Title

# Machine Vision Methods, Natural Language Processing, and Machine Learning Algorithms for Automated Dispersion Plot Analysis and Chemical Identification from Complex Mixtures

Permalink

https://escholarship.org/uc/item/7tt8x45x

Journal

Analytical Chemistry, 91(16)

ISSN

0003-2700

Authors

Yeap, Danny

Hichwa, Paul T

Rajapakse, Maneeshin Y

et al.

Publication Date

2019-08-20

DOI

10.1021/acs.analchem.9b01428

Peer reviewed

Published in final edited form as:

*Anal Chem*. 2019 August 20; 91(16): 10509–10517. doi:10.1021/acs.analchem.9b01428.

# Machine vision methods, natural language processing and machine learning algorithms for automated dispersion plot analysis and chemical identification from complex mixtures

**Danny Yeap**[1], **Paul T. Hichwa**[1,§], **Maneeshin Y. Rajapakse**[1], **Daniel J. Peirano**[1,†], **Mitchell M. McCartney**[1], **Nicholas J. Kenyon**[2,3,4], **Cristina E. Davis**[1,*]

[1]Department of Mechanical and Aerospace Engineering, University of California Davis, Davis, CA 95616, USA

[2]Department of Internal Medicine, 4150 V Street, Suite 3400, University of California, Davis, Sacramento, CA 95817, USA

[3]Center for Comparative Respiratory Biology and Medicine, University of California, Davis, CA 95616, USA

[4]VA Northern California Health Care System, 10535 Hospital Way, Mather, CA 95655, USA

## Abstract

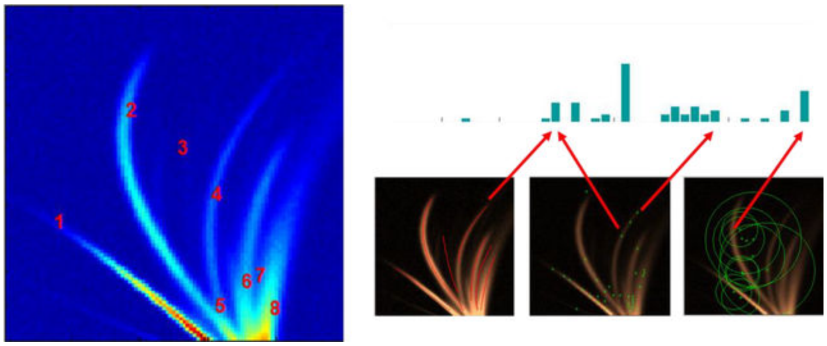
Gas phase trace chemical detection techniques such as ion mobility spectrometry (IMS) and differential mobility spectrometry (DMS) can be used in many settings, such as evaluating the health condition of patients or detecting explosives at airports. These devices separate chemical compounds in a mixture and provide information to identify specific chemical species of interest. Further, these types of devices operate well in both controlled lab environments and in field applications. Frequently, the commercial versions of these devices are highly tailored for niche applications (e.g. explosives detection) because of the difficulty involved in reconfiguring instrumentation hardware and data analysis software algorithms. In order for researchers to quickly adapt these tools for new purposes and broader panels of chemical targets, it is critical to develop new algorithms and methods for generating libraries of these sensor responses. Microelectromechanical system (MEMS) technology has been used to fabricate DMS devices that miniaturize the platforms for easier deployment; however, concurrent advances in advanced data analytics are lagging. DMS generates complex three-dimensional dispersion plots for both positive and negative ions in a mixture. While simple spectra of single chemicals are straightforward to interpret (both visually and via algorithms), it is exceedingly challenging to interpret dispersion plots from complex mixtures with many chemical constituents. This study uses image processing

[*] **Corresponding Author** (CED) cedavis@ucdavis.edu.
[§] presently at Invicta Medical, Inc.
[†] presently at Google, Inc.

ELECTRONIC SUPPLMENTAL INFORMATION AVAILABLE:
The AnalyzeIMS (AIMS) software is available on GitHub, including a new release associated with this manuscript (version xxxx). Please refer to Professor Cristina Davis' webpage for more information. AnalyzeIMS is available as open source for research and personal use under a modified BSD license. Commercial licensing may be available, and a license fee may be required. The Regents of the University of California own the copyrights to the software. Future published scientific manuscripts or reports using this software must cite the original software publication (DOI:10.1007/s12127-016-0200-9) and this publication associated with this specific version release (DOI:10.1007/xxxxxxxxx).

and computer vision steps to automatically identify features from DMS dispersion plots. We used the bag-of-words (BoW) approach adapted from natural language processing and information retrieval to cluster and organize these features. Finally, a support vector machine (SVM) learning algorithm was trained using these features in order to detect and classify specific compounds in these represented conceptualized data outputs. Using this approach, we successfully maintain a high level of correct chemical identification, even when a gas mixture increases in complexity with interfering chemicals present.

## Graphical Abstract



**Keywords**

differential mobility spectrometry (DMS); dispersion plot; machine learning; machine vision; field asymmetric ion mobility spectrometry (FAIMS)

## INTRODUCTION

The growth and development of gas-phase chemical analysis has potential to spark innovation across multiple industries. Mass spectrometry and other gold standard chemical analysis measures have proven to be highly accurate, sensitive, and selective. However, these platforms are often bulky, expensive, and require resource-intensive laboratory conditions[1, 2]. Research advancements in real-time and mobile platforms for chemical sensing and detection have the potential to impact many areas, such as non-invasive health diagnostics[3, 4], agriculture monitoring[5], security applications[6–8], and air quality monitoring[9]. Many of these new mobile technologies have major issues with reliability, calibration, and robust operation in field conditions. Commercially, ion mobility spectrometry[10, 11] (IMS), differential mobility spectrometry[12, 13] (DMS) and metal oxide sensors[14, 15] have so far proven to be some of the most robust technologies, although other sensor technologies continue to gain traction with performance improvements.

Differential mobility spectrometry[16–18] (DMS) differentiates chemicals via ion mobility characteristics and offers some critical features advantageous to a portable system. First, devices operate under normal atmospheric conditions (~1 atm) and at relatively low temperatures (50–100 °C). Second, positive and negative ions can be detected simultaneously, opening options for sophisticated data analysis and chemical prediction methods. Third, the nature of the ion mobility in DMS provides a mechanism to allow select ions to be detected per scan. This creates a means to increase chemical selectivity and overall chemical identification.

Most DMS systems, including commercial devices, are typically limited in automated data analysis with a fixed range of chemical prediction. In the past, DMS has focused on single chemical identification for specific applications[19, 20]. Additionally, there are limited automated methods for developing effective predictive models[21, 22]. Data obtained from DMS devices is represented as a dispersion plot, a 2-dimensional surface plot of charged ion signal intensities (z axis) representing ion mobility in an increasing electric field (y axis) under increasing compensation voltages (x axis). It is exceptionally difficult to identify specific chemicals when mixtures of chemicals are analyzed in a single sample, because differing ions “overlap” in the dispersion plot space, and sometimes complex clusters of ions result from chemical reactions as a sample traverses the device. Traditionally, a trained chemist will visually inspect a dispersion plot or calculate the alpha functions[23] of the intensity peak ions in order to differentiate one or two chemicals. However, these solutions are time consuming and unrealistic for in-field, real-time portable DMS sensors. Additionally, these solutions are laborious when a significant number of samples are needed as well as when samples consist of complex chemical mixtures, such as biological samples which might be comprised of several hundreds of compounds.

