Toxicology Research

A new, multi-disciplinary journal covering the best research in both fundamental and applied aspects of toxicology
Published in partnership with the British Toxicology Society and Chinese Society of Toxicology

The journal will be led by Editor-in-Chief, Nigel Gooderham, professor of molecular toxicology at Imperial College London and will provide a home to communications, full papers and reviews.

Authors will benefit from wide exposure for their work with free online access to content published during 2012 and 2013. Published work can also be made more visible through integration with ChemSpider, the RSC’s free chemical structure database.

Authors can expect a speedy peer review and publication process, with the option of publishing research as an Accepted Manuscript, making work available quickly after acceptance.

The first issue is scheduled for publication around May 2012.

Submit your work now!

www.rsc.org/toxicology

New Journal

Analytical Methods

www.rsc.org/methods

Analytical Methods

PAPER
Attinger et al. Identification of fluid and substrate chemistry based on automatic pattern recognition of stains

Volume 4 | Number 1 | January 2012 | Pages 1–314
Identification of fluid and substrate chemistry based on automatic pattern recognition of stains†

Namwon Kim, a Zhenguo Li, b Cedric Hurth, c Frederic Zenhausern, c Shih-Fu Chang b and Daniel Attinger a,d

Received 9th June 2011, Accepted 19th September 2011
DOI: 10.1039/c1ay05338h

This study proposes that images of stains from 100-nanolitre drops can be automatically identified as signatures of fluid composition and substrate chemistry, for e.g. rapid biological testing. Two datasets of stain images are produced and made available online, one with consumable fluids, and the other with biological fluids. Classification algorithms are used to identify an unknown stain by measuring its similarity to representative examples of predefined categories. The accuracy ranges from 80 to 94%, compared to an accuracy by random assignment of 3 to 4%. Clustering algorithms are also applied to group unknown stain images into a number of clusters each likely to correspond to similar combinations of fluids and substrates. The clustering accuracy ranges from 62 to 80%, compared to an accuracy by random assignment of 3 or 4%. The algorithms were also remarkably accurate at determining the presence or absence of biotin and streptavidin respectively in the liquid and on the glass, the salt composition, or the pH of the solution.

Introduction

Pattern recognition broadens its application in various fields by substituting routine work requiring human senses such as sight, hearing, taste, touch and smell. Specifically developed pattern recognition and machine vision techniques are widely deployed in manufacturing assembly lines for automatic defect detection, in public safety and forensics for recognizing individuals on the basis of fingerprints,1 face,2 voice1 and handwriting.4 Other applications include food analysis5–8 and environmental monitoring9 for identifying physical and chemical data10 using sensor arrays such as electronic noses and tongue.

When a drop of a complex fluid dries on a solid substrate, it leaves a stain, which is a complex signature of the drying conditions, of the morphology and chemistry of the substrate and of the composition of the fluid. The study of formation of patterns during the drying of small liquid droplets is of interest to biotechnology11–16 and materials science.17–19 Recent studies on the self-assembly of colloidal solutions of nanoparticles20,21 explained how some typical stains assume typical shapes such as a peripheral ring, a central bump or a uniform deposit. The structure of the stain was shown to be determined by the relative role of three convective transport phenomena involved in the deposition of the nanoparticles on the solid surface, i.e. the normal flow caused by electrostatic and Van der Waals forces, the radial flow caused by the maximum evaporation rate at the contact line,22 and the Marangoni recirculation caused by surface tension gradient at the air-liquid interface.23 The staining process can be parallelized and occurs within a few seconds because of Marangoni convection coupled with receding of the initial wetting line24 for sub-microlitre drops. A wealth of complex patterns appears in a reproducible manner in stains of fluids and substrates of controlled chemistry, showing multiple length scales, specific periodicity and features like lines, rings, crystals, and various grain sizes.25 Manual inspection of these features has been shown to provide information on the liquid and substrate chemistry.16,24 However, manual identification is tedious and expensive, especially when a lot of images are required to be identified. More critically, the results could be unreliable because of large and unpredictable variations of human factors. Automated identification such as pattern recognition can play a significant role in crime scene analysis and disease diagnosis, which are currently conducted manually by human experts. In contrast to manual identification, the pattern recognition approach is fast, purely data-driven, and not subject to human bias.25

