Pattern Recognition 61 (2017) 234–244

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Pattern recognition and classification of two cancer cell lines by diffraction imaging at multiple pixel distances

He Wang [a], Yuanming Feng [a], Yu Sa [a], Jun Q. Lu [b], Junhua Ding [c], Jun Zhang [a], Xin-Hua Hu [b,*]

[a] Department of Biomedical Engineering, Tianjin University, Tianjin 300072, China
[b] Department of Physics, East Carolina University, Greenville, NC 27858, USA
[c] Department of Computer Science, East Carolina University, Greenville, NC 27858, USA

ARTICLE INFO

Article history:
Received 9 February 2016
Received in revised form
23 May 2016
Accepted 22 July 2016
Available online 25 July 2016

Keywords:
Single-cell assay
Image pattern analysis
Diffraction imaging
Cell classification
Light scattering
Flow cytometry
Cancer cells

ABSTRACT

Rapid and label-free imaging methods for accurate cell classification are highly desired for biology and clinical research. To improve consistency of classification performance, we have developed an approach of pattern analysis by gray level co-occurrence matrix (GLCM) algorithm to extract textural features at multiple pixel distances from cross-polarized diffraction image (p-DI) pairs, which were acquired with a method of polarization diffraction imaging flow cytometry using one time-delay-integration camera for significantly reduced blurring. Support vector machine (SVM) based classification was performed to discriminate HL-60 from MCF-7 cells using the GLCM features and consistency of optimized SVM classifiers was evaluated on three test data sets. It has been shown that the classification accuracy of the best performing SVM classifiers at or above 98.0% can be achieved among all four data sets for each of the three incident beam polarizations. These results suggest that the p-DI pair data provide a new platform for rapid and label-free classification of single cells with high and consistent accuracy.

1. Introduction

Cell classification by recognition of image patterns is of fundamental interest and can have wide applications in life science and clinics [1–8]. Compared to microscopy, flow cytometry (FCM) with imaging capability allows rapid data acquisition and extraction of pattern parameters from large numbers of cells on a single-cell basis [9]. The high rate of image acquisition through FCM demands development of automated image analysis algorithms to process and analyze the big data of acquired images. Based on previous studies of coherent light scattering [10–15], we have developed a polarization diffraction imaging flow cytometry (p-DIFC) method to measure the spatial distribution of light scattered by single cells illuminated by a linearly polarized laser beam [16–22]. Different from fluorescence imaging, the cross-polarized diffraction image (p-DI) pair data acquired by the p-DIFC method record intensity distribution of coherent light scattered by a cell due to the heterogeneous 3D distribution of intracellular refractive index and needs no cell staining. The acquired p-DI pair data are results of coherent superposition of wavefields emitted by the induced dipoles of molecules in the imaged cell and thus present speckle patterns carrying molecular and morphological information of the cell [12,13,15]. We have shown that the p-DI pair data can be used to accurately distinguish two cell lines of high morphological similarity derived from human T and B cancer cells and two types of prostate cells derived from cancer and normal cells [20,22].

Conventional cell images are typically acquired with non-coherent fluorescent light through a microscope or imaging FCM which present 2D projections of a 3D object. Image segmentation is generally needed for further analysis [9]. In contrast, the speckle patterns in a p-DI pair result from superposition of coherent wavefields from all excited intracellular molecules as “digital holograms” and needs no segmentation. A pixel-based global image processing algorithm often suffices to quantify the patterns or textures of such images that can be automated for rapid processing. We have developed a gray-level-co-occurrence-matrix (GLCM) based software to quantitatively characterize textures of p-DI pair data [22–25]. The GLCM parameters were determined from a matrix of elements given by the co-occurring probabilities of paired pixels separated by $\mathbf{d}$ as the displacement vector with $|\mathbf{d}|=1$. A total of 38 parameters were obtained from each p-DI pair to form feature vectors, which were used to train support vector machine (SVM) classifiers with different kernels. With the above algorithms, we have shown that the p-DIFC method performs well for accurate and label-free cell classification through automated image pattern recognition.

Despite the attractive qualities of the p-DIFC method, however,

* Corresponding author.
E-mail address: hux@ecu.edu (X.-H. Hu).

http://dx.doi.org/10.1016/j.patcog.2016.07.035
0031-3203/

we have found that the performances of the trained SVM classifiers are not consistent when they were applied to test sets of p-DI pairs data acquired in different runs of measurements. The causes of inconsistency may relate to the variation of diffraction patterns in different data sets, which are due to cell speed fluctuations leading to different degree of blurring for images acquired with conventional CCD cameras, positioning errors of cells relative to the focus of incident beam and the high sensitivity of extracted GLCM parameters on fine pattern changes among pairs of nearest-neighbor pixels that are unrelated to cell morphology or molecular composition. In our previous study of T versus B cell lines and PC3 versus PCS prostate cells [20,22], SVM classifiers had to be trained and tested with the p-DI pairs acquired in the same measurement run for achieving high values of classification accuracy A. The value of A decreases markedly if the SVM classifier trained with data from one run was applied to the test data acquired in another run on a different day [22]. For example, the value of A reduces from 99.5% to 62.8% in the case of PC3 versus PCS cells. Such reduction in performance prevents application of the p-DIFC method with pre-trained SVM classifiers on p-DI data acquired later.

The current study focuses on pattern analysis of p-DI data and effect of blurring as a part of our research efforts to solve the accuracy-dropping problem. A new configuration of illumination and imaging has been developed to eliminate or significantly reduce motion blur with one time-delay-integration (TDI) CCD camera and decrease the sensitivity of image patterns of acquired data with enlarged focal spot for the incident beam [21]. With this imaging configuration we have investigated the GLCM approach of pattern analysis with $d \geq 1$ to improve the robustness of the SVM classifiers. In this report, we present the results of pattern recognition and classification on two cancer cell lines, HL-60 versus MCF-7, by the p-DI data acquired with the new imaging configuration. The dependence of p-DI parameters and classification accuracy on d has been analyzed to examine the benefits of using different values of d for improved performance. We have also blurred the measured p-DI data with no or little blurring by window smoothing to investigate the effect of blurring on classification. These results demonstrate that the new imaging configuration and GLCM analysis with different d can improve significantly the robustness of SVM based classification.

2. Related work

Automated image analysis for cell classification often proceed by extracting feature parameters related to cell morphology and/or image texture. Morphological parameters can be regarded as quantitative extension of human perception and obtained by characterizing geometric structures of organelles in images acquired by bright-field or fluorescent microscopy, which are typically performed with cells stained with contrast or fluorescent reagents. In deriving these parameters, image segmentation is necessary on the basis of unique correspondence between regions of interest (ROI) and individual organelles. By comparison, characterization of texture can be performed either globally in a given image or locally over clustered pixels to quantify patterns of intensity variation. In the latter case, proper image segmentation may also be needed. With wide availability of algorithms and ever increasing computing power per unit cost, the two types of image parameters are frequently combined to achieve best outcomes of classification. Below, we briefly review different methods of automated cell classification using morphology and texture features as discriminators in terms of their performance.

In an early study of multiclass classification, 9 morphology and 4 texture features have been extracted from 2D images to discriminate 7 *Eimeria* parasite species of different shapes. The parameters were combined as descriptors for training Bayesian classifiers and an overall correct classification rate or classification accuracy of 86.8% was obtained over a data set of 3891 oocyst images [2]. To expand from individual efforts, a contest of HEp-2 Cells Classification has been held in 2012 to evaluate different methods on a common test data set. The contest was designed to discriminate 6 antinuclear autoantibody fluorescence patterns on a given data set of 28 indirect immunofluorescence (IIF) images containing 1457 HEp-2 cells, with half of the IIF images used as the training data [6]. A total of 28 automated methods have been submitted and tested to extract morphological and/or texture features as the input to different classifiers including the SVM and k-nearest-neighbor methods, among others. The test data set consists of manually segmented images of single cells from 14 reserved IIF images. The mean accuracy of classifying the IIF patterns in the segmented cell images into 6 classes ranged from 21.4% to 70.4% for the submitted methods, which were less than the baseline value of 77.5% achieved by a specialist against the ground truth data of the 28 IIF images established by experts [6,26]. Notably, the best performing method utilized the image texture information by quantifying the co-occurrence of local binary differences of neighboring pixel intensities among image pixels [4]. Following the 2012 competition, additional studies have been performed to further improve the performance of automated classification of the HEp-2 cells [7]. It has been reported, for example, that combination of multiple SVM classifiers trained with different sets of sparsely encoded texture features could produce a mean classification accuracy of 87.1% on the IIF patterns of HEp-2 cells on heldout test data sets [8].

