Indoor Localisation using Aroma Fingerprints: Comparing Nearest Neighbour Classification Accuracy using Different Distance Measures

1st Georgy Minaev
Signal Processing
Tampere University of Technology
Tampere, Finland
georgy.minaev@tut.fi

2nd Philipp Müller
Pervasive Computing
Tampere University of Technology
Tampere, Finland
philipp.muller@tut.fi

3rd Ari Visa
Signal Processing
Tampere University of Technology
Tampere, Finland
ari.visa@tut.fi

4th Robert Piché
Automation and Hydraulic Engineering
Tampere University of Technology
Tampere, Finland
robert.piche@tut.fi

***Abstract*—Measurements from an ion mobility spectrometry electronic nose (eNose) can be used for distinguishing different rooms in indoor localisation. An earlier study showed that the Nearest Neighbour classifier with Euclidean distance for features provides reasonable accuracy under certain conditions. In this paper 66 alternative distance measures are compared to the Euclidean distance and principal component analysis (PCA) is applied to the data. PCA shows that the measurements on the various channels of the eNose are correlated and that using principal components 1, 2 and 4 increases the accuracy considerably. Furthermore, the experiments revealed three Pareto optimal distance measures that reduce the misclassification rate between 9-10% while using only 82-88% of the search time compared with Euclidean distance.**

***Index Terms*—Indoor localisation, K Nearest Neighbours, Electronic nose, Ion mobility spectrometry**

I. Introduction

Indoor localisation has received much attention over the last decade. Besides various radio signals, such as cellular networks, wireless local area networks (WLAN), ultra-wideband (UWB), Bluetooth and Bluetooth low energy (BLE), inertial measurement units (IMUs), laser range scanners, floor maps, and magnetic fields have been studied for localisation in the absence of satnav [1], [2]. A source of measurement that just has been started to be studied for localisation are electronic noses (eNoses). To the authors' knowledge, [3] is the only article that has studied the use of eNose measurements for localisation. Here localisation means localising the user/user device, which differs from applications in which eNoses are used for localising the source of an odour (see e.g. [14], [15])

Electronic noses are used in artificial olfaction, for detecting and classifying various gases. For that purpose they mimic the biological sense of smell and its communication with a biological brain [4], using a sensor array, a signal-processing unit, a reference database, and pattern recognition software [5]. Different eNoses exist, which use different sensor types (see e.g. [5]). In [3] an eNose using an ion mobility spectrometry (IMS) sensor was used for localisation. The main reason for choosing an IMS-based eNose was that the sensor element is a metal electrode that does not age. Therefore, the signal drift it experiences is mainly due to environmental changes.

In [3] the K nearest neighbours (KNN) method with different K-values was used for localisation based on IMS measurements. For measuring the closeness between training and test samples the Euclidean distance was used. In this paper, 66 alternative distance measures are analysed and compared with the Euclidean distance using the data from [3] and the nearest neighbour classifier. The aim of the extensive analysis is to find Pareto optimal distance measures, i.e. measures that achieve better localisation accuracy and/or reduce the search time for the nearest neighbour compared with the Euclidean distance. Furthermore, the principal component analysis (PCA) method is used to shorten the classification time and to remove potential correlation from the data. The localisation accuracy when using different sets of principal components is studied.

Our hypothesis is that it is possible to improve indoor localisation accuracy using Aroma Fingerprints, in comparison with [3] by data preprocessing and improved classification methods. Different distance measures with Principal Component analyses are tested in order to find the one that is best in terms of evaluation time and accuracy.

This paper is organised as follows. The ChemPro100i

eNose is described in Section II. Nearest neighbour classification and PCA and its implementation are explained in Section III. The eNose data is described in Section IV. Section V describes the tests, and shows and discusses their results. Finally, Section VI concludes the paper and gives an outlook.

II. ChemPro100i eNose

In this paper measurements from a ChemPro100i [6] eNose from Environics are used for determining in which room the user is. The ChemPro100i has an IMS sensor that ionises the incoming air and separates the resulting ions based on their velocity. Due to differences in their molecular weight, charge and geometry between compounds the mobility of various ions differs [7]. The ChemPro100i measures the ions as a current with seven separate electrode pairs, and the electric field is continuously switched between positive and negative polarities. Thus, it generates a 14-dimensional "fingerprint" (seven variables for positive and seven for negative electrodes) of the air at a specific location, which is used as a measurement in this paper.

