Photonics West 20–25 January 2007 San Jose, California USA

Q:\Manuscripts\2007\SPIE\6441-22.pdf

Automated classification and recognition of bacterial particles in flow by multi-angle scatter measurement and a support-vector machine classifier

Bartek Rajwa[a*], Murugesan Venkatapathi[a,b], Kathy Ragheb[a], Padmapriya P. Banada[c], E. Daniel Hirleman[b], Todd Lary[d], and J. Paul Robinson[a]

[a]Purdue University Cytometry Laboratories, Bindley Bioscience Center, 1203 W. State Street; [b]School of Mechanical Engineering, 585 Purdue Mall; [c]Molecular Food Microbiology Laboratory, Department of Food Science, 745 Agriculture Mall Drive, Purdue University, West Lafayette, IN 47907, USA

[d]Cellular Analysis Technology Center, Beckman Coulter, Inc., 11800 S.W. 147th Avenue, Miami, FL 33196, USA

ABSTRACT

Biological microparticles scatter light in all directions when illuminated. The complex scatter pattern is dependent on particle size, shape, refraction index, density, and morphology. Commercial flow cytometers allow measurement at two nominal angles ($2° \leq \theta_1 \leq 20°$ and $70° \leq \theta_2 \leq 110°$) of scattered light intensity from individual microparticles with a speed varying from 10 to 10000 particles per second. The choice of angle is dictated by the fact that scattered light in the small-angle region is primarily influenced by cell size and refractive index, whereas side scatter intensity is related to the granularity of cellular structures. Obviously, these rudimentary measurements cannot be used to separate populations of cells of similar shape, size, or structure. Hence, there have been several attempts in flow cytometry to measure the entire scatter patterns. However, the published concepts required use of unique custom-built flow cytometers and could not be applied to existing instruments. It was also not clear how much information about patterns is really necessary to separate various populations of cells present in a given sample. The presented work demonstrates application of pattern-recognition techniques to classify particles on the basis of their discrete scatter patterns collected at just five different angles, and accompanied by the measurement of axial light loss. Our approach can be used with existing instruments and requires only the addition of a compact custom-built scatter-detector. Our analytical model of scatter of laser beams by individual bacterial cells suspended in a fluid was used to determine the location for scatter sensors. Experimental results were used to train the SVM-based pattern recognition system. It has been shown that information provided just by six scatter-related parameters was sufficient to recognize various bacteria with 90-99% success rate.

Keywords: flow cytometry, light scattering, machine learning, classification, bacteria.

1. INTRODUCTION

It has been demonstrated that information about forward scatter ($2° \leq \theta_1 \leq 20°$) and side scatter ($70° \leq \theta_2 \leq 110°$) can be used to identify a number of cell subpopulations without the use of any additional information provided by fluorescence stains[1-3]. This was possible owing to the fact that forward-scattered light in the small-angle region ($\theta \leq 2°$) is primarily dependent on the cell size, and is mostly independent of its refractive index or shape[4-6], whereas perpendicular light scatter is sensitive to small internal structures in cells and to refractive index changes.

Researchers agree that full scatter patterns of bioparticles may contain much more information than forward scatter, perpendicular scatter, and extinction signals can reveal. Indeed, several papers published in the '70s and early '80s on complex applications of multi-angle scatter detection in flow involved measurement of the full 180° or 360° scatter patterns from single biological cells[1,7-9]. However, none of the early designs of multi-angle scatter detectors found its

[*]brajwa@purdue.edu; Purdue University Cytometry Laboratories, Bindley Bioscience Building, 1203 W. State Street. West Lafayette, IN 47907-2057; tel. (765) 494 0757

way to commercial systems. Currently only a few research groups still actively investigate applications of multi-angle light-scatter analysis in flow[10-14].