Several methods among those submitted to the HEp-2 cell contest discussed above utilized texture features extracted by the GLCM algorithm. Among these, the highest value of mean classification accuracy was achieved at 63.0% on the common test data set using 4 GLCM parameters coupled with additional features of morphology and gradient and multi-class SVM classifiers [5,6]. Textural and morphological features have also been extracted to classify histopathological images of colon tissues into three classes of normal, low- and high-grade adenocarcinoma and a mean value of 57.1% was achieved for accuracy using only the GLCM features [3]. Other applications of image texture analysis include study of apoptotic cells which identified correlations between the GLCM features in ROI of cell images to the changes of nuclear morphology characterizing apoptosis [1]. It is worth noting that in all studies discussed in this section, the image feature extraction was preceded by manual segmentation to remove regions in an image that do not relate to the organelles of interest. For the images acquired with the p-DIFC method presented here, segmentation is unnecessary since all light signals presented by the pixels in a diffraction image are from the imaged cell. Furthermore, each non-dark pixel represents the intensity of total electromagnetic fields emitted by induced molecular dipoles inside the imaged cell. Hence, the p-DI pairs present a new data platform to investigate cell classification with the attractive qualities of fast acquisition, low background noise and capacity for fully automated processing.

3. Methods

3.1. The p-DIFC system and cell measurement

Details of the p-DIFC system have been published elsewhere for cell positioning by hydrodynamic focusing in a square flow channel and imaging of coherent light scatter [16–18,20,21]. Briefly, a continuous-wave solid state laser (MGL-III-532-100, CNI) was used to produce an incident beam of 532 nm in wavelength. Two cylindrical lenses of 500 mm and 60 mm in focal lengths

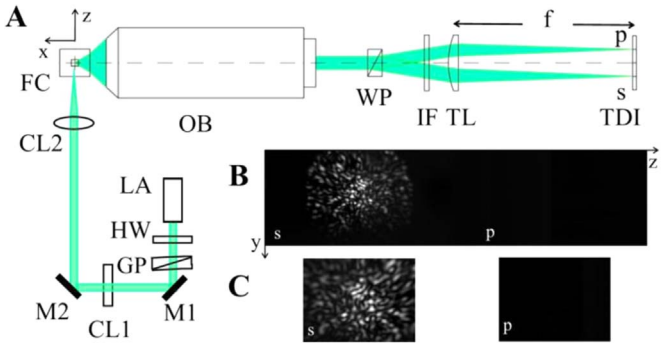

Fig. 1. (A) The schematic of the optical setup for acquisition of p-DI data: BE: beam expander; LA: cw laser of 532 nm in wavelength; HW: half-wave plate; GP: Glan-Thompson polarizer, M1/M2: mirrors; CL1/CL2: cylindrical lenses; FC: flow chamber; OB: infinity-corrected objective; WP: Wollaston prism; IF: 532 nm interference filter; TL: tube lens of f in focal length. The labels of s and p indicate the polarization of the scattered light. (B) A p-DI image of 20,148 × 512 and 12-bit pixel acquired from the TDI camera. (C) The p-DI pair of 400 × 300 pixels for each image obtained from the image in (B).

place the focus of the incident beam on the core fluid carrying the cells in an elliptical cross-section of major diameter of about 300 μm along the flow direction or y-axis and minor diameter of about 50 μm along the *x*-axis. The profile of linearly polarized incident beam propagating along the *z*-axis was close to Gaussian and the power was measured as P_0 before the CL2 lens shown in Fig. 1(A) and adjusted with a half-wave plate. The loss of the optical power by the index-mismatch surfaces of the CL2 lens and flow chamber before the imaged cell was estimated to be about 17%.

The polarizing direction of the incident beam was set to one of three directions of horizontal (hor), vertical (ver) or 45° from horizontal by rotating a polarizing prism. The light scattered by the cell passing through the beam focus was acquired by an infinity-corrected microscope objective of 0.55 in NA and 50x in magnification (378-805-3, Mitutoyo) along the *x*-axis with the scattering angle of $\theta_s = 90°$ for the center of the light collection cone. A Wollaston prism (LSP-3A14, Laser Institute, QFNU) was employed to divide the scattered light into s- and p-polarized beams, separated by an angle of 20°, and projected to the sensor of a TDI camera by a tube lens of 75 mm in focal length. The in-house developed TDI camera provides clock pulse train of adjustable frequency to synchronize the charge transfer speed of the TDI line pixels with the moving speed of the imaged cell for blur-free imaging [21]. The imaging unit as a whole was translated off focus toward the flow chamber by a distance of $\Delta x = 400$ μm from the focusing position conjugate to the imaged cell or core fluid. It has been shown that at the non-conjugate positions of $\Delta x > 0$ the acquired p-DI pairs present patterns of diffraction in high-fidelity because of the unique correspondence between the angular positions of the coherent light scatter and the imager pixel positions [17,27]. The cone angle θ_{wm} of the scattered light collected by the imaging unit with $\Delta x = 400$ μm has been estimated to be about 21° on an input plane located at 150 μm from the imaged cell within the water-filled flow chamber [27]. The throughput of the p-DIFC measurement was maintained at about 1–4 cells/s.

Two cell lines of HL-60 and MCF-7 derived respectively from human chronic myelogenous leukemia and human breast carcinoma were used for this study (CCL-240 and HTB-22, ATCC). The cells were maintained in the RPMI 1640 culture medium supplemented with 10% fetal bovine serum at the Tianjin Medical University. The adherent MCF-7 cells in their logarithmic phases of growth were detached from culture plates with trypsin-EDTA solution, re-suspended in culture medium. The concentrations of the cell suspension samples were adjusted to a value of about 1×10^6 cells/ml and kept on ice before p-DIFC measurement. Three runs or groups of measurements have been performed on the HL-60 and MCF-7 cells in different days. The image data of the two cell types acquired in first group were divided into a training data set and a test data set. All image data acquired in second and third groups were used as two additional test data sets to assess the performance and consistency of trained SVM classifiers in different test data sets measured in different groups. The details of the data sets are provided in Table 1.

Table 1
Acquisition and best performing SVM classifier parameters[1].

Measurement group	Incident polarization	Cell type	N_{tra}	N_{tes}	A_{av}(%)	|**d**| kernel, N_m	Top 2 parameters	A_1(%)	A (%)
#1	ver	HL-60	600	427	99.6	1	s-DIS	95.7	99.4
		MCF-7	600	302		Poly, 14	s-DEN	81.7	
	hor	HL-60	600	202	98.1	22	p-VAR	95.9	98.0
		MCF-7	600	252		Poly, 11	p-DIS	90.2	
	45°	HL-60	600	299	99.8	22	s-IMAX	96.6	99.5
		MCF-7	600	293		Poly, 16	s-MEA	94.0	
#2	ver	HL-60	0	1003		Same as group #1/ver polarization			99.2
		MCF-7	0	1096					
	hor	HL-60	0	841		Same as group #1/hor polarization			98.5
		MCF-7	0	741					
	45°	HL-60	0	908		Same as group #1/45° polarization			99.4
		MCF-7	0	916					
#3	ver	HL-60	0	1044		Same as group #1/ver polarization			98.9
		MCF-7	0	1070					
	hor	HL-60	0	681		Same as group #1/hor polarization			98.5
		MCF-7	0	852					
	45°	HL-60	0	907		Same as group #1/45° polarization			99.3
		MCF-7	0	917					

[1] N_{tra} = number of p-DI pairs in the training data set; N_{tes} = number of p-DI pairs in the test data set; A_{av} = averaged classification accuracy of the best SVM classifier on training data set; |**d**|, kernel and N_m: the pixel distance, kernel function and number of components for feature vector of the best performing SVM classifier; top 2 parameters: the image parameters used as first two components of the feature vector of the best performing SVM classifier; A_1 = highest classification accuracy of single parameter; A = classification accuracy of the best performing SVM classifier applied to a test data set.