The study presented here tests two hypotheses: (1) algorithms trained with existing data linking stain morphology to liquid and substrate chemistry can identify the liquid and substrate chemistry of an unknown stain; and (2) automated pattern recognition methods can group stains in a way that discriminates between specific combinations of liquid and substrate chemistry. The principle of the pattern recognition algorithms used in this
study is as follows: they first perform automatic localization (or cropping) of the stain as a region of interest in the image, then extraction and computation of descriptive features, that are expressed quantitatively as a row vector \( f = [a, \beta, \gamma, \epsilon] \). This vector is assembled from four discriminative feature vectors, with relative weights expressed by \( a, \beta, \gamma, \) and \( \epsilon \). These feature vectors characterize the color distribution as \( f_c \), the local binary patterns (LBP) \( f_l \), the Gabor wavelet \( f_g \), and the relative stain size as \( f_s \) (see more details in the Methods section).

Finally, two specific machine learning techniques are applied, a classification technique to test hypothesis (1), and a clustering technique to test hypothesis (2). Classification is performed by supervised pattern recognition algorithms that compare stain patterns in question with a set of training patterns of which the liquid/substrate category is known, computing the distance between the vector \( f \) of an unknown stain with vector representing labeled stains, and assigning the unknown stain to the class of labeled stains that minimize that distance. Clustering uses unsupervised pattern recognition algorithms to identify patterns common to several images in the stain dataset and group similar stains into a manually determined number of clusters (or classes) without labeled data, such that each cluster corresponds to a distinct visual pattern. Two comprehensive collections of microscopic stain images were constructed from 100 nL drops of consumable and biological fluids to test pattern recognition algorithms. The first dataset contains 480 stain images from 24 consumable fluids (24 classes) such as beer, juice, liquor, milk, red wine, and soda, all deposited on clean glass slides. For each fluid, 20 stains were produced. The second dataset includes 600 stain images from 8 biological fluids (8 classes) such as 10 mM phosphate buffer made from three different ratios of \( K_2HPO_4 \) and \( KH_2PO_4 \), 20 mM phosphate buffer with pH values obtained from three different concentrations of citric acid, a solution of 0.01 mM potassium hydroxide and pure water. All fluids were deposited on two different types of glass slides, clean or coated with streptavidin, which resulted in 16 classes as described in the Methods section. The number of classes was doubled again (32 classes) by adding biotin to each solution. However, pure water on both clean and streptavidin glasses (2 classes) did not generate any stains, thus 30 classes were established for the biological fluids. For both datasets, the drying conditions and the substrate morphology were kept constant. For the sake of comparison with other pattern recognition methods, the two datasets of original stain pictures have been made available online.

**Methods**

**Fluid preparation**

Consumable fluids including beers (Budweiser® Lager, Corona® Extra, Guinness® Extra Stout, Heineken® Lager, and Tsingtao® Lager), juices (Tropicana® Grape, Tropicana® Lemonade, Tropicana® Orange No Pulp, and Campbell’s® Tomato), liquor (Disaronno® Originale Amaretto), milks (Horizon® 1% Low Fat, Horizon® 2% Reduced Fat, Horizon® Chocolate, Horizon® Fat Free, Silk® Soy, Horizon® Strawberry, and Horizon® Whole), red wines (2007 Chateau de Castelneau from Bordeaux, France, Merlot, 2008 Liberty School from Paso Robles, CA, USA, Syrah, and 2008 Graham Beck from Franschhoek, South Africa, Cabernet Sauvignon), and sodas (Coca Cola® Classic, Diet Coke®, Dr Pepper®, and Fanta® Orange) were purchased from local stores. The fluids were used no later than 2 hours after opening of the original containers. For the biological fluids, a series of aqueous buffer solutions were prepared using biotechnology performance certified-grade water (Sigma-Aldrich W3513, Saint-Louis, MO). A first subset consists of phosphate buffer (PB) solutions prepared to a final concentration of 10 mM using three different volume ratios of \( K_2HPO_4 \) and \( KH_2PO_4 \) (Sigma-Aldrich, Saint-Louis, MO) calculated according to the Henderson-Hasselbach formula to yield respective pH values of 6.03, 7.05, and 7.98 as checked by a pH meter (Acorn ph 6, Oakton Instruments, Vernon Hills, IL) calibrated with NIST standard solutions of \( pH = 4.00 \) and 7.00. The second subset of solutions consists of McIlvaine buffers, i.e. a mixture of 20 mM \( K_2HPO_4 \) and 15, 6.7 or 1.7 mM citric acid. This yields final pH values of 4.42, 6.19, and 7.64 respectively, as checked using a freshly calibrated pH meter. The final subset of solutions consists of a \( 10^{-2} \) M KOH solution, and the control, biotechnology performance certified-grade water, to attest of the absence of parasite particles in all solutions prepared. A similar series of biotin-containing solutions was prepared by dissolving biotin powder (Pierce Biotechnology, Inc. 29129, Rockford, IL) to a final concentration of 1.32 mM. All solutions were filtered using a syringe filter with Nylon membranes with 0.2 µm pores (Pall Acrodisc-25, Port Washington, NY) to remove all dust particles and undissolved buffer or biotin crystals.