The presented work demonstrates the application of pattern-recognition techniques to classify microbial particles on the basis of their scatter patterns accompanied by the measurement of axial light loss. Our approach differs from previous reports by the use of an analytical model of scatter of laser beams by individual bacterial cells suspended in a fluid to determine the optimal location for scatter sensors. In contrast to other reports which described direct use of collected scatter signals to characterize bioparticles, our method works in concert with a machine-learning system. The experimental results obtained from the known samples are used to train the support-vector machine, and subsequently samples containing mixtures of unknown particles are classified by the trained algorithm. It has been shown that information provided by just six scatter-related parameters was sufficient for a trained system to recognize various bacteria with a success rate above 90%.

2. MATERIALS AND METHODS

2.1. Bacterial cultures

Four different non-pathogenic bacterial cultures of varying size and shape were selected for the experiments: *Escherichia coli* K12, *Listeria innocua* F4248, *Bacillus subtili*s ATCC 6633, and *Enterococcus faecalis* CG110. The cultures were grown in brain heart infusion (BHI) broth for 16-18h at 37°C, 140 rpm in a shaker incubator. The cultures were washed once and resuspended in sterile phosphate buffered saline (PBS), pH 7.6, before analysis.

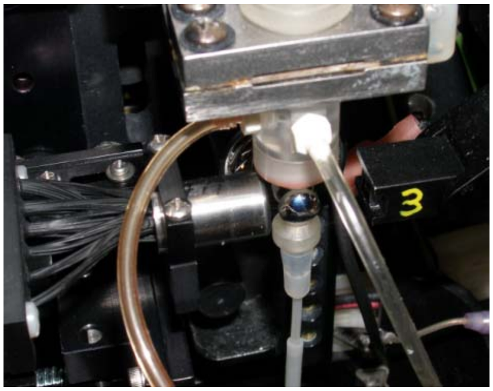

Figure 1. The scatter measurement system consists of four ring detectors and an axial light-loss detector that can be moved towards or away from the laser beam–particle intersection point to change the angles of measurement.

2.2. Analytical model of scatter

The mathematical model of scatter used in this work assumes that the particles (refractive index n=1.44) are in isolation in the sheath fluid (n=1.33), and the angular scatter distribution is calculated and integrated over the area of the forward-scatter detector placed outside the sheath fluid. This assumption is valid if the particles are much smaller than the channel and if the laser beam (10µm x 80µm Gaussian) inside the channel is considerably larger than the particle. The bacterial cells are modeled as homogeneous particles with effective refractive indices using the Discrete Dipole Approximation (DDA).

The discrete-dipole approximation method (DDA) was first formulated by Purcell and Pennypacker, who used it to study interstellar dust grains, and later studied by other researchers such as Draine, and Taubenblatt and Tran[15-17]. In DDA an arbitrarily shaped particle is treated as a three-dimensional assembly of dipoles (j = 1,....,N) on a cubic grid, located at positions $\mathbf{r}_j$. Each dipole is assigned a complex polarisability α_i which can be computed from the complex refractive index of the bulk material and the number of dipoles in a unit volume[18]. The dipole moment or polarization at each dipole is related to the electric field by $\mathbf{P}_j=\alpha_j\mathbf{E}_{to,j}$, where $\mathbf{P}_j$ is the dipole moment at the dipole 'j', $\mathbf{E}_{tot,j}$ is the total electric field at dipole j, at $\mathbf{r}_j$.

Following the notation of Draine[19], the field $\mathbf{E}_{tot,j}$, at each dipole can be decomposed into the electrical field incident upon the features and the electric field contribution from the other interacting dipoles. Hence, the electric field can then be represented as $\mathbf{E}_{tot,j}$=$\mathbf{E}_{inc,j}$+$\mathbf{E}_{dipole,j}$, where $\mathbf{E}_{dipole,j}$ is the electric field contribution from the other N-1 dipoles, and $\mathbf{E}_{inc,j}$ is the known incident field $\mathbf{E}_0\exp(i\mathbf{k}\cdot\mathbf{r}_i\text{-}i\omega t)$. Therefore $\mathbf{E}_{dipole,j}$ can be expressed as:

$$\mathbf{E}_{dipole,j} = \mathbf{E}_{inc,j} - \sum_{k\neq j} \mathbf{A}_{jk}\mathbf{P}_k \,,$$

where -$\mathbf{A}_{jk}\mathbf{P}_k$ is the electric field at $\mathbf{r}_j$ due to dipole $\mathbf{P}_k$. Each element $\mathbf{A}_{jk}$ is a 3x3 matrix:

$$\mathbf{A}_{jk} = \frac{\exp(ikr_{jk})}{r_{jk}} \times \left[k^2(\hat{r}_{jk}\hat{r}_{jk} - \mathbf{1}_3) + \frac{ikr_{jk}-1}{r_{jk}^2}(3\hat{r}_{jk}\hat{r}_{jk} - \mathbf{1}_3) \right], \quad j \neq k,$$

where $k \equiv \omega / c$, $r_{jk} = \left|\mathbf{r}_j - \mathbf{r}_k\right|$, $\hat{r}_{jk} \equiv (\mathbf{r}_j - \mathbf{r}_k)/r_{jk}$, and $\mathbf{1}_3$ is a 3x3 identity matrix. With $\mathbf{A}_{jj} \equiv \alpha_j^{-1}$ the scattering problem is reduced to finding polarizations $\mathbf{P}_j$ that satisfy a system of equations:

$$\sum_{k=1}^{N} \mathbf{A}_{jk}\mathbf{P}_k = \mathbf{E}_{inc,j} \,.$$

These equations can be solved by iterations. By introducing a Green function, the method produces reliable results for extremely rough discretization grids such as 2.22 meshes per wavelength[20]. In the presented study the Quasi-Minimal Residual (QMR) method has been used to solve the problem. Owing to the characteristics of the coefficient matrix, the convergence towards an accurate answer is dependent on scattering feature size and refractive index.

The longer axes of the bacteria are assumed to be aligned with the axis of flow owing to the hydrodynamic forces in a flow cytometer. Since polarization changes the scattering cross section noticeably, especially for long rod-like particles, the model employed takes into account incident laser beam polarization. The proper placing of the detectors was determined by the value of the distinguishability factor D, defined as the ratio of the difference in scattering cross section to the sum of scattering cross section of two different bacteria $D = \sum_{\theta_i} \left[(dCsc_i - dCsc_j) \big/ (dCsc_i + dCsc_j) \right]$, where i,j represent different bacterial species.

In order to predict the feasible classification success for the automated analysis system, the variability in scatter signal resulting from differences in refractive index and sizes of individual particles was considered in an extended scatter model. A normal distribution of size and refractive index was employed in the enhanced model for three species of bacteria (*L. innocua*, *E. faecalis* and, *E. coli*). The 1/e width of the normal distribution of refractive index was 0.033 with a mean of 1.44. The standard deviation for volume of the bacteria was assumed to be 5%.

The overlap of the resultant modeled populations calculated for every pair of bacterial species in the measurement space was used as an estimate of the feasible classification success, for every two-class case (Table 1).

2.3. Flow cytometry

All the analyses were performed with a Cytomics FC500 flow cytometer (Beckman Coulter, Miami, FL) equipped with a 488-nm air-cooled argon laser. The traditional scatter detector in the instrument was replaced by an enhanced detection system (Beckman Coulter) capable of measuring forward-scatter signals at four different angles (Figure 1). The scatter-measurement system consists of four ring detectors and an axial light loss detector that can be moved towards or away from the laser beam–particle intersection point to change the angles of measurement.