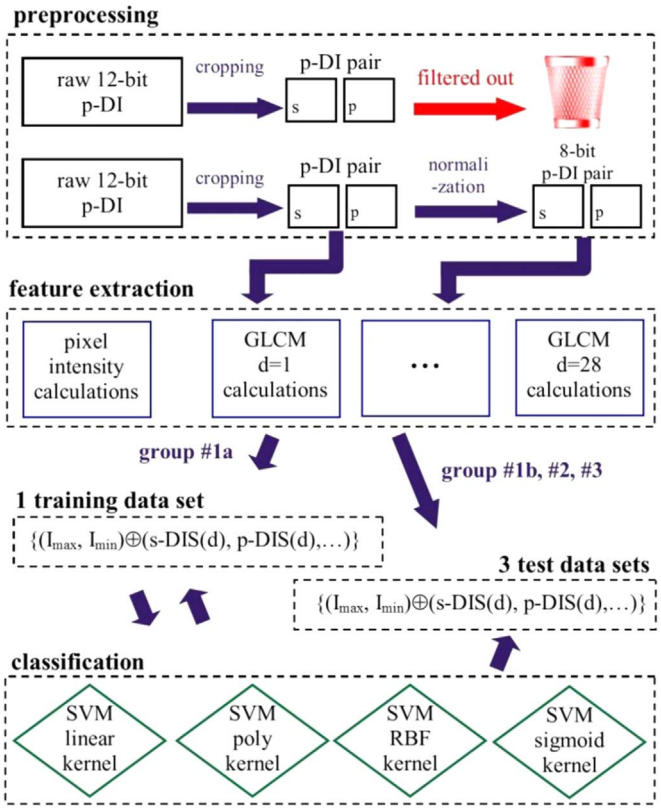

Fig. 2. The block diagram of diffraction image processing and cell classification.

3.2. Data processing

Fig. 2 presents a block diagram of data processing followed by cell classification with 1 training and 3 test data sets. Software for imaging preprocessing and GLCM parameter calculation has been developed in-house on the platform of MATLAB (7.2, MathWorks) for this study. As shown in Fig. 1(B), a raw 12-bit image of 2014×512 pixels acquired from the TDI camera consists of two equal sections recording the s-polarized scattered light intensity on the left and p-scattered light on the right. Each raw image was cropped with the preprocessing software to obtain a p-DI pair as displayed in Figs. 1(C) and 2 by centering the two regions to be cropped on the column of $z=512$ for the left section and $z=1464$ for the right section. Both regions were centered on the same row of maximum total pixel intensity to obtain two cropped regions of 400×300 pixels as one p-DI pair. After cropping, the overexposed and underexposed p-DI pairs were filtered out with the former defined as the pairs of one or both images having saturated pixels (with intensity reaching the maximum value of 4095 for 12-bit images) more than 1% of the total pixels and the latter as the pairs of both images having average pixel intensities less than 2% of the maximum pixel intensity. Additionally, p-DI pairs with strip patterns of high symmetry, large speckles or significant degree of blurring were removed since these have been shown to associate with spherical particles, aggregated small particles or cellular debris instead of intact cells [19] or caused by large mismatch between cell speed and line transfer frequency of TDI sensor due to flow instability [21]. The preprocessing filtered out about 50–70% of the acquired raw images in different groups of measurements. Table 1 lists the total numbers of p-DI pairs used as training and test data for subsequent GLCM calculations and classification study.

3.3. Feature extraction

After preprocessing, the 12-bit p-DI pairs were transformed linearly into 8-bit pairs by normalizing the minimum and maximum 12-bit pixel intensities into 0 and 255 of the 8-bit intensities. Each image was processed by the GLCM software to store in a square matrix $\mathbf{p}(\mathbf{d})$ the co-occurring probabilities of intensities or gray levels of paired pixels separated by $\mathbf{d}$ with the row and column positions of matrix elements given by the intensities [22–25]. The reduction from 12-bit to 8-bit in p-DI data can significantly speed up GLCM calculations without significant loss of dynamic range since the size of $\mathbf{p}(\mathbf{d})$ is determined by the bit depth of pixel intensities. The vector $\mathbf{d}$ has 4 possible directions of 0°, 45°, 90° and 135° and each matrix $\mathbf{p}(\mathrm{d})$ was obtained as an average of $\mathbf{p}(\mathbf{d})$ of same d over the 4 directions. After the determination of $\mathbf{p}(\mathrm{d})$, a total of 17 GLCM parameters were obtained from each input images of 400×300 pixels, which were combined with 2 pixel intensity parameters of minimum and maximum to characterize the 12-bit image. This led to a set of 38 parameter, $\{g_{im}(d),\ m=1,\ \ldots,\ 38\}$, to form different feature vectors as representations of one p-DI pair acquired from the ith imaged cell. For this study, the values of d was varied from 1 to 28. The definitions of the GLCM parameters are based on [23] and a list of parameter definitions is provided online [28].

3.4. SVM based classification

The GLCM based image processing allows a representation of the ith imaged cell by a feature parameter given by $\boldsymbol{c}_i(N_m, d) = \sum_{m=1}^{N_m} g_{im}(d)\mathbf{u}_m$ for each d in a parameter space Σ_p of N_m-dimension, where $\boldsymbol{u}_m$ are unit vectors in Σ_p. The vector $\boldsymbol{c}_i(N_m, d)$ consists of either a portion or all of the parameter set $\{g_{im}(d), m=1, \ldots, 38\}$ with $1 \le N_m \le 38$. As a machine learning algorithm widely used for classification and regression, SVM establishes a feature space E by mapping $\{\boldsymbol{c}_i(N_m, d), i=1, \ldots, N\}$ for a data set of N cells using a kernel function $K(\boldsymbol{c}_i(N_m, d), \boldsymbol{c}_j(N_m, d))$ [8,29–31]. We employed a previously developed classification software for this study based on an open-source code package of SVM (LIBSVM 2.86), which provides choices of four kernel functions: linear, polynomial (poly), Gaussian radial basis function (RBF) and sigmoid [31].

To assess the performance of an SVM classifier on a data set, an accuracy of classification A is defined to measure the outcomes of classifier prediction against the ground truth established by the known cell type, which is given by

$$A = \frac{TP + TN}{TP + TN + FP + FN}. \quad (1)$$

where TP is the number of correctly predicted p-DI pairs acquired from HL-60 cells, TN the number of correctly predicted p-DI pairs from MCF-7 cells, FP the number of p-DI pairs of MCF-7 cells but incorrectly predicted as of HL-60 cells and FN the number of p-DI pairs of HL-60 cells but incorrectly predicted as of MCF-7 cells.

3.5. Blurring of measured p-DI data

To analyze the effect of blurring on classification, we employed a MATLAB function (fspecial) to "blur" p-DI pair data acquired with the TDI camera's pixel lines transfer synchronized with the moving "rays" of scattered light along the y-axis. The MATLAB function performs window smoothing on an input image by a width of w pixels along a selected direction. The window smooth or blur width w was set to vary between 5 and 30 pixels for this study. The use of window smoothing on a p-DI pair to generate a sequence of images with different degrees of blur is advantageous in

comparison to the use of p-DI data acquired by different cell speeds or TDI's line transfer frequencies [21]. The artificially added blurring allows investigation of the blurring effect on image parameters and classification performance on a p-DI pair from the same cell. We have previously demonstrated the equivalence of image parameter variations of p-DI data between the two cases of blurring made by window smoothing or flow speed change [32].

4. Experimental results

4.1. Analysis of diffraction images

Three groups of measurements have been carried out in different days to acquire p-DI pairs from the HL-60 and MCF-7 cell suspension samples using a prototype system (DIFC2-P3, WavMed Technologies Corp.) equipped with one TDI camera as shown in Fig. 1(A). During each measurement, cell suspensions in volumes of about 0.1 mL were loaded multiple times into the core fluid tube