**Glass slides preparation and droplet deposition**

Microscope glass slides (12-544-1, Fisher Scientific, Pittsburgh, PA) were cleaned by immersion in a 3:1 volume mixture of sulfuric acid (\( H_2SO_4 \)) and 30% hydrogen peroxide (\( H_2O_2 \)) for 2 minutes, then rinsed extensively with filtered deionized water, and blown dry with compressed nitrogen gas in a class 1,000 cleanroom. To reduce wettability for deposition of the red wine drops, glass slides rinsed with filtered deionized water and dried with a stream of nitrogen gas were used. Streptavidin-coated glass slides (SMS, Arrayit Corporation, Sunnyvale, CA) were also rinsed with filtered deionized water, and subsequently dried with a stream of filtered nitrogen. Then, 0.1 µL droplets were deposited on the glass slide using recently calibrated micro-pipettes (0.1–2.5 µL, Eppendorf, Hauppauge, NY and P2, Gilson, Inc., Middleton, WI) by making slight contact between the surface and the liquid protruding from the pipette tip and subsequently pulling the micropipette away from the surface. The pipetting accuracy was determined by measuring the area occupied by 20 spots of a solution made of 1.32 mM biotin in water relative to the entire image field of view. This solution was chosen because it corresponded to one of the largest spreads of all solutions tested while still leaving a residue dense enough to be easily distinguished from the rest of the surface. The dried residue area was approximated by a box stretched to fit the residue and the area was measured by software (ImageJ, NIH, Bethesda, MD) and normalized to the entire image area. The area averaged at 7.5% of the entire field of view with a standard deviation of 0.8%. These numbers do not only reflect manual pipetting errors but also local differences in the glass surface. The slides were arrayed at room temperature of 20–23 °C with a relative
humidity (RH) of 20–50% and immediately placed in a desiccator filled with anhydrous calcium chloride (CaCl$_2$) or calcium sulfate (CaSO$_4$) powders where they were allowed to dry for 24 h before imaging.

### Image acquisition

The images of the stains from consumable fluids shown in Fig. 1 were acquired using an inverted microscope (IX 71, Olympus, Center Valley, PA) equipped with a 2X objective (U1S2 PLN, NA = 0.06, WD = 5.8 mm, Olympus) and a color complementary metal oxide semiconductor (CMOS) camera (PL-A776, PixeLINK, Ottawa, ON, Canada) under bright field transillumination from a halogen lamp (U-LH100L-3, Olympus) through frost (LP453900, Olympus) and day light filters (9-U115). The acquired image with $2,040 \times 1,536$ pixel resolution at 20 ms exposure time resulted in a picture scale of 1.61 $\mu$m/pixel. Images of the stains from biological fluids were acquired using a Nikon Eclipse Ti-U inverted microscope equipped with a Nikon 2X objective (Plan Achromat UW, Nikon, NA = 0.06, WD = 7.5 mm) and a color CMOS camera (PL-E425CU, PixeLINK, Ottawa, ON, Canada). The acquired image resolution was set to $2,040 \times 1,536$ pixels with a 20 ms exposure time. Calibration with an etched glass ruler produced a calculated resolution of 1.03 $\mu$m/pixel. The intensity of the incident light (100 W Halogen) was adjusted to deliver a uniform background, are removed from the histogram. The size of the vector $\mathbf{f}_L$ used in the calculations was $10 \times 1$ for the consumable fluid dataset and $243 \times 1$ for the biological fluid dataset.

