects to see an increase in cell activation at increased concentration of agonist. By using the classifier created from the first

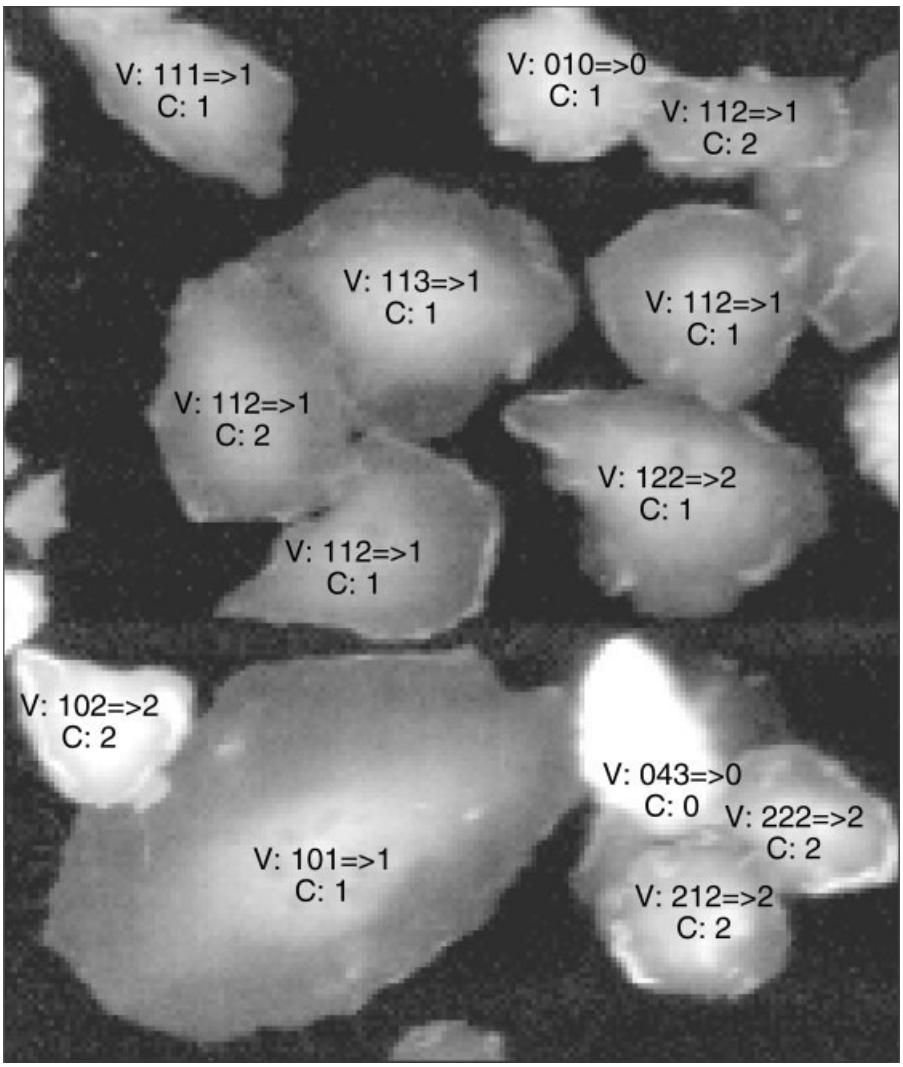

FIG. 4. Classification results superimposed on cytoplasm image after addition of agonist. Classes: −1, no GFP-Rac1 expression; 0, no ruffle formation; 1, low to medium response; 2, high response; 3, unclear; 4, no visual assessment. V, XXX⇒X corresponds to visual classifiers A1, A2, and B and the weighted summary, *W*. C, X corresponds to the automatic classification provided by the described method.

data set, the cells of the second data set were classified. The results show that the proportion of activated cells increases with concentration of agonist, indirectly verifying the performance of the automated image analysis system.

The classification result may be improved by finding more precise features that better describe Rac1 activation and ruffle formation. However, this is of limited value if the reliability of the gold standard does not improve. Use of a classifier that takes the non-normality of some of the included features into account also may be of interest.