Several emerging topics in the area of “big data analytics” can potentially be applied to automated dispersion plot analysis to streamline chemical detection in both simple and complex mixtures. Generally, there are algorithm approaches for major steps in data processing, such as: removing noise, feature extraction from the signal space, machine learning to build a model against the features of the data, and finally prediction of unknowns using the model. Specifically, we are interested in using existing computer vision techniques to extract features and apply machine learning to generate prediction models.

This paper outlines the development of a data analysis methodology as well as the software and algorithm application for generating chemical prediction models for portable DMS devices. We apply image processing techniques to the DMS data followed by computer vision feature extraction methods and the generation of a representation model from natural language processing in order to finally build predictive models for use in portable DMS devices. Specifically, we start with a grayscale DMS dispersion plot[24] allowing a computer vision algorithm to identify features such as peak ions, corners, and points-of-interest. Phase symmetry[25, 26], phase congruency[27], and morphological erosion[28–30] are used to detect the peak ions from the DMS dispersion plot. Features from accelerated segment test[31] (FAST) and speeded up robust features[32–34] (SURF) algorithms were used to detect corners and points-of-interest. To find similarities between the features, hierarchical clustering was used to generate the bag-of-words model for each sample. Support vector machine[35–37] (SVM) was used on the bag-of-words to find a relationship between the samples, and subsequent detection accuracies were generated from each sample to show the robustness of our methodology. SVM models are built for each compound. We tested the prediction methods in the presence of other interfering chemicals that frequently confound dispersion plot analysis. The methodology is useful because it is a more comprehensive method utilizing wide range of compensation voltage and RF voltage to develop machine learning models.

These models provide a quantitative approach to detect the presence of a compound in a mixture. In practice, this methodology can be used to alert people of dangerous chemicals in close proximity such as explosives or gas leaks. For explosive detection, common compounds found in the airport environment and explosive compounds would be mixed together. Using computer vision and machine learning, models would be built to identify the presence of different explosive chemicals in the mixture. These models can be used to create a virtual library of models that can be used to detect explosive compounds in an actual airport environment by using the same computer vision methodology and trained SVM models to provide predictions.

## EXPERIMENTAL

### Gas Sample Preparation.

Individual samples were prepared by injecting desired liquid volumes of 2-butanone, 2-hexanone, ethyl acetate and 4-methyl-2-pentanone (Sigma Aldrich, St Louis, MO) into each sample bag (SKC Tedlar® Sample bag, SKC Inc., Eighty Four, PA) filled with 3 L of ultra high purity nitrogen (Airgas, Radnor, PA). These bags were maintained for more than 10 min at room temperature to equilibrate analytes into the gas phase to achieve the stock concentrations (1000 ppm). Serial dilution was performed from the stock concentration to achieve 100 ppm concentration of each chemical. Each gas phase chemical was sampled into a gas tight 1 mL glass syringe (Hamilton Co., Reno, NV) from the final concentration (100 ppm) and introduced separately into the inlet of the differential mobility spectrometer using a syringe pump with a dilution nitrogen gas flow. The final concentration of each chemical at the DMS cell inlet was 500 ppb. The bags with chemical mixtures were prepared by adding each new chemical to the bag followed by serial dilution. Therefore, the final concentration of each chemical in a mixture was ~ 500 ppb.

### Instrumentation.

A MicroAnalyser™ differential mobility spectrometer (Sionex Corp., Bedford, MA) with a 5 mCi, $^{63}$Ni ionization source was used in this study. The drift cell was operated with at 80 °C for all data collection. To avoid chemical memory effects between analyses, the drift cell was purged with ultra-pure nitrogen while heating at 100 °C. Sample lines, made of PTFE tubing, were purged with ultra-pure nitrogen during this cleaning cycle and background reactant ion peak (RIP) dispersion data were recorded prior each sample injection to verify the drift cell was free of memory effects from previous chemicals. Ultra-pure nitrogen (Air gas, Radnor, PA) was used as the carrier gas (300 mL min$^{-1}$) and sample gas (20 mL min$^{-1}$). DMS dispersion plots were recorded for each sample by scanning the compensation voltage (CV) from +10 to −30 V range (step size = 100; step duration = 10 msec; step settle time = 3 msec) and scanning the RF voltage from 500 to 1490 V in 10 V increments (RF voltage steps = 100). Each sample was repeated n=12 times. Both positive and negative polarity DMS dispersion data were used in the study.

### Data Analysis.

To visualize raw data, AnalyzeIMS (AIMS)[21] v1.301 was used. AIMS was developed using MATLAB 2017 and has been made available for open source for research and non-profit personal use (please see access information in software section below), and the original version did not require specific toolboxes. The old version of AIMS from a paper in 2016 has the capabilities of plotting and visualizing raw DMS data and performing some noise reduction and baseline correction on the raw data. It is also written with an older version of Matlab 2014. The new version published along with this paper has been upgraded to a more recent version of Matlab 2017. All the computer vision and machine learning methodologies such as peak ion, corner, points-of-interest and SVM outlined in this paper are also implemented in the new version. This current release utilizes the standard statistics and machine learning and computer vision toolboxes provided by MATLAB. Using these tools, each data sample in the present work was visualized to ensure data quality and reproducibility.

The dispersion plot analysis algorithms developed herein were written in MATLAB 2017. The Matlab platform allows for computer vision techniques to be applied directly to DMS output data for feature characterization, or “bag-of-words” compilation, for different compounds. Once the bag-of-words model was extracted, a one-versus-all multiple classification support vector machine (SVM) classifier was trained to build a robust model to detect the presence of chemical compounds. The purpose of chemical mixtures in this study is to determine if SVM can accurately discern a single compound from a mixture.

## RESULTS AND DISCUSSION

### DMS Dispersion Plots.

A dispersion plot of the DMS background in positive polarity (no analytes present) is shown (Figure 1A). The reactant ion peak (RIP), hydrated protons appears dominant with a low intense peak left to the RIP likely due to some very low abundance impurities.

Positive polarity DMS dispersion plots were obtained for individual chemicals: 2-butanone, 2-hexanone, ethyl acetate and 4-methyl-2-pentanone (Figure 1B–E). The individual chemicals show clear evidence of monomer and dimer ion formations that are named as “m” for monomer ion and “d” for dimer ion in each plot. Ethyl acetate shows an intense ion peak “f1” (Figure 1D) at higher RF voltages while 4-methyl-2-pentanone shows low intense peak “f2” additionally to the monomer and dimer ions (Figure 1E). These new ion peaks may have occurred due to the fragmentation of parent ions by the RF field heating inside the drift cell, as observed by previous researchers[38]. The trace impurity peak presence in the background spectra can be visually observed in all the individual plots except in 2-butanone (Figure 1B). The high intensity of the 2-butanone monomer ion peak may have occurred by overlapping the impurity peak with 2-butanone monomer peak, a possible reason for the disappearance of the impurity peak in 2-butanone spectrum.