**Gabor features** are represented by the Gabor vector $\mathbf{f}_G$. Gabor filters are a set of filters, also called Gabor wavelets, designed to describe the local texture properties of an image in various directions and scales. In our case, we designed Gabor filters in 4 color channels, i.e., Y, Cb, or Cr, we first compute a pixel histogram (see Fig. 2b), and then calculate its mean, standard deviation, skew, energy, and entropy. Consequently, each of the three color components is represented by five features, so that $\mathbf{f}_C$ is a $15 \times 1$ vector. The range of each dimension was normalized linearly to [0, 1] in order to balance the importance among each feature.

**Local binary patterns (LBP)** are represented by the vector $\mathbf{f}_L$. LBP is a popular texture operation which labels the pixels of an image by thresholding its neighborhood with the gray level value of the center pixel. Each pixel is associated with a sequence of binary number, and the histogram of the decimal numbers that correspond to the binary sequence of these binary numbers is taken as the LBP features for the image. For example, if a neighborhood of 8 pixels is used, as in the consumable fluid dataset, each center pixel is assigned with a sequence of 8 binary numbers, and the histogram will be of length $2^8 = 256$. One example of LBP with $3 \times 3$ neighborhoods (8 pixels) and uniform pattern is shown in Fig. 1c. If a neighborhood of 16 pixels is used, as in the biological fluid dataset, the histogram is of length $2^{16} = 65,536$. To reduce the length of the feature vector, and more importantly, to implement rotation-invariant descriptor, patterns that typically reflect noise, from e.g. a uniform background, are removed from the histogram. The size of the vector $\mathbf{f}_L$ used in the calculations was $10 \times 1$ for the consumable fluid dataset and $243 \times 1$ for the biological fluid dataset.

### Preprocessing of images and stains

The automatic localization of the stain was performed as shown in Fig. 2a, where the region containing the stain was cropped into a rectangle slightly larger than the stain to reduce the adverse effect of irrelevant background. Raw input images were then converted into grayscale and binary formats, sequentially. All the holes in the inverted images were filled up to have complete objects and an object of the largest area was treated as a stain. A bounding box was fitted to this area and the area of the stain was cropped from the original image with an offset on all sides in order to ensure the entire stain was selected. This cropped image was then resized to a $256 \times 256$ pixel image. Complete sets of the cropped stain images from consumable and biological fluids are shown in Fig. 1a and Fig. 4.

### Feature extraction: color, LBP, gabor wavelet and combined features

This section explains how a stain is described by a combined feature vector $\mathbf{f} = [\mathbf{f}_C, \mathbf{f}_G, \mathbf{f}_L, \mathbf{f}_S]$, where $\mathbf{f}_C$, $\mathbf{f}_L$, $\mathbf{f}_S$, and $\mathbf{f}_G$ are row vectors representing respectively the color features, local binary patterns, Gabor features and size of the stain.

**Color features** were described by the vector $\mathbf{f}_C$. Color features are among the simplest yet extensively used low-level features for content-based image analysis. They are robust against noise, resolution, orientation and resizing, although they convey little semantic meaning. The color feature of an image can be captured by its pixel distribution (histogram), in some color space. Here, we adopt the 3-component YCbCr color space which is perceptually more meaningful than the original RGB space. For each
scales and 6 directions, giving 24 filters in total. Each filter returns two values or responses. We divide each cropped image along the vertical and horizontal symmetry axis into 4 sub-images, and perform Gabor transform on each sub-image. With 24 filters, there are 48 response results for each sub-image, corresponding to a Gabor vector \( \mathbf{f}_i \) of size 192\( \times \)1 for each image. Fig. 2d–g shows some Gabor transform results on a sub-image. The size feature vector \( \mathbf{f}_s \), of size 1\( \times \)1, was simply set as the total number of pixels in the cropped image.

Description of the pattern recognition algorithms

For classification purposes, the k-nearest neighbor algorithm was used. For clustering, three algorithms were compared, the k-means, average linkage and spectral clustering algorithms. These algorithms can be described as follows.

k-nearest neighbor classification algorithm\(^{10,11} \) (k-NN)

The k-nearest neighbor algorithm (k-NN) is a supervised learning algorithm for classifying objects based on closest training examples in the feature space. A training dataset \( X = \{(x_i,y_i)\}^{n-1} \) of \( n \) labeled images is first built with each image represented by its feature vector \( x_i \in \mathbb{R}^d \), and by its known integer class label \( y_i \in \{1,\ldots,c\} \), with \( c \) the chosen number of classes. Then a test object \( z \in \mathbb{R}^d \) is classified to the majority class of its \( k \) nearest neighbors in the training set to the minimum \( \| z - x_i \| \). In the case of \( k = 1 \), an object is simply assigned to the class of its nearest neighbor in the training dataset.