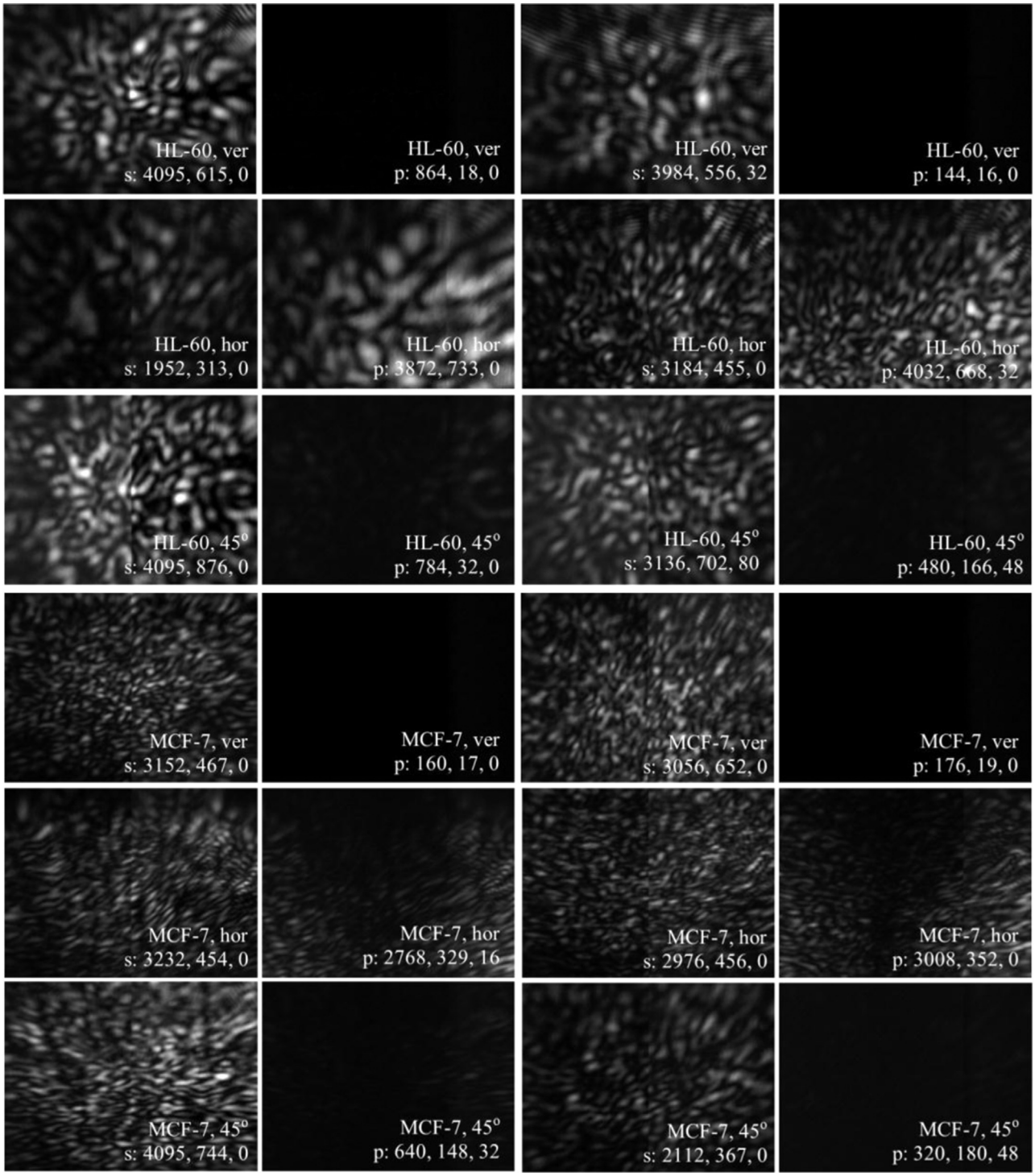

Fig. 3. Representative 12-bit p-DI pairs acquired from two HL-60 and two MCF-7 from the training data set with white for pixel intensity 4095 and black for 0. Each image is labeled with the cell type, incident beam polarization, polarization of scattered light, maximum, average and minimum pixel intensities.

at the top of the flow chamber. A loaded sample was driven by a syringe pump into the core fluid nozzle inside the flow chamber with a channel of 1 mm in side length [18]. The flow speed v of the cells carried by the core fluid was controlled by the syringe pump speed and the air pressure in the sheath reservoir maintained with an air pump. The value of v was determined before cell measurement with microspheres of 9.6 μm in diameter by the length of their shadows under white light illumination by a monitoring CCD camera during a fixed exposure time of 1 ms and off-focus distance Δx set to 0. The values of v was found to be in a range given by 10.6 ± 0.5 mm/s. At the mean value of v, the cell image on the TDI sensor became "stationary" to the pixel lines of the TDI sensor which were transferred at a frequency of $f = 16.67$ kHz and the acquired p-DI data were blur-free or close to blur-free [21,32]. The power of incident beam was measured and kept at 19 mW, 50 mW and 39 mW for beam polarizations of ver, hor and 45° respectively for the HL-60 cells for all measurements. The incident beam power was adjusted and kept at 9 mW, 80 mW and 23 mW for the MCF-7 cells. About 2000 cells were imaged for each cell type at one of three incident beam polarizations of the vertical (ver), horizontal (hor) and 45° relative to the incident plane of x–z.

Fig. 3 shows examples of the 12-bit p-DI pairs of two HL-60 and MCF-7 cells acquired with each of three incident beam polarizations. It is clear from the images that on average cells scatter more light with s-polarization for an incident laser beam of s-polarization for the case of ver or equally among the s- and p-polarization directions for the case of 45°. This is due to the fact that s-polarized incident beam induces molecular dipoles of the imaged cell oriented along the y-axis and p-polarized incident beam induces molecular dipoles along the x-axis. For side scattering directions centered along the x-axis, or $\theta_s = 90°$, those dipoles induced by the s-polarized beam (ver) have much higher contribution to light signals than those induced by the p-polarized beam (hor) for the transverse nature of light wavefields. By examining the maximum, average and minimum pixel intensities of the s- and p-polarized DI data in Fig. 3, one can observe that the actual distribution of scattered light energy among the two images of a p-DI pair depends on the detailed features of morphology and molecular dipoles of the imaged cell. Therefore, the p-DI pair carries rich information for single cell assay and classification. We also note that some measured p-DI pairs after preprocessing still show some degree of blurring as a result of flow speed variation.

4.2. GLCM parameter extraction and SVM classification at different $|d|$

After preprocessing, we performed GLCM calculations on all normalized p-DI pairs listed in Table 1 at different d to obtain 38 GLCM and pixel intensity parameters as $\{g_{im}(d)\}$ from each pair. The associated ith cell can then be represented by a set of feature vectors of $\boldsymbol{c}_i(N_m, d)$ of N_m components selected from the 38 parameters according to their performance on single-parameter classification. As examples, Fig. 4 presents four GLCM parameters calculated from the four data sets as functions of d to illustrate their dependences on d and differences between the two cell lines. The four parameters were chosen for their good performance in classification of cells measured with incident beam polarizations denoted as ver and hor and are listed as the top ranked parameters in Table 1. The GLCM parameters of dissimilarity (DIS) extracted from s- and p-polarized diffraction images, s-DIS and p-DIS, exhibit opposite dependences among the two cell lines on d. The parameter s-DIS has larger values for MCF-7 cells and differs significantly from those for HL-cells for small values of d between 4 and 13. The trends change for p-DIS which are of larger values for HL-60 cells and rise faster than those for MFH-7 cells for large values of d between 16 and 28. It took about 0.42 s on average to obtain the 38 feature parameters from each p-DI pair at a particular value of d on a computer with one Intel i5 CPU (4 cores) of 3.1 GHz with no optimization for multi-core execution.

With the 38 parameters, SVM classification and optimization were carried out in two steps for a particular kernel function and value of d. The first step was to rank the parameters with the training data set formed by HL-60 and MCF-7 cells of $N_{tra} = 600$ as shown in Table 1 for each of the three incident beam polarizations of ver, hor or 45°. Cells were classified by their feature vectors of single component as $\{\boldsymbol{c}_i(N_m = 1, d), i = 1, \ldots N_{tra}\}$ by forming SVM