We also show the DMS dispersion plots for the chemical mixtures (Figure 2), and the dimer ions “D” for all chemicals in the mixtures are well overlapped, appearing around 0 V compensation voltage. The mixture of two ketones, 2-butanone and 2-hexanone clearly show the two monomer ions “m1” and “m2” (Figure 2A). The addition of ethyl acetate to the mixture of two ketones: 2-butanone and 2-hexanone (Figure 2B). Additional peaks of m3, monomer ion peak of ethyl acetate and f1, fragment ion peak from ethyl acetate are visible (Figure 2B) other than the dominant monomer peaks “m1 and m2” of the two ketones. Addition of the third ketone, 4-methyl-2-pentanone into this mixture alters the peak intensities, specifically the low ion intensity of the monomer ion peak of ethyl acetate. However, the characteristic peaks for each chemical are visible at different strengths that include the monomer ion peaks of 2-butanone, 2-hexanone “m1, m2” and possible fragment ions from ethyl acetate “f1” and from 4-methyl-2-pentanone “f2”. The monomer ion of 4-methyl-2-butanone may have overlapped with the monomer ion of 2-hexanone (Figure 2C).

## Peak Ion, Corner, and Points-of-Interest Detection of Dispersion Plot.

In order to identify and capture the peak ions (Figure 3A) of each DMS sample, a series of computer vision algorithm were used: phase symmetry, phase congruency, morphological erosion, and flood-and-fill. The peak ions are outlined as red curves (Figure 3A). For completeness of this process, the Supporting Information provides from more detail. This method of peak ion detection generates a single pixel width of all peak ions within the sample.

Corners are another important structure in computer vision that can be used to generate robust features vectors to differentiate images. We implemented this in our method, by using Feature from Accelerated Segment Test (FAST), an algorithm that iteratively checks every pixel in a DMS dispersion plot to detect a corner, represented as green crosses (Figure 3B). FAST determines if a pixel is a corner by first drawing a three pixel radius circle around any particular pixel of interest. This circle is composed of 16 pixels, labeled 1–16 in a clockwise manner starting from the top. If the intensity of the center of pixel is greater than three out of four pixels mentioned earlier then the center pixel is deemed a corner. This process is repeated for all pixels in the image. The advantage to using FAST is that it is computationally quick compared to other similar algorithms, such as the scale-invariant feature transform[32], because it does not need to check the intensities of all 16 pixels. FAST only generates the locations of all the corners.

Points-of-interests is another important feature detection approach that can be used to describe an image. These are specific pixels in an image that are important and can be accurately found even if the image were zoomed in or rotated. The algorithm we used to find points-of-interest is the Speed Up Robust Feature (SURF). SURF identifies key points which allows the algorithm to compute the key points more quickly compared to other key points algorithms such as Scale Invariant Feature Transform. SURF is used on the grayscale DMS dispersion plot, and the key points are represented as green crosses (Figure 3C). These points-of-interest are themselves scale invariant, meaning a zoomed version of the same image will produce similar if not the same results. The green circles represent the scale of the key points; a larger circle means that the key point is more scale invariant. SURF only generates the location of all points of interest.

## Feature Vectors Extraction.

The previous section outlined how to detect peak ions; however, feature vectors must be generated from these peak ions. Peak Ion feature vectors are always a fixed length of size

14 where each feature describes the numerous features such as perimeter and area of peak ions. The features are generated by drawing a bounding box over each ridge. The bounding box is the smallest rectangle that encompasses the peak ion. The area of the peak ion, its perimeter and centroid are approximated by using that of the bounding box. The eccentricity and the orientation of the peak ion are generated by drawing the smallest ellipse that can encompass the peak ion. This ellipse's eccentricity describes the curvature of the ellipse and the ellipse's orientation describes the tilt with respect to the major axis and x axis which in turn describes the peak ion's eccentricity and orientation.

The feature vectors for the corners and point of interest are generated in the same fashion. The previous section outlined the method used to generate the locations of corners and points-of-interest. In order to generate feature vectors, a SURF descriptor algorithm[39] is used to generate a fixed size of 64. A descriptor is used to help find relationships between the corners and points-of-interest. The advantage with a SURF descriptor is that it generates feature vectors that are robust to rotations and scale.

### Bag-of-words (BoW) Representation Model.

At this point, a series of algorithms refine a DMS dispersion plot into three descriptors: peak ions (Figure 3A), corners (Figure 3B) and points-of-interest (Figure 3C). The next step is to compress these three descriptors into a single new representation of the DMS sample; this is performed using a bag-of-words technique. Words are generated by using clustering algorithms on the peak ion, corner, and points-of-interest feature vectors. All features have scalar values. The size of features generated by peak ions is always size 14 and the size of corners and points-of-interest is always size 64. As a result, two separate clustering algorithms are used to generate words for feature vectors of length 14 and 64.

The peak ion words are generated by using k-mean clustering where k is selected as the max number of peak ions found amongst all the samples. The inputs of k-mean clustering are all the peak ions feature vectors in the samples and the output is an assignment of which cluster the peak ion belongs to. Peak ion features that are similar perimeter, area, eccentricity, etc will be put in the same cluster or word.

Each peak ion in the sample are placed into a cluster and one cluster denotes a word in the bag-of-words techniques.

The corner and points-of-interest feature vectors yield a list of highly dimensional feature vectors for each chemical sample. In order to find similarities between these feature vectors, a hierarchical clustering technique called agglomerative clustering[40]. Agglomerative clustering is a bottom-up empirical clustering algorithm that creates clusters until all corners, and points-of-interest in the sample have been assigned to a cluster. As the features are being clustered, the algorithm ensures that all clusters are similar size to prevent one cluster from getting too large and grouping features that are not similar together. The purpose of clustering is to capture any relationship between the different corners and points-of-interest in the sample, thus every cluster represents this relationship.

At this point, peak ion, corner, and points-of-interest clusters have been assigned to every feature vector. These clusters comprised together create the dictionary of words for the samples. A frequency graph of the number of features in each cluster is generated creating what is commonly called the bag-of-words (Figure 4). The frequency plot is one representation of bag of words model for each pure compound. The number of clusters is different across all chemicals and can be attributed to the relationships between image feature vectors when clustered. Clustering is performed on all samples in training set. In this case most of the cluster size were about 60 because only the pure compounds were used in the training set. However, when mixtures are added to the training set there can be different number of clusters due to the different sets of features vectors generated.

Bag-of-words models (or frequency graphs) are unique for every sample collected. Dispersion plots of the same chemical should produce similar frequency graphs, with slight differences caused by inherent noise from the DMS detector. As shown visually, representation models appear quite different for spectra of four different analytes (Figure 4).

The next step is to build classification algorithms to distinguish the bag-of-words models between the four analyte, creating a library of each chemical species to be used later for chemical identification of "unknown" samples.

## Support Vector Machine (SVM) Learning and Classification.

With the bag-of-words model completed for every dispersion plot, the next step is to identify a relationship between the different bag-of-words models for all samples of the same chemical. This algorithm would seek a common pattern that define a specific chemical, analogous to way a mass-to-charge $m/z$ fragmentation pattern is unique for chemical identification on mass spectrometers. We use a machine learning algorithm termed Support Vector Machine[35–37] (SVM) to achieve this, and the SVM outputs a model that can be used for chemical prediction of unknown samples in the future.