2.4. Classification using Support Vector Machines

Support vector machine (SVM) algorithms allow for non-linear decision boundaries in the input space. SVMs are based on the concept of decision hyperplanes that define decision boundaries. A decision hyperplane is one that separates a set of objects having different class memberships. SVMs are able to construct hyperplanes in a multidimensional space that separates cases of different class labels. An optimal decision hyperplane is here defined as the linear decision function with maximal margin between the vectors of the two classes. It has been demonstrated that to construct such hyperplanes one has to take into account only a small amount of the training data, the so-called support vectors, which determine this margin[21]. For $\mathbf{w}_0 \cdot \mathbf{z} + b_0 = 0 \;\; | \;\; \mathbf{w} \in R^N, b \in R$. which the optimal hyperplane, it has been shown that the weights $\mathbf{w}_0$

can be expressed as linear combination of support vectors is $\mathbf{w}_0 = \sum_{\substack{\text{support}\\ \text{vectors}}} \alpha_i \mathbf{z}_i$. Therefore, the linear decision function I(**z**) will be in the form of $I(\mathbf{z}) = sign\left(\sum_{\substack{\text{support}\\ \text{vectors}}} \alpha_i \mathbf{z}_i \cdot \mathbf{z} + b_0 \right)$, where $\mathbf{z}_i \cdot \mathbf{z}$ is the dot product between support vectors $\mathbf{z}_i$ and vector **z** in feature space. SVM is a linear classifier in the parameter space, but it is easily extended to a nonlinear classifier by mapping the space $S=\{\mathbf{x}\}$ of the input data into a high-dimensional (possibly infinite-dimensional) feature space $F=\{\varphi(\mathbf{x})\}$. If one chooses an adequate mapping φ, the data points become linearly separable or mostly linearly separable in the high-dimensional space, so that one can easily apply the structure risk minimization[22]. To avoid working in the potentially high-dimensional space F, one tries to pick a feature space in which the dot product can be evaluated directly using a nonlinear function in input space, i.e. by means of the kernel trick: $\kappa(\mathbf{x}_1,\mathbf{x}_2) = \langle\varphi(\mathbf{x}_1), \varphi(\mathbf{x}_2)\rangle$. Therefore, instead of making a non-linear transformation of the input vectors followed by dot-products with support vectors in feature space, one can first compare two vectors in input space, and then make a non-linear transformation of the value of the result[23]. A kernel can be also understood as a similarity measure between two observations. A large value for $\kappa(\mathbf{x}_1,\mathbf{x}_2)$ indicates similar points, where smaller values indicate dissimilar points. Typical kernels include the linear kernel, $\kappa(\mathbf{x}_1,\mathbf{x}_2) = \mathbf{x}_1^T\mathbf{x}_2$, the polynomial kernel, $\kappa(\mathbf{x}_1,\mathbf{x}_2) = (\mathbf{x}_1^T\mathbf{x}_2 + 1)^d$, or the RBF kernel, $\kappa(\mathbf{x}_1, \mathbf{x}_2) = exp(-\gamma\|\mathbf{x}_1-\mathbf{x}_2\|^2)$. It has been shown that all these kernels are functions of dot products[24].

Supervised classification was performed in this report using an implementation of a multi-class support vector machine based algorithm by Chih-Chung Chang and Chih-Jen Lin[25,26].

3. RESULTS AND DISCUSSION

Measurements of forward and side scatter did not allow distinguishing between the microbial particles of *Escherichia coli* K12, *Listeria innocua* F4248, *Bacillus subtilis* ATCC 6633, and *Enterococcus faecalis* CG110. Figure 2 represents example of one of the experiments in which two samples containing *E. faecalis* and *B. subtilis* were analyzed using an unmodified commercial flow cytometer. The results have been scatter-plotted to demonstrate that the two populations completely overlap in the measurement space.

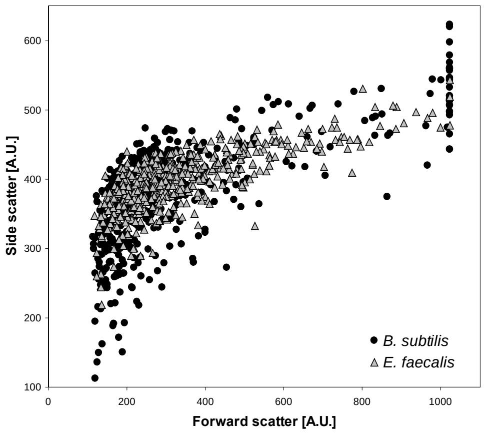

Figure 2. Scatterplot representing *E. faecalis* (circles) and *B. subtilis* (triangles) analyzed using an unmodified commercial flow cytometer. The results demonstrate that the two populations completely overlap in the measurement space. 1000 event have been plotted.