Feature normalization

To balance the importance of each feature used in classification, we linearly normalize each feature dimension in the training data into the interval [0, 1], and then apply the same transformation to the associated feature dimension in the test data. Specifically, let \( f_i \) be the \( i \)-th feature dimension of the \( n \) training data, and let \( f_{i,\text{min}} = \min(f_i) \) and \( f_{i,\text{max}} = \max(f_i) \). The goal is to apply a linear transformation:

\[
T_i : f_i \rightarrow [0, 1] \\
x \rightarrow a_i x + b_i
\]

so that \( T(f_{i,\text{min}}) = 0 \) and \( T(f_{i,\text{max}}) = 1 \). Note that this can be achieved with:

\[
a_i = \frac{1}{f_{i,\text{max}} - f_{i,\text{min}}} \quad \text{and} \quad b_i = -a_i f_{i,\text{min}}.
\]

For a test image \( x = (x_1,\ldots,x_d) \), we then apply the transformation \( T(x) = (T_1(x_1),\ldots,T_d(x_d)) \) before making the classification. For clustering, each feature dimension of the data is linearly normalized to interval [0, 1].

k-means clustering algorithms\(^{32} \)

For an unlabeled dataset \( z_1,\ldots,z_n \in \mathbb{R}^d \), the k-means algorithm aims to find \( k \) centers \( c_1,\ldots,c_k \in \mathbb{R}^d \) to minimize the following quantization loss:

\[
\sum_{i=1}^{n} \min_j \| z_i - c_j \|^2.
\]

The integer number \( k \) is a manually set number that corresponds to the number of clusters into which the images are to be sorted. An object is then grouped to its nearest center, using the nearest neighbor algorithm. Solving this optimization problem exactly is computationally expensive. In practice, one alternates between the data partition and center update. First, \( k \) centers are randomly selected. Second, each object is grouped to its nearest center. Third, the current center in a cluster is replaced by the means of the objects in that cluster. The latter two steps are repeated until convergence.

Average linkage algorithm\(^{33} \)

The average linkage algorithm belongs to the family of linkage clustering algorithms. It yields a cluster hierarchy over the data, in a bottom-up manner. Each object is treated as an individual cluster first. Then, the two closest clusters are merged. This procedure is repeated until some desired conditions are met. In our case, we stop the process when \( k \) clusters of the data are obtained. In average linkage, the distance between two clusters is taken as the average distance between objects across clusters, \( i.e. \):

\[
d(C_i, C_j) = \frac{1}{n_i n_j} \sum_{x \in C_i, y \in C_j} \| x - y \|
\]

where \( n_i \) and \( n_j \) denote the numbers of objects in clusters \( C_i \) and \( C_j \), respectively.

Spectral clustering algorithm\(^{34} \)

The spectral clustering algorithm is a graph-based method. Given a dataset, a graph is built with each object as a node, and the similarity between two objects serves as the weight on the...
edge joining the associated nodes. Let $W = (w_{ij})$ be the similarity matrix of the graph, with $w_{ij}$ capturing the similarity between nodes $i$ and $j$. Typically, one takes:

$$w_{ij} = \exp(-\|x_i - x_j\|^2/\sigma),$$

where $\sigma$ is a scale factor. The degree of node $i$ is defined as:

$$d_i = \sum w_{ij},$$

and $D = \text{diag}(d_1, \ldots, d_n)$ is called the degree matrix. The Laplacian matrix of the graph is defined as $L = D - W$. We use Normalized Cuts (NCuts), a state-of-the-art spectral clustering, for our clustering problem. The NCuts aims to find a balanced partitioning of the graph, leading to the following generalization of eigenvalue problem:

$$Lv = \lambda Dv$$

where the eigenvector $v_2$ corresponding to the second smallest eigenvalue is the relaxed indicator vector for two-way partition. For K-way partition, the relaxed cluster indicator matrix can be obtained as $F = (v_2, \ldots, v_K)$ with $v_i$ denoting the unit eigenvector corresponding to the $i$th smallest eigenvalue. To derive $K$ clusters, one usually applies k-means to the rows of $F$, and object $i$ is grouped to the cluster of row $i$.