SVM is a binary supervised machine learning algorithm that uses a technique called labeling to facilitate learning. This means each bag-of-words must be labelled, 0 or 1, indicating whether or not the pure compound is present. With all the bag-of-words labelled, SVM will find the best hyperplane to abstractly bisect the data such that one side will dictate the pure compound is present. This will be the basis for generating the different models for each pure compound. The training of SVM is performed by using the data clustered from k-means and agglomerative clustering. The input of SVM are all the samples' bag-of-word models or clusters and the chemical labels that indicate presence of the pure compound of interest. How the data is split is described in the following sections. This method allows SVM use the knowledge of the labels to find relationship between the words and distinguish between chemicals.

Before applying machine learning the data must be split into proper sets for: training data to go into the SVM and generate multiple models for consideration; validation data that can test SVM models to choose the best one to move forward; and data not in either prior step that is blinded testing data for prediction accuracy. The training data uses the bag-of-words for a given number of samples to find a relationship between the samples.

Once the SVM is trained, the validation data set is used to determine how well the algorithm has learned from the training bag-of-words data based on an accuracy rate of 0–100%, where a higher percentage means that the algorithm has successfully identified the presence of the compound in the samples. If the accuracy rate is low, it means that the algorithm is not suitable with the dataset and a different algorithm must be used. Once a sufficiently high accuracy rate is found for the validation set, the algorithm will build a single model using the training and validation data. This model is then tested for prediction accuracy using the last "blinded" portion of the data that the model has never seen. The prediction accuracies tell us how robust the model is for chemical identification. This prediction accuracy rate is often thought to be the most realistic representation of how the algorithm will perform under real-world field conditions.

### SVM Models and Prediction Accuracies.

To start the analysis outlined the previous section the data must be split into training/ validation and test set using the pure compounds of butanone, hexanone, ethyl acetate, and 4-methyl-2-pentanone and their respective mixtures. In our previous work using PLS-DA classification, we found the inclusion of mixtures can improve the prediction accuracies because mixtures will provide different features compared to only using repeated samples of pure compounds[22]. A total of 150 samples for each compound was used a 70/30 split was implemented. In other words, 70% of the data was used for the training/validation set and 30% of the data was set aside for the test set.

The training/validation set of 105 butanone samples comprised of the following: 26 samples of pure butanone and the 2–4 mixtures (Figure 2), 26 samples of hexanone and 2–4 mixtures, 26 samples of ethyl acetate and the 2–4 mixtures, and 26 samples of 4-methyl-2-pentanone and the 2–4 mixtures. Butanone and its 2–4 mixtures are labelled as "butanone" and the other pure compounds and its 2–4 mixtures are labelled as "other". The test set is divided and labelled the same fashion except 12 samples of each pure compound and 2–4 mixtures are used.

A SVM was used, which outputs a binary decision, 0 or 1, indicating whether or not a pure compound is present in the sample once SVM is trained using the training data. In SVM, the hyperparameters needed are C and the kernel. C denotes the amount of regularization. The kernel denotes the type of kernel used in SVM which can be linear or non-linear kernels. In this study, we tried different kernels and regularization to find which parameters produced the best accuracies. The optimal parameters found were a linear kernel and a C value of 1. The supporting information gives more detail on how C was chosen. We used a standard 5-fold cross validation[41] to provide a reliable overall accuracy. Briefly, a 5-fold cross validation combines the sample of the training and validation set and randomly divides the combined dataset into 5 equal partitions. Four of the partitions will be used for to train SVM and the last partition will be used in the validation set. Each partition takes turns acting as the validation set, generating 5 different accuracies. The average of all 5 accuracies is then reported (Figure 5, blue histograms for each pure chemical). Thus, there is only one accuracy shown in the figure. Observing the accuracies, it can be seen that most of the accuracies are above 90% suggesting that SVM is a good algorithm that can be used to detect the presence of a compound within a mixture.

A final test to show that SVM is a robust model utilized the test set to observe how a truly blind dataset would look like. The training and validation set were combined to train the SVM model and the test was used to observe how well SVM could detect the presence of a particular compound. When testing new samples, the same methodology outlined in this paper is used with the exception of cluster assignment. Since there is already a bag-of-words model built, the feature vectors will use the clusters and be assigned the cluster that is closest to the feature vector. This way no bag-of-words will be skipped and will always be assigned a cluster that in the trained model. The accuracies (Figure 5, red histograms for each pure chemical) of butanone dropped 5% compared to the pure compound analysis but the accuracies of the other compounds improved by 10–20%. All the accuracies were above 85% and all the test set accuracies were approximately 5–10% lower than the validation set accuracies. The accuracies are sufficient enough to suggest that SVM is a robust algorithm in being able to detect the presence of a specific compound in an unknown sample.

In order to show that the methodology can used with many confounding compounds, a mixture of seven chemicals was also used to evaluate how well the models perform. In addition, different chemicals such as toluene and heptane were used to show that it can be applied to other compounds not just the four previously mentioned. The accuracies (Figure 6) show similar results when compared with that of four compound accuracies (Figure 5). Some possible applications of this are discussed in the application portion of this paper.

It is important to acknowledge that at this point, our dispersion plot analysis method is only binary. It will report a “yes” or “no” presence of a target chemical in mixtures when a query is performed. However, in the future, it is possible to expand our method to also report approximate quantities of chemicals present. Future work can address this by modeling the dispersion plot changes that occur when concentrations of discrete chemicals increase/decrease in mixtures. It is also important to continue to increase the accuracies of the chemical matches in the dispersion plots. Traditionally trained chemists will note the analogous visual similarity of the bag-of-words (BoW) models (Figure 4) to mass spectrometry mass-to-charge ratio (*m*/*z*) plots. Although the physical mechanisms to generate these plots are very different, the mass spectrometry field has witnessed great advances in deconvolution and chemical identification via *m*/*z* library matching algorithms. We expect equivalent advances in matching our BoW models can also improve dispersion plot analysis.

### Application.

One possible application for this methodology is to detect a specific compound of interest from a mixture on real-time mobile devices. To do this, the machine learning portion of developing SVM models can be trained on a more powerful machine such as a desktop. The models can be put onto a mobile device such as a tablet and data from a portable DMS can be fed to the tablet. The tablet can then be used to detect if chemicals of interest are present in an environment. Our results include several chemicals such as heptane and toluene to show the application importance. Heptane is an important paraffin in fossil fuel industry and detection of heptane during combustion is necessary to improve the efficiency of current combustors and reduce the environmental polluting species[42].Toluene represents aromatic group VOCs that include chemicals such as Benzene and Xylene. These are widely used organic solvents for paints, adhesives, detergents, dyes, and preservatives. Trace detection of these chemicals are vital as they are some of the most hazardous chemicals to the human health and pollutants to the environment[43]. These applications can improve the quality of human life by informing people about possible hazardous chemicals in the ambient air. In terms of software run time, the longest runtime on the real-time device would be the feature detection and extraction. This run time of this is around 8.39 seconds for approximately 50 samples. Once the features vectors are extracted then the centroids from clustering algorithms will be used to generate the bag of words for each sample. Then SVM will be used to detect if the compound is present.