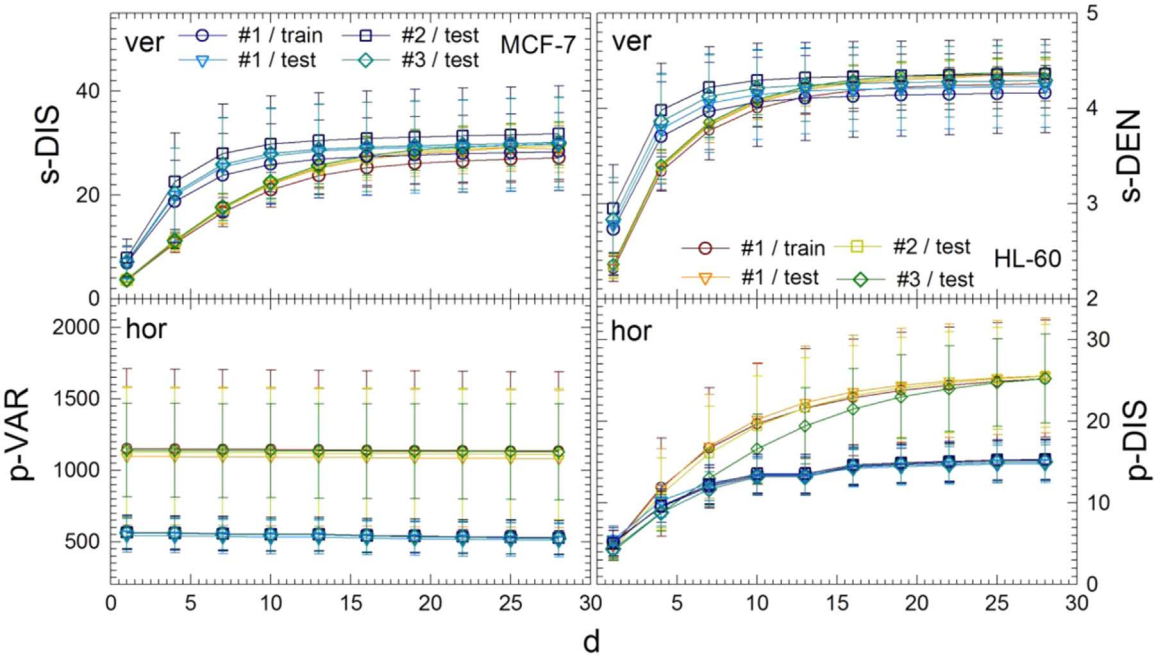

Fig. 4. The values of four GLCM parameters listed in Table 1 as functions of pixel distance d. The GLCM parameters of s-DIS and p-DIS are, respectively, dissimilarity of s- and p-polarized DI; s-DEN is the difference entropy of s-polarized DI; p-VAR is variance of p-polarized DI. The symbols and error bars represent the means and standard deviations of the parameters calculated from each data set. The lines are for visual guide.

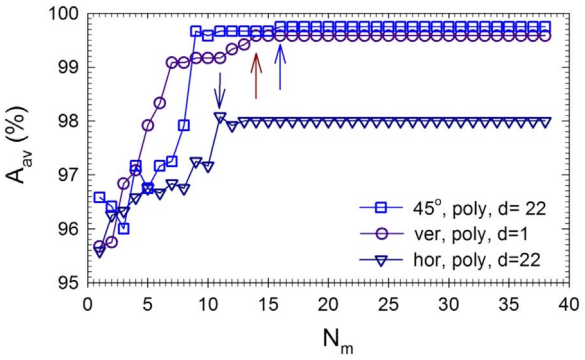

Fig. 5. Three examples of classification accuracy A_{av} versus the number of components N_m of feature vectors for SVM classifiers determined with polynomial kernel function and the training data set acquired with three incident beam polarizations. The GLCM parameters were calculated at two values of pixel distance d as labeled and the arrows indicate the optimized SVM classifiers. The lines are for visual guide.

classifiers consisting of one parameter at a time and a kernel function. A standard scheme of five-fold cross-validation was employed by dividing the training data set into five equal parts randomly with four parts being used as training data and one part as validation data. The procedure was iterated 5 times to obtain an averaged accuracy as A_{av}, which was used to rank the 38 parameters in the order of decreasing A_{av}.

The second step was to compose feature vectors of multiple components arranged in the sequence of parameters' rankings as $\{\boldsymbol{c}_i(N_m, d), i=1, \ldots N_{tra}\}$ with $2 \leq N_m \leq 38$. Different SVM classifiers formed with feature vectors of increasing N_m from 2 to 38 and a kernel function were evaluated by the training data. The same scheme of five-fold cross-validation was used to determine A_{av} for evaluation. Fig. 5 presents three series of A_{av} determined by the training data versus N_m for the polynomial kernel function and different combinations of incident beam polarization and d. An optimized SVM classifier was determined for each combination with the maximum value of A_{av} and, in the cases of multiple maxima, the smallest N_m, as shown by those indicated with arrows in Fig. 5.

After training, the optimized classifier was chosen for each of 120 combinations of incident beam polarization (3), pixel distance d (10) and kernel function (4) on the three test data sets. Fig. 6 presents the values of A_{av} for the training data set and A for the three test data sets of these optimized SVM classifiers versus different d. To identify the best performing SVM classifiers for the sets of p-DI data measured with the same incident beam polarization, we defined a performance index PI to measure a classifier's performance and consistency of cell classification among the four

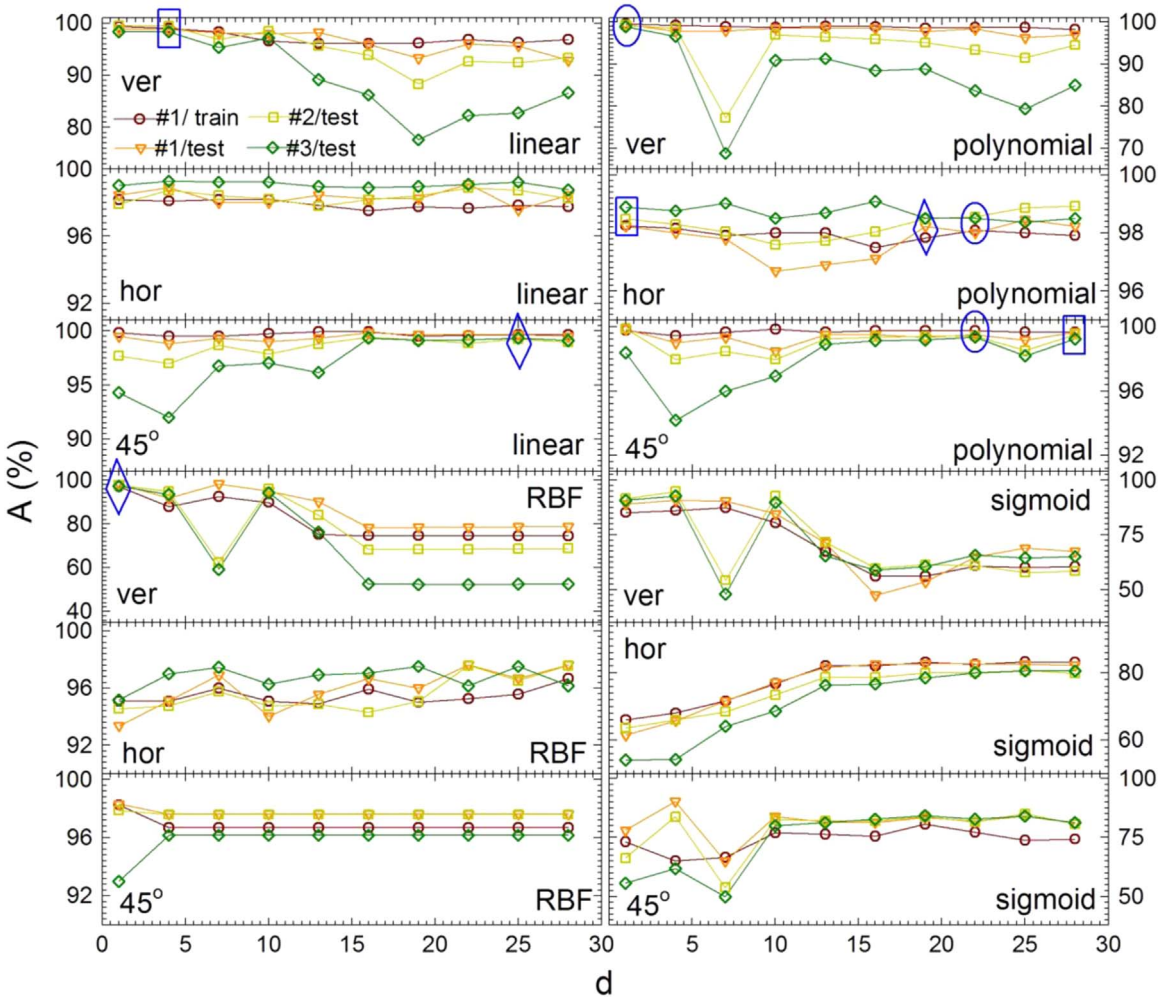

Fig. 6. The classification accuracy A of optimized SVM classifiers obtained with the training data set and three test data sets versus pixel distance d for different combinations of incident beam polarization as labeled on the left and kernel function labeled on the right in each sets of plots. For each incident beam polarization, three enclosed symbols are marked to indicate PI values of corresponding models: those by blue ellipses have the highest values, those by rectangles and diamonds have the second and third highest values and all PI values are given in texts. The lines are for visual guide. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Confusion matrices of three best SVM classifiers with ver polarization.[1]