Normalized mutual information (NMI)

To measure the quality of the clustering process, the NMI method is applied as follows. For a clustering of the data, denoted as $P_1 = (C_1, \ldots, C_K)$, we can define a discrete random variable $X$ as the cluster-membership of a randomly selected object. Thus $X$ can take on $K$ values, and

$$P(X = C_i) = \frac{n_i}{n},$$

where $n_i$ is the number of objects in cluster $C_i$. Suppose $Y$ is the random variable associated with another partition of the same data $P_2 = (A_1, \ldots, A_M)$. The joint distribution of $X$ and $Y$ is:

$$P(X = C_i, Y = A_j) = \frac{n_{ij}}{n},$$

where $n_{ij}$ is the number of objects in $C_i \cap A_j$. The NMI between the partition $P_1$ and $P_2$ is defined as:

$$\text{NMI}(P_1, P_2) = \frac{I(X, Y)}{\sqrt{H(X)H(Y)}},$$

where $I(X, Y)$ is the mutual information between $X$ and $Y$, and $H(X)$ and $H(Y)$ are the entropies of $X$ and $Y$, respectively. NMI is within the closed interval $[0, 1]$, and the larger the better.

Results and discussion

Visual inspection of stains from consumable fluids

Fig. 1 shows the stains obtained from the consumable fluid drops. All stains had approximately the same size, roughly equal to the initial wetted area. The stains were highly reproducible for a given consumable fluid and distinct among fluid types. Beer stains showed no significant differences in terms of shape and color except for the Guinness Extra Stout, which showed a browner annulus in the vicinity of the wetting line as the stout itself was darker than the other beers. Some waves or fingering were visible along the wetting line in the Tropicana grape juice stain, while other stains from juice showed no fingering at the wetting line. Circular black spots and fibrous deposits were observed in the stains from orange and tomato juices, respectively. The stains from seven types of milks showed different levels of brown color, which seemed proportional to the nominal concentration of fat. On the other hand, crack patterns were observed on the peripheral annulus for the samples with the lowest fat concentration. Milk stains with additional ingredients like chocolate, strawberry, as well as soy milk, showed randomly distributed spots. Wine stains from Merlot and Cabernet Sauvignon showed radial wrinkles, while stains from Syrah did not. The four kinds of soda stains showed similar ring deposits, with a thick ring for the sugary sodas and a thin ring for the sugar-free diet Coke.

Pattern recognition of stains from consumable fluids

We first applied a simple method, the k-nearest neighbor (k-NN) algorithm with $k = 1$ (see Methods) for the classification of the dataset of stains from consumable fluids. The Euclidean distance metric was used to measure the distances between feature vectors. In the set of the consumable fluids, shown in Fig. 1, there were 24 classes, with 20 images in each class. We used 10 images from each class as training data, and treated the remaining 10 images as test data. We tested classification accuracy based on each extracted feature vector like color distribution, LBP, Gabor wavelet and size feature and their combination, according to the image processing steps shown in Fig. 2. Despite its simplicity, the 1-NN algorithm based on the extracted features significantly improved the classification accuracy compared to random assignments, which was only 4.2% (Fig. 3a). Especially, color distribution feature played a dominant role in classification with an accuracy reaching 94%, while other features, LBP, Gabor wavelet and size features returned accuracies of 89%, 75%, and

![Fig. 3](https://example.com/fig3.png)

The use of (supervised) classification and (unsupervised) clustering pattern recognition algorithms identifies stains much more accurately than by random assignment: (a) Classification accuracy of the consumable and biological fluids based on the random assignment and the 1-nearest neighbor algorithm using the color distribution, local binary pattern, Gabor wavelet, size and combination features. (b) Clustering accuracy of consumable and biological fluids based on the random assignment and in the normalized mutual information (NMI) using the color distribution feature. RA: Random Assignment. CD: Color Distribution. LBP: Local Binary Pattern. GW: Gabor Wavelet. S: Size. C: Combination. KM: K-Means. AL: Average Linkage. SP: Spectral Clustering.
17%, respectively. Most of the inaccuracy of the classification based on the color distribution feature arose from beer stains as shown with the confusion matrix (see Fig. S1, ESI†). This is because the beer stains looked very similar. The classification accuracy of the beer stains based on the color distribution feature was 72% and excluding beer stains from the dataset increased the classification accuracy of consumable fluids stains to 99%.