The cell analysis methodology described in this paper is very generic in its nature and applicable in a great variety of situations. The suggested cytoplasm segmentation method has shown to be robust and versatile and should be useful in a wide range of similar situations. The described problem-specific features based on ruffle regions and internal drainage regions are designed explicitly for capturing the ruffling of the cells and thus may be of limited value for studying other events. The use of water-

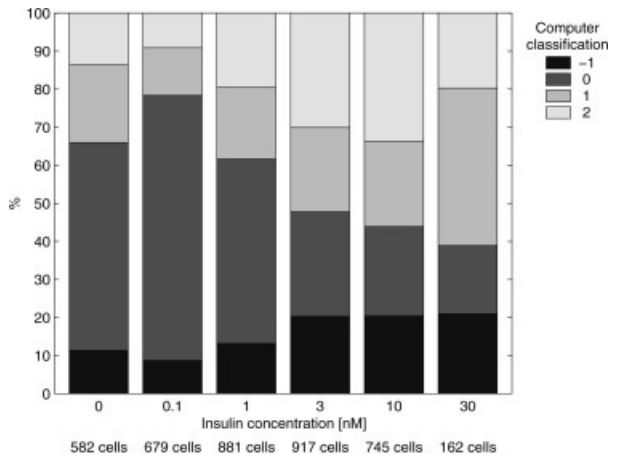

FIG. 5. Relative distribution of the different classes for increasing insulin concentration. The proportion of activated cells (classes 1 and 2) increases with increasing insulin concentration.

shed segmentation as a tool for finding and defining subcellular regions should, however, be important for other studies.

Analysis of time-dependent events, such as ruffle formation, in single cells is of great importance for cell-based drug screening and evaluation. Fully automatic image-based systems, such as the one described in this paper, promise a rapid route for the analysis of biological events that demand high spatial and temporal resolution.