An optimal location of the multi-angle detector has been determined using a computer model of light scatter based on DDA. Table 1 shows computed distinguishabilities for pairs of every two of the analyzed bacteria at the two sets of angles 7.8°, 11.3°, 17°, and 22.5° (Set I), and 4°, 5.8°, 8.7°, and 11.5° (Set II). These two sets of angle represent two positions of the multi-angle detector. As Figure 3 demonstrates, Set I (represented by light-gray lines) offered better distinguishability than Set II (dark-grey lines) for most of the analyzed two-component mixtures of bacteria. Table 2 shows theoretically estimated classification rates for the designed system using angle Set I.

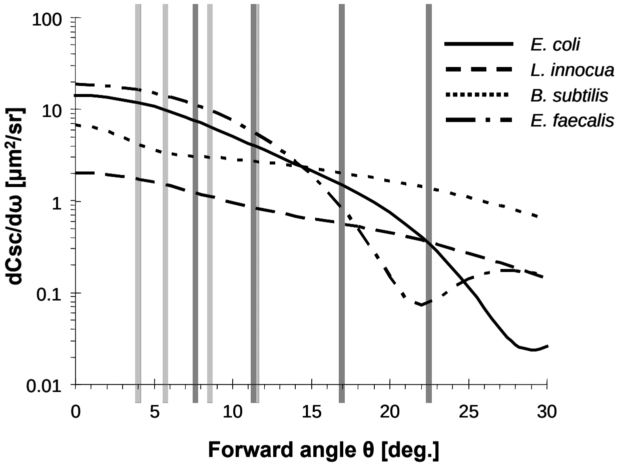

Figure 3. The nominal variation of average differential scattering cross section (dCsc/dω) with forward angle for the four bacteria species.

	E. coli		*L. innocua*		*B. subtilis*		*E. faecalis*	
	I	II	I	II	I	II	I	II
E. coli			1.88	2.82	2.15	1.49	1.34	0.67
L. innocua	1.88	2.82			2.11	2.14	2.42	3.12
B. subtilis	2.15	1.49	2.11	2.14			2.25	2.01
E. faecalis	1.34	0.67	2.42	3.12	2.25	2.01		

Table 1. Distinguishabilities calculated for scatter measurement at nominal 7.8°, 11.3°, 17°, and 22.5° (Set I) and 4°, 5.8°, 8.7°, and 11.5° (Set II).

Collected scatter signals at the four forward angles, side scatter, and axial light-loss measure for each particle from every group of bacteria formed multidimensional data vectors describing all the bioparticles. The classification problem was then to determine the species of analyzed particles on the basis of these data vectors. Plotting the results of all the measured angles of scatter and axial light loss, as well as visualization of principal components, did not allow for clean manual gating to separate the subpopulations (data not shown).