Classifiers	Test #1		Test #2		Test #3		Parameters; $\mu_A \pm \sigma_A$
1st best	HL-60	MCF-7	HL-60	MCF-7	HL-60	MCF-7	$d=1$, poly, N_m
HL-60	425	2	994	9	1035	9	$=14$;
MCF-7	2	300	7	1089	14	1056	$99.2\% \pm 0.22\%$
2nd best	HL-60	MCF-7	HL-60	MCF-7	HL-60	MCF-7	$d=4$, linear,
HL-60	426	1	1003	0	1041	3	$N_m=19$;
MCF-7	8	294	11	1085	35	1035	$98.8\% \pm 0.52\%$
3rd best	HL-60	MCF-7	HL-60	MCF-7	HL-60	MCF-7	$d=1$, RBF,
HL-60	418	9	966	37	1005	39	$N_m=7$;
MCF-7	4	298	12	1084	18	1052	$97.7\% \pm 0.38\%$

[1] Rows represent ground truth. Test data set #1, #2 and #3 were obtained respectively from measurement group #1, #2 and #3 with total cell numbers given by N_{tes} in Table 1. The mean and standard deviation values of A, μ_A and σ_A, were obtained with the 3 test data sets.

Table 3
Confusion matrices of three best SVM classifiers with hor polarization.[1]

Classifiers	Test #1		Test #2		Test #3		Parameters; $\mu_A \pm \sigma_A$
1st best	HL-60	MCF-7	HL-60	MCF-7	HL-60	MCF-7	$d=22$, poly,
HL-60	199	3	828	13	675	6	$N_m=11$;
MCF-7	6	246	10	731	17	835	$98.4\% \pm 0.24\%$
2nd best	HL-60	MCF-7	HL-60	MCF-7	HL-60	MCF-7	$d=1$, poly,
HL-60	199	3	834	7	679	2	$N_m=16$;
MCF-7	5	247	17	724	15	837	$98.5\% \pm 0.27\%$
3rd best	HL-60	MCF-7	HL-60	MCF-7	HL-60	MCF-7	$d=19$, poly,
HL-60	199	3	828	13	675	6	$N_m=10$;
MCF-7	5	247	11	730	17	835	$98.4\% \pm 0.12\%$

[1] Same as noted in Table 2.

Table 4
Confusion matrices of three best SVM classifiers with 45° polarization.[1]

Classifiers	Test #1		Test #2		Test #3		Parameters; $\mu_A \pm \sigma_A$
1st best	HL-60	MCF-7	HL-60	MCF-7	HL-60	MCF-7	$d=22$, poly,
HL-60	298	1	901	7	896	11	$N_m=16$;
MCF-7	2	291	4	912	1	916	$99.4\% \pm 0.06\%$
2nd best	HL-60	MCF-7	HL-60	MCF-7	HL-60	MCF-7	$d=28$, poly,
HL-60	299	0	906	2	897	10	$N_m=12$;
MCF-7	2	291	7	909	4	913	$99.5\% \pm 0.18\%$
3rd best	HL-60	MCF-7	HL-60	MCF-7	HL-60	MCF-7	$d=25$, linear,
HL-60	298	1	899	9	897	10	$N_m=18$;
MCF-7	1	292	5	911	2	915	$99.4\% \pm 0.18\%$

[1] Same as noted in Table 2.

data sets as

$$PI = \frac{\mu_A}{\sigma_A}, \tag{2}$$

where μ_A and σ_A is respectively the mean value and standard deviation of A_{av} and A for selected data sets. For each incident beam polarization, three best performing SVM classifiers with three largest values of PI are marked in Fig. 6 by enclosed symbols and the corresponding confusion matrices are given in Tables 2 to 4. Table 1 lists the best performing classifier for each polarization. For p-DI data acquired with ver as the incident beam polarization, the values of PI of the marked SVM classifiers with (d, kernel function) are given by 386 with (1, poly), 242 with (4, linear) and 229 with (1, RBF). The same for the case of hor are given by 408 with (22, poly), 370 with (1, poly) and 365 (19, poly) while those for the case of 45° are given by 636 with (22, poly), 565 with (28, poly) and 517 with (25, linear). In some cases, the values of accuracy A_{av} or A show no or little variations for different d if one of the top performing image parameter is given by the maximum pixel intensity of the 12-bit p-DI pair. Overall, the optimized SVM classifiers with linear and polynomial kernels functions yield the most consistent performance at most of the d values among the four choices. It should be noted that the values of μ_A and σ_A given in Tables 2–4 were obtained from the three test data sets, which are slightly different from those used in ranking the classifiers in Fig. 6 obtained from the four sets of training and test data.

The training and test of SVM classifiers were performed with the same computer used to calculate GLCM parameters. The first step of training with the data set of 600 p-DI pairs of each cell type to rank the 38 feature parameters was the most time consuming part and the execution time ranged from 25 to 40 s and the times increased with the classifiers in the order of RBF, sigmoid, polynomial and linear kernel functions. In contrast, the second step of training and testing were much faster that took 1 s or less to complete on one data set.

4.3. Effect of blurring on SVM classification

Finally, we have investigated the effect of blurring on classification by performing window smoothing on the p-DI pair data acquired with two incident beam polarizations of ver and 45°. The acquired p-DI data from the training and three test data sets were "blurred" with different window width w between 5 and 30 pixels along the flowing direction of the cells or the y-axis. In a previous study of blurring effect based on frequency analysis, we have shown that the blurring with $w=30$ on blur-free images was equivalent to the measured images with a cell speed varied from the matched case of 10.6 mm/s to 13.0 mm/s [32]. Fig. 7 (A) presents a set of images for one HL-60 cell and one MCF-7 cell blurred with different w and Fig. 7(B) shows two confocal fluorescent images of two cells to illustrate their morphological differences in size and structures of nucleus and mitochondria. The GLCM parameters were extracted from the blurred p-DI pairs at selected values of d and SVM classifiers were trained with the linear kernel function using the same training data set. The SVM classifier with the highest A_{av} was applied to each of the three test data sets to evaluate the consistency of the optimized model for cell classification. The same procedures were repeated for four values of w and the values of A are plotted in Fig. 7(C) as functions of width w. It can be seen from these results that blurring can reduce significantly the accuracy and consistency of the optimized SVM classifiers on the p-DI data acquired with ver incident beam polarization. For these two cases, the optimized SVM classifiers have very good performance on p-DI data without blurring and the PI value decreases monotonically from 218 at $w=0$ to 20.0 at $w=30$ in the case of $d=1$ and similarly from 242 to 45.7 in case of $d=4$. In contrary, the effect of blurring on A is not significant for p-DI pairs acquired with incident beam polarization of 45° in the cases of $d=16$ and 19 where the PI values fluctuate between 570 and 179 among different w values.

5. Discussion

By acquiring paired p-DI data from imaged cells, the p-DIFC method provides a means to perform image based pattern recognition and cell classification that is significantly different from conventional approaches of microscopy and FCM. As the result of coherent superposition of wavefields emitted by induced molecular dipoles, the p-DI pair data carry rich morphological and molecular information in the forms of complex diffraction patterns. To extract and to utilize the information, however, one has to develop robust algorithms for automated processing of large amount of image data. We have presented here the initial results