## CONCLUSION

This paper uses specific advances in computer vision, natural language processing and machine learning to detect the presence of specific chemicals from a dispersion plot of a complex mixture. Computer vision was used to identify important features such peak ions, corners, and points-of-interests in DMS dispersion plots. Clustering was used to group similar features together to generate the bag-of-words model for each sample. Support vector machine was used to find a relationship between the bag-of-words model, and we found that SVM is a robust algorithm in predicting the presence of a specific compound. It is worth noting that any supervised machine learning (i.e. classification) model can be applied at this step. SVM is a known, reliable classification method commonly used within computer vision and bag-of-words classification, making it a good choice for initial tests. Future work might include comparing different supervised machine learning algorithms at this step, and assessing the various methods for improved accuracies. Another future topic to explore is feature engineering, where standardizing features are artificially imposed and/or selected from the dispersion plots using chemistry knowledge. This can help to improve accuracies, although great care has to be taken to ensure the models of the data are not over trained.

Overall, our method of using computer vision and machine learning can be applied to any data collected by DMS and is applicable in many areas of research. The computer vision and machine learning algorithms used are also computationally fast and have the potential of being applied to portable platforms in the future.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## ACKNOWLEDGEMENTS

Partial support was provided by: NIH award U01 EB0220003-01 (CED, NJK); NIH award UG3-OD023365 (CED, NJK); the NIH National Center for Advancing Translational Sciences (NCATS) through grant #UL1 TR000002 (CED, NJK); and NIH award 1P30ES023513-01A1 (CED, NJK). Student support was partially provided by the US Department of Veterans Affairs, Post-9/11 GI-Bill (DJP), and NSF award 1343479 Veteran's Research Supplement (DJP). The authors would like to thank Sean Pham for assistance with paper revisions. The contents of this manuscript are solely the responsibility of the authors and do not necessarily represent the official views of the funding agencies.

## REFERENCES

1. Katajamaa M; Orešič M, Data processing for mass spectrometry-based metabolomics. Journal of Chromatography A 2007, 1158 (1), 318–328. [PubMed: 17466315]
2. DeHaven CD; Evans AM; Dai H; Lawton KA, Organization of GC/MS and LC/MS metabolomics data into chemical libraries. Journal of Cheminformatics 2010, 2 (1), 9. [PubMed: 20955607]
3. Shnayderman M; Mansfield B; Yip P; Clark HA; Krebs MD; Cohen SJ; Zeskind JE; Ryan ET; Dorkin HL; Callahan MV; Stair TO; Gelfand JA; Gill CJ; Hitt B; Davis CE, Species-specific bacteria identification using differential mobility spectrometry and bioinformatics pattern recognition. Analytical Chemistry 2005, 77 (18), 5930–5937. [PubMed: 16159124]
4. Nakhleh MK; Amal H; Jeries R; Broza YY; Aboud M; Gharra A; Ivgi H; Khatib S; Badarneh S; Har-Shai L; Glass-Marmor L; Lejbkowicz I; Miller A; Badarny S; Winer R; Finberg J; Cohen-Kaminsky S; Perros F; Montani D; Girerd B; Garcia G; Simonneau G; Nakhoul F; Baram S; Salim R; Hakim M; Gruber M; Ronen O; Marshak T; Doweck I; Nativ O; Bahouth Z; Shi D.-y.; Zhang W; Hua Q.-l; Pan Y.-y.; Tao L; Liu H; Karban A; Koifman E; Rainis T; Skapars R; Sivins A; Ancans G; Liepniece-Karele I; Kikuste I; Lasina I; Tolmanis I; Johnson D; Millstone SZ; Fulton J; Wells JW; Wilf LH; Humbert M; Leja M; Peled N; Haick H, Diagnosis and Classification of 17 Diseases from 1404 Subjects via Pattern Analysis of Exhaled Molecules. ACS Nano 2017, 11 (1), 112–125. [PubMed: 28000444]
5. Aksenov AA; Pasamontes A; Peirano DJ; Zhao W; Dandekar AM; Fiehn O; Ehsani R; Davis CE, Detection of Huanglongbing Disease Using Differential Mobility Spectrometry. Analytical Chemistry 2014, 86 (5), 2481–2488. [PubMed: 24484549]
6. Miller RA; Zapata A; Nazarov EG; Krylov E; Eiceman GA, High performance micromachined planar field-asymmetric ion mobility spectrometers for chemical and biological compound detection. Mater Res Soc Symp P 2002, 729, 139–147.
7. Eiceman GA; Krylov EV; Nazarov EG; Miller RA, Separation of ions from explosives in differential mobility spectrometry by vapor-modified drift gas. Analytical Chemistry 2004, 76 (17), 4937–4944. [PubMed: 15373426]
8. Pasupuleti D; Eiceman GA; Pierce KM, Classification of biodiesel and fuel blends using gas chromatography - differential mobility spectrometry with cluster analysis and isolation of C18:3 me by dual ion filtering. Talanta 2016, 155, 278–88. [PubMed: 27216685]
9. Eiceman GA; Young D; Schmidt H; Rodriguez JE; Baumbach JI; Vautz W; Lake DA; Johnston MV, Ion mobility spectrometry of gas-phase ions from laser ablation of solids in air at ambient pressure. Appl Spectrosc 2007, 61 (10), 1076–83. [PubMed: 17958958]
10. Cumeras R; Figueras E; Davis CE; Baumbach JI; Gracia I, Review on Ion Mobility Spectrometry. Part 1: current instrumentation. Analyst 2015, 140 (5), 1376–1390. [PubMed: 25465076]
11. Cumeras R; Figueras E; Davis CE; Baumbach JI; Gracia I, Review on Ion Mobility Spectrometry. Part 2: hyphenated methods and effects of experimental parameters. Analyst 2015, 140 (5), 1391–1410. [PubMed: 25465248]
12. Krebs MD; Zapata AM; Nazarov EG; Miller RA; Costa IS; Sonenshein AL; Davis CE, Detection of biological and chemical agents using differential mobility spectrometry (DMS) technology. IEEE Sensors Journal 2005, 5 (4), 696–703.
13. Kolakowski BM; Mester Z, Review of applications of high-field asymmetric waveform ion mobility spectrometry (FAIMS) and differential mobility spectrometry (DMS). Analyst 2007, 132 (9), 842–864. [PubMed: 17710259]
14. Zappa D; Galstyan V; Kaur N; Munasinghe Arachchige HMM; Sisman O; Comini E, "Metal oxide -based heterostructures for gas sensors"- A review. Anal Chim Acta 2018, 1039, 1–23. [PubMed: 30322540]
15. Burgues J; Marco S, Low Power Operation of Temperature-Modulated Metal Oxide Semiconductor Gas Sensors. Sensors (Basel) 2018, 18 (2).
16. Miller RA; Eiceman GA; Nazarov EG; King AT, A novel micromachined high-field asymmetric waveform-ion mobility spectrometer. Sensor Actuat B-Chem 2000, 67 (3), 300–306.
17. Eiceman GA; Tadjikov B; Krylov E; Nazarov EG; Miller RA; Westbrook J; Funk P, Miniature radio-frequency mobility analyzer as a gas chromatographic detector for oxygen-containing

volatile organic compounds, pheromones and other insect attractants. Journal of Chromatography A 2001, 917 (1–2), 205–217. [PubMed: 11403471]