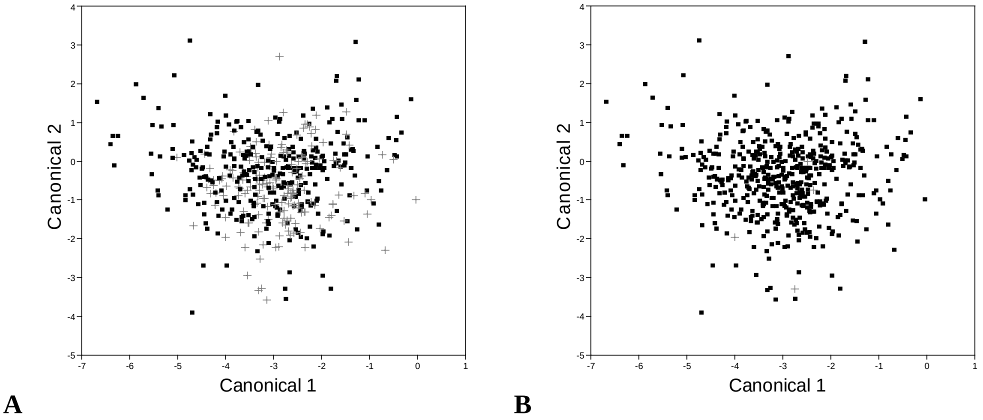

Figure 4. Canonical plot showing an example of LDA (A) and SVM (B) automated classifications of *B. subtilis* and *E. faecalis*. 1000 events have been plotted. Legend: black "■" – correctly classified events, gray "+" – missclassfied events.

Automated classification was performed by two supervised methods: linear discriminant analysis (LDA) and support-vector machine algorithm (SVM) (See Figure 4 A, B). The 5x2 cross-validation and bootstrap algorithms were used to determine the classification success. Table 2 summarizes the results.

	E. coli		*L. innocua*		*B. subtilis*		*E. faecalis*	
	R	***E***	***R***	***E***	***R***	***E***	***R***	***E***
E. coli	-	-	86.30%	95.8%	99.10%	-	68.70%	77.1%
L. innocua	86.30%	95.8%	-	-	99.60%	-	81.60%	95.6%
B. subtilis	99.10%	-	99.60%	-	-	-	98.50%	-
E. faecalis	68.70%	77.1%	81.60%	95.6%	98.50%	-	-	-

Table 2. Average real (R), and estimated (E) classification success rates for six-parameter (7.8°, 11.3°, 17.7°, 22.5°, 90°, and axial light loss) scatter system employing SVM classifier.

4. CONLUSIONS

The results obtained in the presented study clearly demonstrate that scatter signals alone can be successfully used to automatically recognize various species of bacteria, provided that a sufficient training sample is available. The data showed also that whereas manual gating using multi-parameter scatter plots or utilizing a linear classifier did not provide satisfactory results, the SVM-based classification produced a remarkably high success rate (See Figure 4 and Table 2).

Comparison of the classification success with the estimated distinguishabilities shows a high level of agreement except for one of the classification cases for each set of angles. This is encouraging, considering the fact that the model does not take into account the intra-population variance in size and refractive indices. The classification success predicted by the extended model has not been reached, suggesting that instrument noise seriously limits the feasible classification rates. Despite that, the obtained rates demonstrate that label-free classification of microorganisms is indeed feasible, even with a system collecting a very limited number of scatter parameters. This, in turn, suggests that simple multi-angle scatter detectors may find a niche in systems where fluorescent labels cannot be use, e.g. on-line flow cytometry monitoring of bioreactors.

REFERENCES

1. G.C. Salzman, J.M. Crowell, J.C. Martin, P.M. Labauve and P.F. Mullaney, "Classification of Human Leukocytes by Multiangle Laser Light-Scattering in A Flow System," *Biophysical Journal* 15(2), A240 (1975).
2. J.W.M. Visser, G.J.V.D. Engh and D.W.V. Bekkum, "Light-Scattering Properties of Murine Hematopoietic-Cells," *Blood Cells* 6(3), 391-407 (1980).
3. G.B.J. Dubelaar, J.W.M. Visser and M. Donze, "Anomalous Behavior of Forward and Perpendicular Light-Scattering of A Cyanobacterium Owing to Intracellular Gas Vacuoles," *Cytometry* 8(4), 405-412 (1987).
4. P.F. Mullaney and P.N. Dean, "Cell Sizing - A Small-Angle Light-Scattering Method for Sizing Particles of Low Relative Refractive Index," *Applied Optics* 8(11), 2361-& (1969).
5. P.F. Mullaney, M.A. Vandilla, J.R. Coulter and P.N. Dean, "Cell Sizing - A Light Scattering Photometer for Rapid Volume Determination," *Review of Scientific Instruments* 40(8), 1029 (1969).
6. M. Kerker, H. Chew, P.J. Mcnulty, J.P. Kratohvil, D.D. Cooke, M. Sculley and M.P. Lee, "Light-Scattering and Fluorescence by Small Particles Having Internal Structure," *Journal of Histochemistry & Cytochemistry* 27(1), 250-263 (1979).
7. G.C. Salzman, J.M. Crowell and P.F. Mullaney, "Flow-System Multi-Angle Light-Scattering Instrument for Biological Cell Characterization," *Journal of the Optical Society of America* 65(10), 1170-1171 (1975).
8. M.R. Loken, R.G. Sweet and L.A. Herzenberg, "Cell Discrimination by Multiangle Light-Scattering," *Journal of Histochemistry & Cytochemistry* 24(1), 284-291 (1976).
9. M. Bartholdi, G.C. Salzman, R.D. Hiebert and M. Kerker, "Differential Light-Scattering Photometer for Rapid Analysis of Single Particles in Flow," *Applied Optics* 19(10), 1573-1581 (1980).