LITERATURE CITED

1. Didsbury J, Weber RF, Bokoch G, Evans T, Snyderman R. *Rac*, a novel *ras*-related family of proteins that are botulinum toxin substrates. J Biol Chem 1989;264:16378–16382.
2. Kjoller L, Hall A. Signaling to rho GTPases. Exp Cell Res 1999;253: 166–179.
3. Azuma T, Witke W, Stossel TP, Hartwig JH, Kwiatkowski DJ. Gelsolin is a downstream effector of rac for fibroblast motility. EMBO J 1998; 17:1362–1370.
4. Miki H, Suetsugu S, Takenawa T. WAVE, a novel WASP-family protein involved in actin reorganization induced by Rac. EMBO J 1998;17: 6932–6941.
5. Wählby C, Lindblad J, Vondrus M, Bengtsson E, Björkesten L. Algorithms for cytoplasm segmentation of fluorescence labelled cells. Anal Cell Pathol 2002;24:101–111.
6. Beucher S. The watershed transformation applied to image segmentation. Scanning Microsc 1992;6:299–314.
7. Vincent L, Soille P. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. IEEE Trans Pattern Anal Mach Intell 1991;13:583–597.
8. Vincent L. Morphological grayscale reconstruction in image analysis: Applications and efficient algorithms. IEEE Trans Image Process 1993; 2:176–201.
9. Simon I, Pound CR, Partin AW, Clemens JQ, Christens-Barry WA. Automated image analysis system for detecting boundaries of live prostate cancer cells. Cytometry 1998;31:287–294.
10. Wu K, Gauthier D, Levine MD. Live cell image segmentation. IEEE Trans Biomed Eng 1995;42:1–12.
11. Ortiz de Solorzano C, Malladi R, Lelievre S, Lockett S. Segmentation of nuclei and cells using membrane related protein markers. J Microsc 2001;201:404–415.
12. Beliën JAM, van Ginkel HAHM, Tekola P, Ploeger LS, Poulin NM, Baak JPA, van Diest PJ. Confocal DNA cytometry: a contour-based segmentation algorithm for automated three-dimensional image segmentation. Cytometry 2002;49:12–21.
13. Malpica N, Ortiz de Solorzano C, Vaquero JJ, Santos A, Vallcorba I, Garcia-Sagredo JM, del Pozo F. Applying watershed algorithms to the segmentation of clustered nuclei. Cytometry 1997;28:289–297.
14. Umesh Adiga PS, Chaudhuri BB. An efficient method based on watershed and rule-based merging for segmentation of 3-D histo-pathological images. Pattern Recog 2001;34:1449–1458.
15. Breu H, Gil J, Kirkpatrick D, Werman M. Linear time Euclidean distance transform algorithms. IEEE Trans Pattern Anal Mach Intell 1995;17:529–533.
16. Ranefall P, L. Egevad BN, Bengtsson E. A new method for segmentation of color images applied to immunohistochemically stained cell nuclei. Anal Cell Pathol 1997;15:145–156.
17. Rodenacker K, Bengtsson E. A feature set for cytometry on digitized microscopic images. Anal Cell Pathol 2003;25:1–36.
18. Duda RO, Hart PE. Pattern classification and scene analysis. New York: John Wiley & Sons; 1973.
19. Raudys SJ, Jain AK. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. IEEE Trans Pattern Anal Mach Intell 1991;13:252–264.
20. Dash M, Liu H. Feature selection for classification. Intell Data Anal 1997;1:131–156.
21. Queiros CE, Gelsema ES. A note on some feature selection criteria. Pattern Recog Lett 1989;10:155–158.
22. Pudil P, Ferri FJ, Novovičová J, Kittler J. Floating search methods for feature selection with nonmonotonic criterion functions. In: Proceedings of the 12th International Conference on Pattern Recognition (ICPR). 1994. p 279–283.
23. Pudil P, Novovičová J, Kittler J. Floating search methods in feature-selection. Pattern Recog Lett 1994;15:1119–1125.
24. Ferri F, Pudil P, Hatef M, Kittler J. Comparative study of techniques for large-scale feature selection. In: Gelsema ES, Kanal LN, editors. Pattern recognition in practice IV: multiple paradigms, comparative studies and hybrid systems. New York: Elsevier; 1994. p 403–413.
25. Jain AK, Zongker D. Feature selection: Evaluation, application, and small sample performance. IEEE Trans Pattern Anal Mach Intell 1997; 19:153–158.
26. Kudo M, Somol P, Pudil P, Shimbo M, Sklansky J. Comparison of classifier-specific feature selection algorithms. In: Ferri F, Iñesta J, Amin A, Pudil P, editors. Advances in pattern recognition. Joint IAPR International Workshops SSPR 2000 and SPR 2000; LNCS, Alicante, Spain. Volume 1876. New York: Springer-Verlag; 2000. p 677–686.
27. Habbema JDF, Hermans J. Selection of variables in discriminant analysis by F-statistic and error rate. Technometrics 1977;19:487–493.
28. Fleiss JL, Cohen J, Everitt BS. Large sample standard errors of kappa and weighted kappa. Psychol Bull 1969;72:323–327.
29. Costanza MC, Afifi AA. Comparison of stopping rules in forward stepwise discriminant analysis. J Am Stat Assoc 1979;74:777–785.
30. Gilles S, Brady M, Declerck J, Thirion J, Ayache N. Bias field correction of breast MR images. In: Proceedings of the 4th International Conference on Visualization in Biomed Comput (VBC); Hamburg, Germany. New York: Springer-Verlag; 1996. p 153–158.
31. Lindblad J, Bengtsson E. A comparison of methods for estimation of intensity nonuniformities in 2D and 3D microscope images of fluorescence stained cells. In: Proceedings of the 12th Scandinavian Conference on Image Analysis (SCIA); Bergen, Norway. 2001. p 264–271.
32. Lancaster P, Šalkauskas K. Curve and surface fitting, an introduction. London: Academic Press; 1986.
33. Arcelli C, Sanniti di Baja G. Computing Voronoi diagrams in digital pictures. Pattern Recog Lett 1986;4:383–389.
34. Schulerud H, Albregtsen F. Many are called, but few are chosen. feature selection and error estimation in high dimensional spaces. Comput Methods Programs Biomed. Forthcoming.
35. Mardia KV. Measures of multivariate skewness and kurtosis with applications. Biometrika 1970;57:519–530.
36. Prokop RJ, Reeves AP. A survey of moment-based techniques for unoccluded object representation and recognition. CVGIP Graph Models Image Process 1992;54:438–460.
37. Dorst L, Smeulders AWM. Length estimators for digitized contours. Comput Vis Graphics Image Process 1987;40:311–333.
38. Kulpa Z. Area and perimeter measurement of blobs in discrete binary pictures. Comput Graphics Image Process 1977;6:434–454.
39. Barber CB, Dobkin DP, Huhdanpaa H. The quickhull algorithm for convex hulls. ACM Trans Math Software 1996;22:469–483.
40. Rencher AC, Larson SF. Bias in Wilks's Λ in stepwise discriminant analysis. Technometrics 1980;22:349–356.
41. Guorong X, Peiqi C, Minhui W. Bhattacharyya distance feature selection. In: Proceedings of the 13th International Conference on Pattern Recognition (ICPR). Volume 2. New York: IEEE; 1996. p 195–199.
42. Stenkvist B, Bengtsson E, Eriksson O, Jarkrans T, Nordin B, Westman-Naeser S. Histopathological systems of breast cancer classification: reproducibility and clinical significance. J Clin Pathol 1983;36:392–398.
43. Murray GD. A cautionary note on selection of variables in discriminant analysis. Appl Stat 1977;26:246–250.
44. Stone M. Cross-validation choice and assessment of statistical predictions. J R Stat Soc 1974;B36:111–147.