18. Miller RA; Nazarov EG; Eiceman GA; King AT, A MEMS radio-frequency ion mobility spectrometer for chemical vapor detection. Sensor Actuat a-Phys 2001, 91 (3), 301–312.
19. Krylov EV; Coy SL; Vandermey J; Schneider BB; Covey TR; Nazarov EG, Selection and generation of waveforms for differential mobility spectrometry. Review of Scientific Instruments 2010, 81 (2), 024101. [PubMed: 20192506]
20. KREBS M; ZAPATA A; NAZAROV E; al., e., Detection of biological and chemical agents using differential mobility spectrometry (DMS) technology. IEEE SENSORS JOURNAL 2005, 5 (4), 696–703.
21. Peirano DJ; Pasamontes A; Davis CE, Supervised Semi-Automated Data Analysis Software for Gas Chromatography / Differential Mobility Spectrometry (GC/DMS) Metabolomics Applications. Int J Ion Mobil Spectrom 2016, 19 (2), 155–166. [PubMed: 27799845]
22. Rajapakse MY; Borras E; Yeap D; Peirano DJ; Kenyon NJ; Davis CE, Automated chemical identification and library building using dispersion plots for differential mobility spectrometry. Anal Methods-Uk 2018, 10 (35), 4339–4349.
23. Eiceman GA; Krylov EV; Krylova NS; Nazarov EG; Miller RA, Separation of ions from explosives in differential mobility spectrometry by vapor-modified drift gas. Anal Chem 2004, 76 (17), 4937–44. [PubMed: 15373426]
24. Kanan C; Cottrell GW, Color-to-grayscale: does the method matter in image recognition? PLoS One 2012, 7 (1), e29740. [PubMed: 22253768]
25. Shajudeen PMS; Righetti R, Spine surface detection from local phase-symmetry enhanced ridges in ultrasound images. Med Phys 2017, 44 (11), 5755–5767. [PubMed: 28786479]
26. Rouco J; Azevedo E; Campilho A, Automatic Lumen Detection on Longitudinal Ultrasound B-Mode Images of the Carotid Using Phase Symmetry. Sensors (Basel) 2016, 16 (3).
27. Kovesi P, Phase congruency: a low-level image invariant. Psychol Res 2000, 64 (2), 136–48. [PubMed: 11195306]
28. Aghito SM; Forchhammer S, Efficient coding of shape and transparency for video objects. IEEE Trans Image Process 2007, 16 (9), 2234–44. [PubMed: 17784597]
29. Gu L; Xu J; Peters TM, Novel multistage three-dimensional medical image segmentation: methodology and validation. IEEE Trans Inf Technol Biomed 2006, 10 (4), 740–8. [PubMed: 17044408]
30. Shao L; Liu L; Li G, Optical intrinsically fuzzy mathematical morphology for gray-scale image processing. Appl Opt 1996, 35 (17), 3109–16. [PubMed: 21102688]
31. Rosten E; Porter R; Drummond T, Faster and better: a machine learning approach to corner detection. IEEE Trans Pattern Anal Mach Intell 2010, 32 (1), 105–19. [PubMed: 19926902]
32. Naqvi SAG; Zafar HMF; Haq I, Hard exudates referral system in eye fundus utilizing speeded up robust features. Int J Ophthalmol 2017, 10 (7), 1171–1174. [PubMed: 28730125]
33. Sun M; Yang S; Jiang J; Wang Q, Detection of Perlger-Huet anomaly based on augmented fast marching method and speeded up robust features. Biomed Mater Eng 2015, 26 Suppl 1, S1241–8. [PubMed: 26405883]
34. Stanciu SG; Hristu R; Stanciu GA, Influence of confocal scanning laser microscopy specific acquisition parameters on the detection and matching of speeded-up robust features. Ultramicroscopy 2011, 111 (5), 364–74. [PubMed: 21349249]
35. Amari S; Wu S, Improving support vector machine classifiers by modifying kernel functions. Neural Netw 1999, 12 (6), 783–789. [PubMed: 12662656]
36. Kwok JY, Moderating the outputs of support vector machine classifiers. IEEE Trans Neural Netw 1999, 10 (5), 1018–31. [PubMed: 18252604]
37. Keerthi SS; Shevade SK; Bhattacharyya C; Murthy KK, A fast iterative nearest point algorithm for support vector machine classifier design. IEEE Trans Neural Netw 2000, 11 (1), 124–36. [PubMed: 18249745]
38. Kolakowski BM; Mester Z, Review of applications of high-field asymmetric waveform ion mobility spectrometry (FAIMS) and differential mobility spectrometry (DMS). Analyst 2007, 132 (9), 842–64. [PubMed: 17710259]

39. Bay H; Ess A; Tuytelaars T; Van Gool L, Speeded-up robust features (SURF). Computer vision and image understanding 2008, 110 (3), 346–359.

40. Gowda KC; Krishna G. J. P. r., Agglomerative clustering using the concept of mutual nearest neighbourhood. 1978, 10 (2), 105–112.

41. Meijer RJ; Goeman JJ, Efficient approximate k-fold and leave-one-out cross-validation for ridge regression. Biom J 2013, 55 (2), 141–55. [PubMed: 23348970]

42. Ding J; Zhang L; Zhang Y; Han K-L, A reactive molecular dynamics study of n-heptane pyrolysis at high temperature. The Journal of Physical Chemistry A 2013, 117 (16), 3266–3278. [PubMed: 23544797]

43. Mirzaei A; Kim J-H; Kim HW; Kim SS, Resistive-based gas sensors for detection of benzene, toluene and xylene (BTX) gases: a review. Journal of Materials Chemistry C 2018, 6 (16), 4342–4370.