Classification based on a combined feature vector $f = [\alpha f_c, \beta f_L, \gamma f_G, f_s]$, with optimal value of the weighting factors $\alpha = 1 - \beta - \gamma - \epsilon$, was determined using the leave-one-out cross-validation (LOOCV) method described in the Methods section. Each of the weighting factors ranged between 0 and 1. For the consumable fluid data, the best weighting factors determined by the LOOCV method were $\alpha = 0.8, \beta = 0.1$ and $\gamma = 0.1$, and $\epsilon = 0$; this corresponded to a classification accuracy of 93%, which is slightly lower than the accuracy based on the color distribution feature alone (94%). Though this result appears to be counter intuitive, it is possible in practice. When a single feature is dominant in classification performance, combination of that feature with less dominant features may not improve or even degrade the accuracy. The confusion matrix for the classification accuracy of consumable fluids based on the 1-nearest neighbor classification showed no improvement for the color distribution feature (94%). Though this result appears to be counter intuitive, it is possible in practice. When a single feature is dominant in classification performance, combination of that feature with less dominant features may not improve or even degrade the accuracy. The confusion matrix for the classification accuracy of consumable fluids based on the color distribution feature was shown in Fig. S1, ESI†.

We also applied several popular clustering algorithms to the dataset including k-means, average linkage, and spectral clustering (described in the Methods section). Clustering is the process of grouping data into clusters, such that objects within a cluster are similar to each other while those across clusters are dissimilar, under certain criteria. In order to measure the quality of the clustering performance, we adopt the Normalized Mutual Information (NMI) by measuring the normalized mutual information between the clustering result and the ground-truth clusters (see Methods). The maximum clustering accuracy of 87% for the consumable fluid dataset in NMI was achieved when the color distribution feature was used as shown in Fig. 3b.

**Visual inspection of stains from biological fluids**

Another collection of stain images was prepared using biological fluids (Fig. 4) such as phosphate buffer ($K_2HPO_4/KH_2PO_4$) at different volume ratio, phosphate solutions ($K_2HPO_4$) added of different volumes of citric acid to control pH, and KOH solutions. To detect the effects of specific molecular interactions between the solution and the solid substrate, two versions of each fluid were prepared (one including biotin, and the other not), and two versions of the glass slides were prepared (one coated with streptavidin and the other not). The biotin-streptavidin system was chosen because it is known to be the strongest non-covalent interaction between a protein and its cofactor binding constant of the system, i.e. $K_D = 4 \times 10^{-14} \text{M}^{36}$ or, equivalently, a force of 160 pN as measured by atomic force microscopy.37 This system also has a widespread use in biotechnology,38 in particular in high affinity sensitive immunoenzymatic assays (e.g. ELISA).

All the prepared biological fluids deposited on the clean glass and streptavidin-coated glass slides formed unique stains as shown in Fig. 4, except the pure water droplet deposited on a clean glass slide (which is not shown because it did not leave a visible stain and attests of the purity of the aqueous solutions made). The variation in diameter of the biological stains was on the order of one order of magnitude, much larger than for the consumable stains. When drops of 10 mM phosphate buffer with different volume ratio were deposited on the clean and streptavidin-coated glass slides, a small-diameter bump was formed, probably because of Marangoni convection coupled with receding of the initial wetting line.23 Visual inspection revealed the following about the effect of pH: the most acid solutions (15 mM citric acid solution) left crystallized patterns in the bump on the clean glass slide and scattered granular patterns on the streptavidin coated glass slide, while more basic solutions (6.7 mM and 1.7 mM of citric acid) formed more homogeneous bumps on both glass slides. Interestingly, the KOH solution deposited on the streptavidin-coated glass slide showed circular snowflake stains whereas it formed random spots on the clean glass slide. Adding 1.32 mM biotin to each solution drastically changed the stain pattern, as shown by comparing the first with the second column in Fig. 4 for the clean glass surface and the third with the fourth column for the streptavidin-coated glass case. The presence of biotin in the phosphate buffer with 15 mM citric acid produced needle patterns in the stains when deposited on the clean glass slide, whereas it formed globular patterns when biotin was not in solution. Addition of biotin to KOH and pure water deposited on the clean glass slide produced thin needles pointing inward from the periphery. Stains from all the solutions added with biotin maintained circular shapes after the drying process.