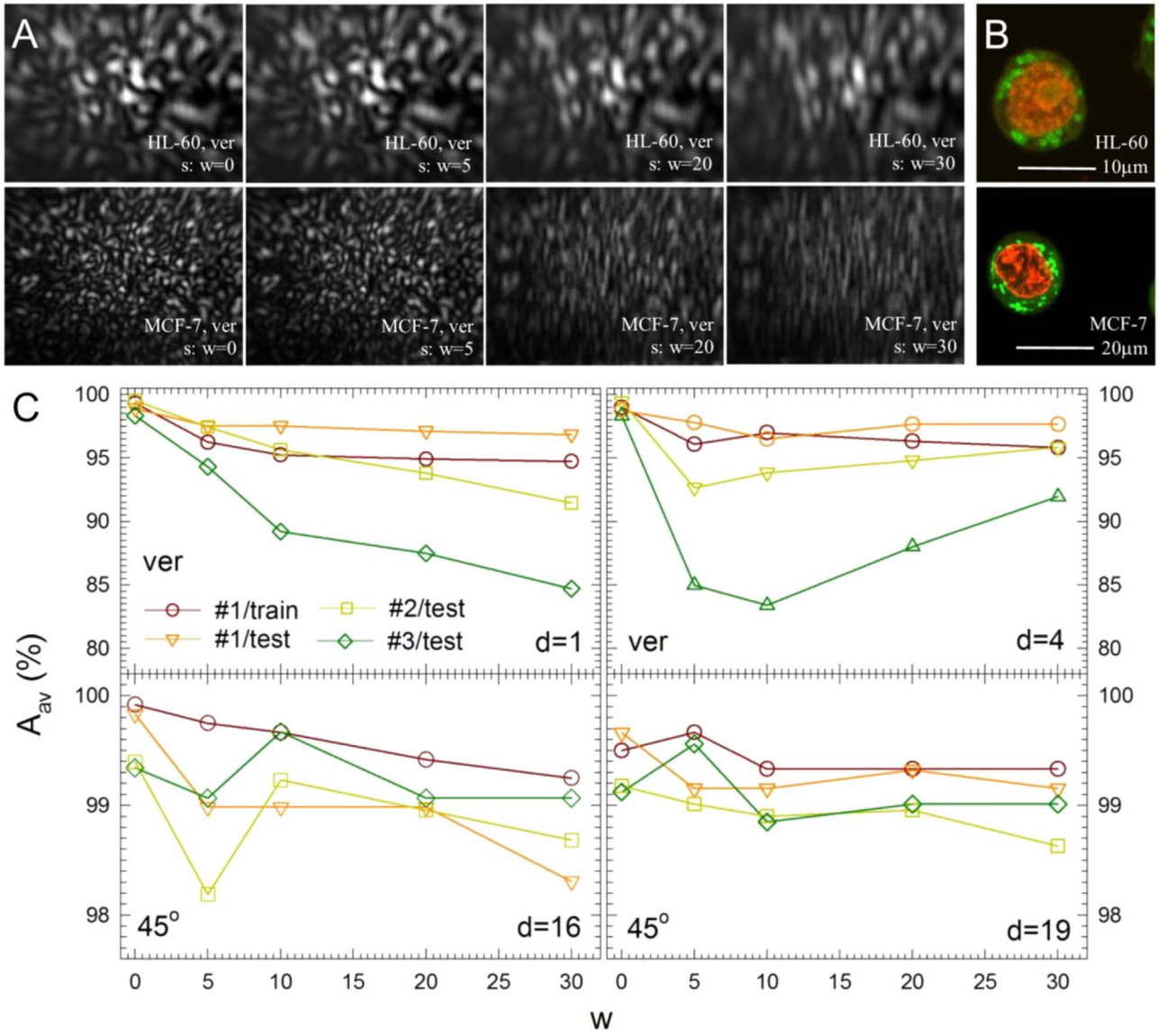

Fig. 7. (A) Two sets of acquired diffraction images "blurred" by window smoothing with different width *w*. Each image is labeled by the cell type, incident beam polarization, polarization of scattered light and *w*. (B) Confocal image slices of suspended cells double stained with Syto-61 for nucleus and MitoTracker-orange for mitochondria. The cell type and scale bar sizes are labeled. (C) The classification accuracy *A* of optimized SVM classifiers obtained with the training data set and three test data sets versus w with linear kernel functions. GLCM parameters were calculated at two values of pixel distance d with the incident beam polarization labeled on the left side and d labeled on the right side. The lines are for visual guide.

of investigation on the performance and consistency of SVM based classification using textural and intensity parameters extracted from the p-DI data with no or sufficiently small blur using one TDI camera. These data demonstrate significantly enhanced consistency of SVM classification with high performance with GLCM parameters calculated at variable distance between the paired pixels from the TDI acquired p-DI data.

GLCM employs a statistical approach for textural characterization of an input image which fits well to the analysis of p-DI data. Still, it is often difficult to relate the values and variations of GLCM parameters to the image patterns visible to human observers. By comparing the images in Fig. 3 to the plots in Fig. 4, one can gain insights on these parameters. As one of the top-performing GLCM parameters for classification, the mean values of dissimilarity (DIS) plotted in Fig. 4 present opposite differences between the two cell types in the s-polarized images acquired with vertical incident beam polarization and p-polarized images acquired with horizontal polarization. Since DIS is defined as the sum of the off-diagonal GLCM elements $p_{i,j}$ weighted by $|i–j|$ [23], it measures the total intensity variations among the pixels paired at d. Clearly, DIS of an image depends on d through two factors: (a) the intensity "dissimilarity" of $|i–j|$ between paired pixels; (b) the value of $p_{i,j}$ or the number of paired pixels with large intensity dissimilarities. For diffraction images consisting of bright or dark "spots" or speckles, it is obvious that either s-DIS or p-DIS should increase with d since the co-occurring probability of different intensities rises for pairs of increasing *d*, as confirmed by data in Fig. 4. A careful examination of the p-DI pairs in Fig. 3 further indicates that the MCF-7 images with sufficiently large average intensities exhibit fine speckles in comparison to those of HL-60 images. This fact can be used to understand that the values of s-DIS (ver) of MCF-7 images are larger than those of HL-60 images if we assume that DIS for small d values is dominated by factor (b). For the opposite differences between the two cell types in the case of p-DIS (hor) shown in Fig. 4, one has to assume the factor (a) becomes more important than (b), especially for large *d* values.

Based on the classification accuracy data presented in Fig. 7 and confusion matrices in Tables 2–4, one can observe that SVM based

classification of the two cell lines can be achieved with excellent and consistent performance among the data sets acquired in different days with the TDI camera. In addition, the high performance with A_{av} or A larger than 95% is possible for the three kernel functions of linear, polynomial and RBF with appropriately chosen d values, which demonstrate the potential to develop a robust data platform for rapid assay of single cells with the p-DIFC method. Compared to the methods reviewed in Section 2, the p-DI data yielded classification accuracies markedly higher than those by the fluorescent or other conventional approaches. Moreover, these results were statistically significant because of the large number of cells imaged for the three test data sets ranging from 3569 cells with the incident beam of hor polarization to 4942 cells with ver polarization.

It is interesting to note that the different patterns of change among the values of A_{av} or A of blurred p-DI data displayed in Fig. 7(C). While the case of (ver, $d=1$) shows expected decrease of accuracy and performance consistency for increasing blurring or w, the other three cases display mostly very small changes in A_{av} and A and somewhat reduced performance consistency of the optimized SVM classifiers. These may be explained by the fact that the two cell types of HL-60 and MCF-7 are of significantly difference in cell size and nucleus-to-cell ratio as demonstrated by the two confocal images of cells in Fig. 7(B). These differences are responsible for the different diffraction patterns in the p-DI data in Fig. 3 and thus the blurring with w up to 30 pixels is still not sufficiently large to affect classification accuracy in these cases.

6. Conclusion

A cell classification study on HL-60 and MCF-7 cells has been carried out to assess the performance of pattern recognition and machine learning algorithms on p-DI data acquired in different days with a new imaging configuration using one TDI camera. The results demonstrate that the significant reduction of blurring in p-DI data and choice of different pixel distances can significantly improve the consistency of best performing SVM classifiers. Additional studies should be pursued to cell types of similar origin and 3D morphology to further test and improve the positioning accuracy of the flowing cells, performance and consistency of machine learning algorithms. With further improvements on throughput rate and image processing speed with GPU computation, we expect that the p-DIFC method has the potential to be translated from a research novelty into a powerful tool for label-free and rapid cell assay.

Conflict of interest

None declared.

Acknowledgments

We wish to thank Dr. Fu Zheng of Tianjin Medical University for preparing the HL-60 and MCF-7 cell samples and Dr. Wenhuan Jiang for helps on the development of classification software. X.H. Hu acknowledges grant support by the Tianjin Science and Technology Commission and Y. Feng acknowledges grant supports from National Natural Science Foundation of China (#81041107 and #81171342).