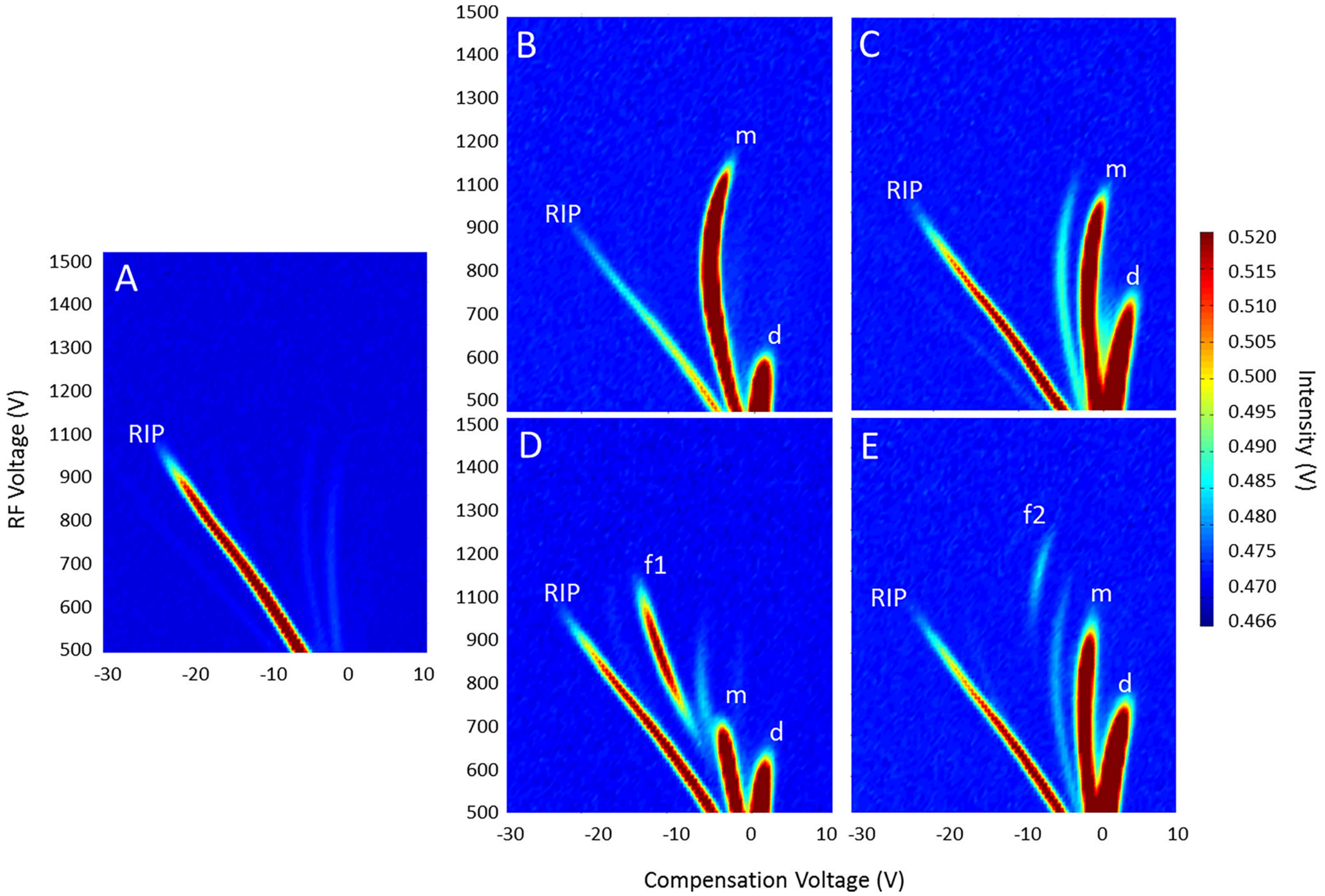


**Figure 1.**

Example DMS dispersion plot of pure compounds: (A) RIP alone, (B) 2-butanone, (C) 2-hexanone, (D) ethyl acetate, (E) 4-methyl 2-pentanone. In each image, the monomer (m) and dimer (d) are noted. Fragment ion peaks are also noted (f1 and f2).

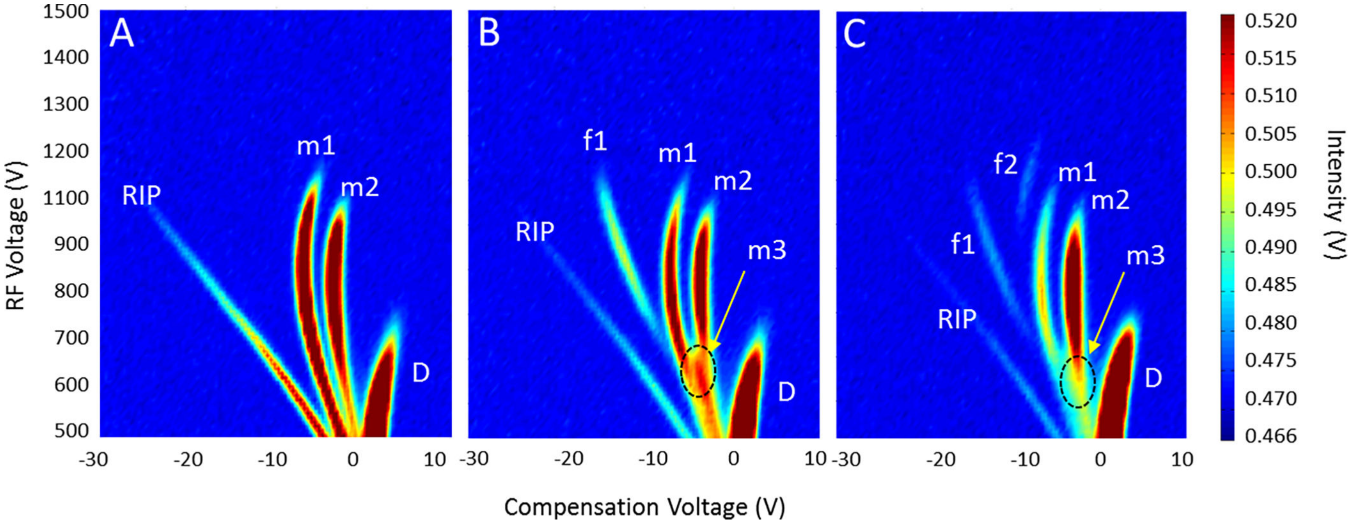


**Figure 2.**

DMS dispersion plot of chemical mixture of the following: (A) 2-butanone and 2-hexanone; (B) 2-butanone, 2-hexanone, and ethyl acetate; (C) 2-butanone, 2-hexanone, ethyl acetate, and 4-methyl 2-pentanone. Monomers are as follows: m1 (2-butanone), m2 (2-hexanone), m3 (ethyl acetate), m2 (4-methyl 2-pentanone overlaps with 2-hexanone).

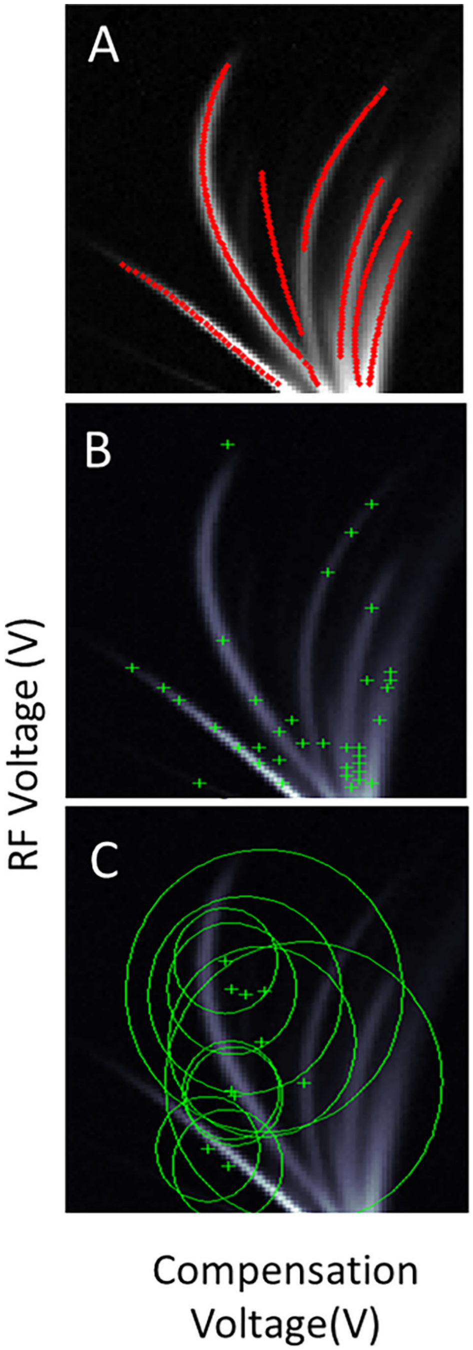


**Figure 3.**

Computer vision algorithms are used to detect the following: (A) peak ion detection outlined by red curve, (B) corner detection outlined by green crosses, and (C) points of interest outlined by green crosses and circles.