![Stains of biological fluids show complex patterns that can be used for identifying the fluid and substrate chemistry.](image-url)
process, possibly due to its strong non-covalent interaction between biotin in solution and streptavidin coated on the glass slide. However, different patterns were observed at the interior of the stains depending on the solution used. Altering the volume ratio of the phosphate buffers resulted in stains with short needles at the wetting line and multiple concentric lines inside the wetting line, coarse long needles pointing inward at the periphery without inside concentric lines and scattered granular pattern, as the volume ratio increased. The 10 mM citric acid in phosphate solution with biotin showed thorny crown patterns when deposited on the streptavidin coated glass slide. Stains from 6.7 mM and 1.5 mM citric acid in phosphate solution with biotin on the streptavidin glass slide formed more apparent annular ring patterns compared to when deposited on the clean glass slide. Pure water and KOH drops added with biotin showed almost identical stains when deposited on the streptavidin coated glass slide, i.e. shorter needles around the periphery than when deposited on the clean glass slide.

**Pattern recognition of the images from biological fluids**

The dataset of biological fluids includes 30 classes with different fluid composition and substrate chemistry. 20 images were taken per class, with 10 images per class used as training data and the remaining images used to test the pattern recognition algorithms. The accuracy of the classification process is shown in Fig. 3a. The best accuracy of 81% is obtained by the combination of the four features (color, LBP, Gabor and size), and compares favorably with an accuracy of 3.3% obtained by random assignment. As with the dataset of consumable fluids, the color distribution was the most discriminative single feature, followed by the LBP, Gabor and size feature. Contrary to the dataset of consumable fluids, the combination of four features provided higher accuracy than the single feature of color distribution, probably because the features observed in the biological fluids were richer than the ones observed in the consumable fluids, see Fig. S2, ESI.† The classification accuracy with the biological fluids was not as high as the accuracy with the consumable fluids.

Fig. 3b shows the clustering accuracy, using three different clustering algorithms. The spectral clustering algorithm delivered the best result, with 66% accuracy for 30 classes.

Fig. 5 describes the performance of classification algorithms to determine the presence or absence of biotin and streptavidin respectively in the liquid and on the glass (4 classes), the salt composition (3 classes), or the pH of the solution (3 classes). For each of the three tasks, pattern recognition algorithms using combined features returned a classification accuracy larger than 90%. This is a remarkable result since the influence on the stain image of pH, biotin/streptavidin, and citric acid is difficult to determine by visual inspection.

**Conclusions**

Biological and consumable fluids were classified and clustered by pattern recognition techniques based on the descriptive features (color distribution, local binary pattern, Gabor wavelet, and size) extracted from photographic images of the stains. For both datasets, the color distribution was the most discriminative single feature. The nearest neighbor classification based on the combined feature achieved the highest classification and clustering accuracy for biological fluids. However, for consumable fluids, the color distribution feature alone achieved the highest accuracy. Results with the dataset of stains from consumable fluids showed a slightly higher accuracy than that from biological fluids. This is probably due to the large variations of composition of the biological fluids considered. The algorithms were also remarkably accurate at determining the presence or absence of biotin and streptavidin respectively in the liquid and on the glass, the salt composition, or the pH of the solution. The pattern recognition scheme developed in this work showed feasibility of pattern recognition in recognizing specific fluids from the raw stain images of the fluids. This work is the first example of computer-based classification and clustering of stains from consumable and biological fluids using pattern recognition algorithms. The application of this methodology for clinical diagnostics, such as the classification of pleural effusions for malignancy, congestive heart failure or lung infection, has the potential to enable rapid screening at the point-of-care. The interest of the method lies in the possibility of investigating other protein systems than avidin/biotin in order to: 1) quickly determine protein content of clinical samples such as pleural effusions or cerebrospinal fluid for early diagnosis, and 2) investigate specific interactions between the fluid and biomarkers patterned on the surface. Future improvements of the method include increasing the throughput by involving the use of a micro-arrayer to form the droplets.

**Acknowledgements**

NK, DA, ZL and SFC acknowledge support from NSF award #1034349. CH and FZ acknowledge support from the iBis Foundation of Arizona.
Notes and references


29. The datasets used in this work are available for download, distributed for non-commercial research purpose only, at: http://www.ee.columbia.edu/dvmm/staindata Please cite this work in publications of any work that uses the datasets.