References

[1] G.A. Losa, C. Castelli, Nuclear patterns of human breast cancer cells during apoptosis: characterisation by fractal dimension and co-occurrence matrix statistics, Cell Tissue Res. 322 (2005) 257–267.
[2] C.A.B. Castanon, J.S. Fraga, S. Fernandez, A. Gruber, L. da, F. Costa, Biological shape characterization for automatic image recognition and diagnosis of protozoan parasites of the genus Eimeria, Pattern Recogn. 40 (2007) 1899–1910.
[3] D. Altunbay, C. Cigir, C. Sokmensuer, C. Gunduz-Demir, Color graphs for automated cancer diagnosis and grading, IEEE Trans. Bio-Med. Eng. 57 (2010) 665–674.
[4] R. Nosaka, Y. Ohkawa, K. Fukui, Feature extraction based on co-occurrence of adjacent local binary patterns, in: Y.-S. Ho (Ed.), Advances in Image and Video Technology, Springer, Berlin, 2011, pp. 82–91.
[5] S. Ghosh, V. Chaudhary, Feature analysis for automatic classification of HEp-2 florescence patterns: computer-aided diagnosis of auto-immune diseases, in: International Conference on Pattern Recognition (ICPR), 2012.
[6] P. Foggia, G. Percannella, P. Soda, M. Vento, Benchmarking HEp-2 cells classification methods, IEEE Trans. Med. Imaging 32 (2013) 1878–1889.
[7] B.C. Lovell, G. Percannella, M. Vento, A. Wiliem, Performance evaluation of indirect immunofluorescence image analysis systems, in: International Conference on Pattern Recognition (ICPR), 2014.
[8] S. Manivannan, W. Li, S. Akbar, R. Wang, J. Zhang, S.J. McKenna, An automated pattern recognition system for classifying indirect immunofluorescence images of HEp-2 cells and specimens, Pattern Recogn. 51 (2016) 12–26.
[9] T.C. George, P.J. Morrissey, C. Cui, S. Singh, P. Fitzgerald Bocarsly, Measurement of Cytoplasmic to Nuclear Translocation, Current Protocols in Cytometry, John Wiley & Sons, Inc, New York 2009, p. 9.28.
[10] S. Holler, Y. Pan, R.K. Chang, J.R. Bottiger, S.C. Hill, D.B. Hillis, Two-dimensional angular optical scattering for the characterization of airborne microparticles, Opt. Lett. 23 (1998) 1489–1491.
[11] J. Neukammer, C. Gohlke, A. Hope, T. Wessel, H. Rinneberg, Angular distribution of light scattered by single biological cells and oriented particle agglomerates, Appl. Opt. 42 (2003) 6388–6397.
[12] J.Q. Lu, P. Yang, X.H. Hu, Simulations of light scattering from a biconcave red blood cell using the FDTD method, J. Biomed. Opt. 10 (2005) 024022.
[13] R.S. Brock, X.H. Hu, D.A. Weidner, J.R. Mourant, J.Q. Lu, Effect of detailed cell structure on light scattering distribution: FDTD study of a B-cell with 3D structure constructed from confocal images, J. Quant. Spectrosc. Radiat. Transf. 102 (2006) 25–36.
[14] K. Singh, X. Su, C. Liu, C. Capjack, W. Rozmus, C.J. Backhouse, A miniaturized wide-angle 2D cytometer, Cytom. A 69 (2006) 307–315.
[15] H. Ding, J.Q. Lu, R.S. Brock, T.J. McConnell, J.F. Ojeda, K.M. Jacobs, X.H. Hu, Angle-resolved Mueller matrix study of light scattering by B-cells at three wavelengths of 442, 633 and 850 nm, J. Biomed. Opt. 12 (2007) 034032.
[16] K.M. Jacobs, J.Q. Lu, X.H. Hu, Development of a diffraction imaging flow cytometer, Opt. Lett. 34 (2009) 2985–2987.
[17] K.M. Jacobs, L.V. Yang, J. Ding, A.E. Ekpenyong, R. Castellone, J.Q. Lu, X.H. Hu, Diffraction imaging of spheres and melanoma cells with a microscope objective, J. Biophotonics 2 (2009) 521–527.
[18] Y. Sa, Y. Feng, K.M. Jacobs, J. Yang, R. Pan, I. Gkigkitzis, J.Q. Lu, X.H. Hu, Study of low speed flow cytometry for diffraction imaging with different chamber and nozzle designs, Cytometry A 83 (2013) 1027–1033.
[19] J. Zhang, Y. Feng, M.S. Moran, J.Q. Lu, L.V. Yang, Y. Sa, N. Zhang, L. Dong, X.H. Hu, Analysis of cellular objects through diffraction images acquired by flow cytometry, Opt. Express 21 (2013) 24819–24828.
[20] Y. Feng, N. Zhang, K.M. Jacobs, W. Jiang, L.V. Yang, Z. Li, J. Zhang, J.Q. Lu, X.H. Hu, Polarization imaging and classification of Jurkat T and Ramos B cells using a flow cytometer, Cytometry A 85 (2014) 817–826.
[21] H. Wang, Y. Feng, Y. Sa, Y. Ma, R. Pan, J.Q. Lu, X.H. Hu, Acquisition of cross-polarized diffraction images and study of blurring effect by one time-delay-integration camera, Appl. Opt. 54 (2015) 5223–5228.
[22] W. Jiang, J.Q. Lu, L.V. Yang, Y. Sa, Y. Feng, J. Ding, X.H. Hu, Comparison study of distinguishing cancerous and normal prostate epithelial cells by confocal and polarization diffraction imaging, J. Biomed. Opt. 21 (2016) 071102.
[23] R.M. Haralick, Statistical and structural approaches to texture, Proc. IEEE 67 (1979) 786–804.
[24] R. Jobanputra, D.A. Clausi, Preserving boundaries for image texture segmentation using grey level co-occurring probabilities, Pattern Recogn. 39 (2006) 234–245.
[25] K. Dong, Y. Feng, K.M. Jacobs, J.Q. Lu, R.S. Brock, L.V. Yang, F.E. Bertrand, M.A. Farwell, X.H. Hu, Label-free classification of cultured cells through diffraction imaging, Biomed. Opt. Express 2 (2011) 1717–1726.
[26] 21th International Conference on Pattern Recognition, Contest on HEp-2 Cells Classification: Results, ⟨http://mivia.unisa.it/hep2contest/results.shtml⟩, 2016.
[27] R. Pan, Y. Feng, Y. Sa, J.Q. Lu, K.M. Jacobs, X.H. Hu, Analysis of diffraction imaging in non-conjugate configurations, Opt. Express 22 (2014) 31568–31574.
[28] X.H. Hu, 3D morphology and GLCM parameter definition tables, ⟨http://bmlaser.physics.ecu.edu/literature/3D_GLCM_Para_Def_Tables.pdf⟩, 2015.
[29] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.
[30] C.J.C. Burges, A tutorial on support vector machines for pattern recognition, Data Min. Knowl. Disc 2 (1998) 121–167.
[31] C.C. Chang, C.J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011) 1–27.
[32] H. Wang, C. Jin, Y. Feng, D. Qi, Y. Sa, X.H. Hu, Quantitative assessment of image motion blur in diffraction images of moving biological cells, Opt. Eng. 55 (2016) 023103.

He Wang received his BS degree from the Shandong University in 2010 and PhD degree from the Tianjin University in 2016, both are in biomedical engineering. His current research interests include biological cell imaging and development of TDI cameras.

Yuanming Feng is a professor of biomedical engineering at the Tianjin University. His research interests include diffraction imaging flow cytometry and measurement technique of radiation-induced apoptosis.

Yu Sa received his BS and Ph.D degrees from Tianjin University and he is a lecturer at Department of Biomedical Engineering of Tianjin University. His research interests include instrument development, diffraction imaging, and computational fluid dynamics modeling studies.

Jun Qing Lu received her BS/MS degrees in physics from Nankai University and Ph.D degree from University of California, Irvine. She is currently an associate professor at East Carolina University and mainly interested in numerical studies of light scattering in turbid media and with biological cells.

Junhua Ding is an associate professor of computer science with East Carolina University. He received his BS, MS and Ph.D, all in computer science, in 1994, 1997, and 2004, respectively. His research interests are in software engineering and data engineering, and published over 60 peer reviewed papers in these fields.

Jun Zhang received his BS degree in 2010 and PhD degree in 2016 from the Tianjin University, both are in biomedical engineering. His current research interests are in image analysis and medical physics.

Xin-Hua Hu received his BS and MS degrees from Nankai University, Tianjin, China, in 1982 and 1985, an MS degree in physics from Indiana University in 1986, and PhD degree in physics in 1991 from the University of California at Irvine. He joined the physics faculty in 1995 and is currently a professor at East Carolina University. His main research interests relate to the investigations of light scattering and their applications in probing tissues and cells.