10. A.N. Shvalov, I.V. Surovtsev, A.V. Chernyshev, J.T. Soini and V.P. Maltsev, "Particle classification from light scattering with the scanning flow cytometer," *Cytometry* 37(3), 215-220 (1999).
11. A.N. Shvalov, J.T. Soini, I.V. Surovtsev, G.V. Kochneva, G.F. Sivolobova, A.K. Petrov and V.P. Maltsev, "Individual Escherichia coli cells studied from light scattering with the scanning flow cytometer," *Cytometry* 41(1), 41-45 (2000).
12. M.A. Yurkin, K.A. Semyanov, P.A. Tarasov, A.V. Chernyshev, A.G. Hoekstra and V.P. Maltsev, "Experimental and theoretical study of light scattering by individual mature red blood cells by use of scanning flow cytometry and a discrete dipole approximation," *Appl. Opt.* 44(25), 5249-5256 (2005).
13. H.B. Steen, "Flow cytometer for measurement of the light scattering of viral and other submicroscopic particles," *Cytometry Part A* 57A(2), 94-99 (2004).
14. V.P. Maltsev, "Scanning flow cytometry for individual particle analysis," *Review of Scientific Instruments* 71(1), 243-255 (2000).
15. E.M. Purcell and C.R. Pennypacker, "Scattering and Absorption of Light by Nonspherical Dielectric Grains," *Astrophysical Journal* 186(2), 705-714 (1973).
16. B.T. Draine, "The Discrete-Dipole Approximation and Its Application to Interstellar Graphite Grains," *Astrophysical Journal* 333(2), 848-872 (1988).
17. M.A. Taubenblatt and T.K. Tran, "Calculation of Light-Scattering from Particles and Structures on A Surface by the Coupled-Dipole Method," *Journal of the Optical Society of America A-Optics Image Science and Vision* 10(5), 912-919 (1993).
18. T. Wriedt, "A review of elastic light scattering theories," *Particle & Particle Systems Characterization* 15(2), 67-74 (1998).
19. B.T. Draine and P.J. Flatau, "Discrete-Dipole Approximation for Scattering Calculations," *Journal of the Optical Society of America A-Optics Image Science and Vision* 11(4), 1491-1499 (1994).
20. T. Wriedt, "A review of elastic light scattering theories," *Particle & Particle Systems Characterization* 15(2), 67-74 (1998).
21. C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning* 20(3), 273-297 (1995).
22. S. Amari and S. Wu, "Improving support vector machine classifiers by modifying kernel functions," *Neural Networks* 12(6), 783-789 (1999).
23. C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning* 20(3), 273-297 (1995).
24. B. Scholkopf and A.J. Smola, "A short introduction to learning with kernels," *Advanced Lectures on Machine Learning* 2600 41-64 (2002).
25. Chih-Chung Chang ,Chih-Jen Lin. "LIBSVM: a library for support vector machines," http://www.csie.ntu.edu.tw/~cjlin/libsvm. 2001.
26. C.J.C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery* 2(2), 121-167 (1998).