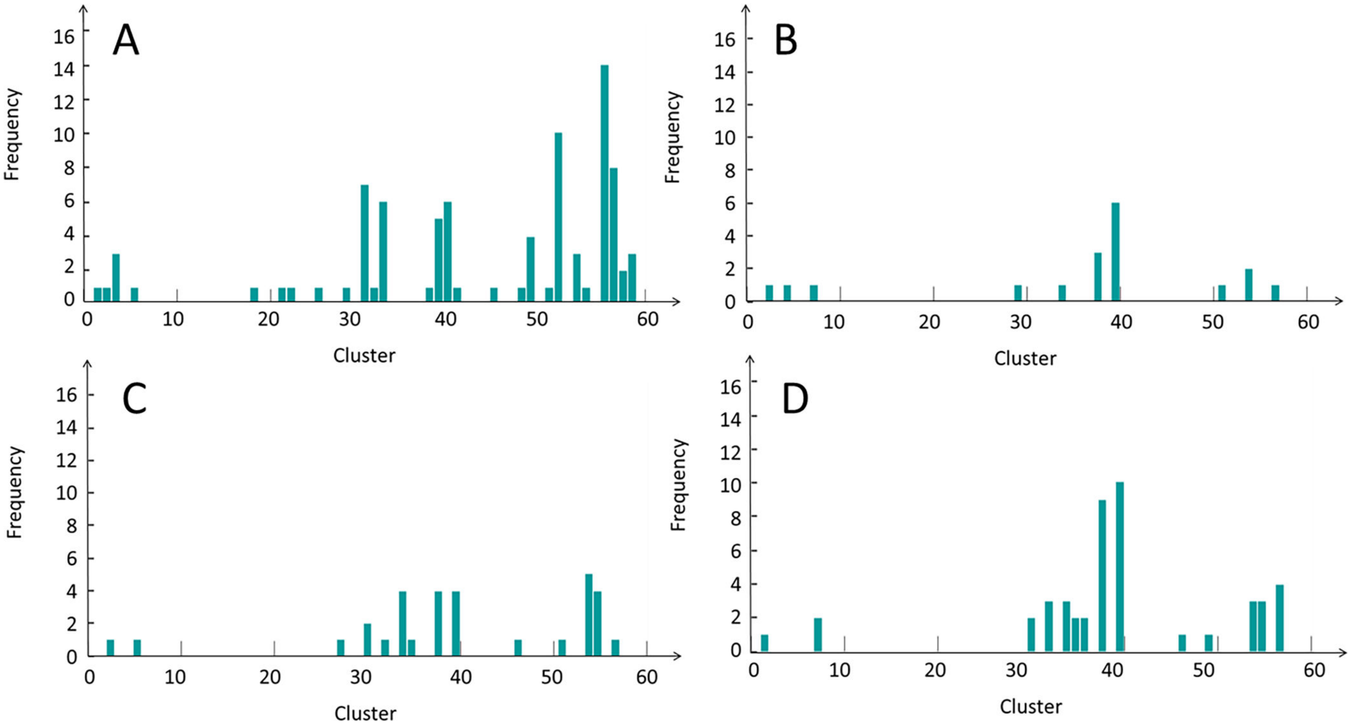


**Figure 4.**

The bag-of-visual-words models are shown for example dispersion plots of the following pure compounds: (A) 2-butanone, (B) 2-hexanone, (C) ethyl acetate, (D) 4-methyl 2-pentanone.

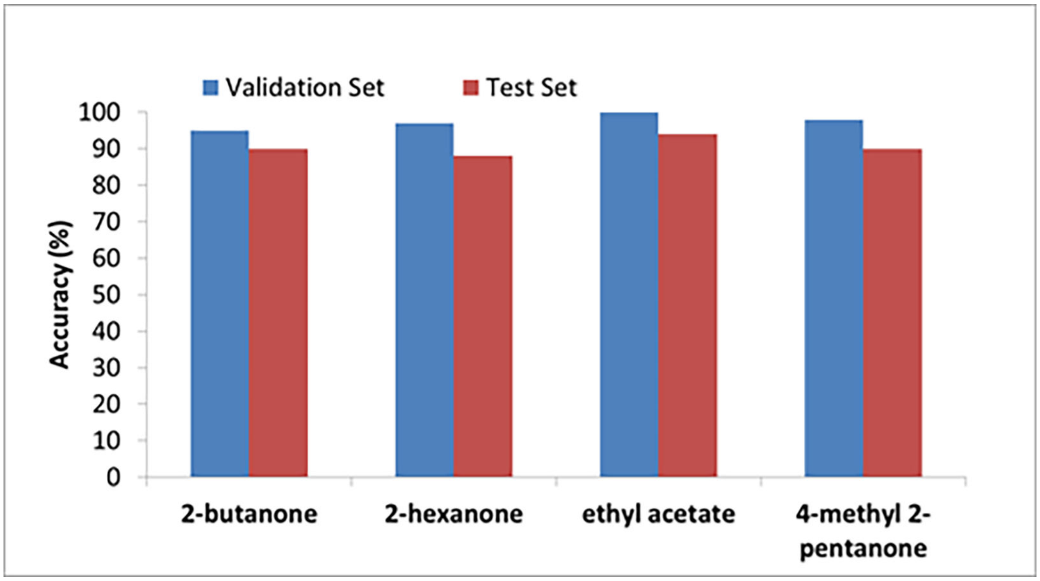


**Figure 5.**

Accuracy for validation and test sets when the pure compounds and respective mixtures are included in the training, validation, and test sets.

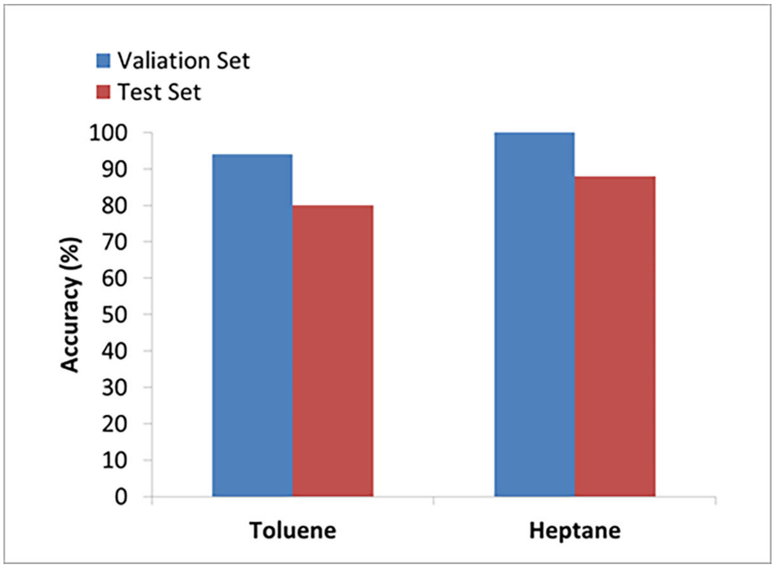


**Figure 6.**

Accuracy for validation and test sets when the pure compounds and seven mixtures are included in the training, validation, and test sets.