The major drawbacks of the ChemPro100i as a potential mass market device are its size and price. However, for the localisation only its measurement cell would be needed, whose dimensions are around 4 cm by 2 cm by 1 cm (i.e. approximately the size of a match box) and which contributes only a small part to the price of the ChemPro100i. Alternative IMS chips have recently appeared, for example by Owlstone, whose field asymmetric ion mobility spectrometer is "fabricated on a single microchip with dimensions under a centimetre" [8].

III. IMS fingerprint-based localisation

A. Nearest neighbour classification

For localisation only the nearest neighbour (NN) classifier is used in this paper, because in [3] it yielded the same performance as the KNN with $K = \{3, 5, 7\}$. The idea behind NN is to find the training fingerprint $\mathbf{x}_\text{i} = [x_{\text{i},1} \ \ldots \ x_{\text{i},14}]$ from a set of N fingerprints that is closest to a test fingerprint $\mathbf{x}_\text{test}$. For the training fingerprints the location at which they were taken is known, while for the test fingerprint the location is the unknown parameter that has to be determined. The closeness of the test fingerprint and any training fingerprint can be measured by various distance measures. In this paper 67 different measures are analysed and compared.

B. Principal Component Analyses

One drawback of KNN classifiers is that they can be fooled by irrelevant features. Furthermore, the measurements of the 14 electrodes in the used eNose might be correlated. In order to address these two drawbacks, in this paper principal component analysis (PCA) is used. PCA converts the set of 14 potentially correlated variables into a lower-dimensional set of linearly uncorrelated variables [9, p. 580].

For the training data $\mathbf{X} = \{\mathbf{x}_1, .., \mathbf{x}_\text{N}\}$ with $d = 14$ dimensions PCA works as follow [9, p. 568]:

1) Compute d-dimensional mean vector $\boldsymbol{\mu}$ and d-by-d covariance matrix $\mathbf{C}$ of data set $\mathbf{X}$.
2) Compute eigenvectors and eigenvalues of $\mathbf{C}$, and sort them according to decreasing eigenvalues.
3) Choose a subset of these eigenvalues, for example, the first k eigenvalues and form d-by-k matrix $\mathbf{A}$ (k eigenvectors as columns of $\mathbf{A}$).
4) PCA-transformed data $\mathbf{Y} = \{\mathbf{y}_1, .., \mathbf{y}_\text{N}\}$ is now defined as $\mathbf{y}_i = \mathbf{A}^\text{T}(\mathbf{x}_i - \boldsymbol{\mu})$, where each $\mathbf{y_i}$ has $\mathbf{k}$ variables.

The choice of the k principal components affects how much of the total variance in $\mathbf{X}$ is explained by the transformed data $\mathbf{Y}$. In principle all d principal components could be used. However, in general a small subset is chosen that explains at least 95% or 99% of the total variance in the training data.

In order to find the nearest neighbour of $\mathbf{x}_\text{test}$ the test sample has to be transformed into the same format as the training data. This can be achieved using mean vector $\boldsymbol{\mu}$ and matrix $\mathbf{A}$. The PCA-transformed test sample is then defined as $\mathbf{y}_\text{test} = \mathbf{A}^\text{T}(\mathbf{x}_\text{test} - \boldsymbol{\mu})$.

IV. Data

The eNose data was collected from seven different indoor locations at Tampere University of Technology, Finland, in May 2017. Table I summarises the locations and their types [3]. There was one office room for four people (location 1), a coffee room (location 2), a corridor connecting two buildings (location 5), and four large open areas (locations 3, 4, 6, 7). Locations 6 and 7 were in close proximity to cafeterias. Fig. 1 shows approximate positions of all seven locations.

At each location two measurements of 10 minutes' duration were made. The first set was collected during a Saturday, when the buildings were almost empty and the cafeterias were closed, and the second set was collected either two or three days later, when staff and students of the university were present and lunch buffets were on display at the two cafeterias. The measurement frequency was 1 Hz. Table I shows the sizes of both measurement sets for all seven locations.

TABLE I
Measurement descriptions

id	location type	# measurements		total
		empty	crowded	
1	office room	629	618	1247
2	coffee room	643	631	1247
3	open space	616	618	1234
4	open space	609	614	1223
5	corridor	630	646	1276
6	open (cafeteria) space	608	637	1245
7	open (cafeteria) space	626	611	1237
$\sum$		4361	4375	8736

V. Experiments

In the experiments both IMS data and PCA-transformed data with varying sets of principal components were used. For the experiments the data was split into two sets: training set

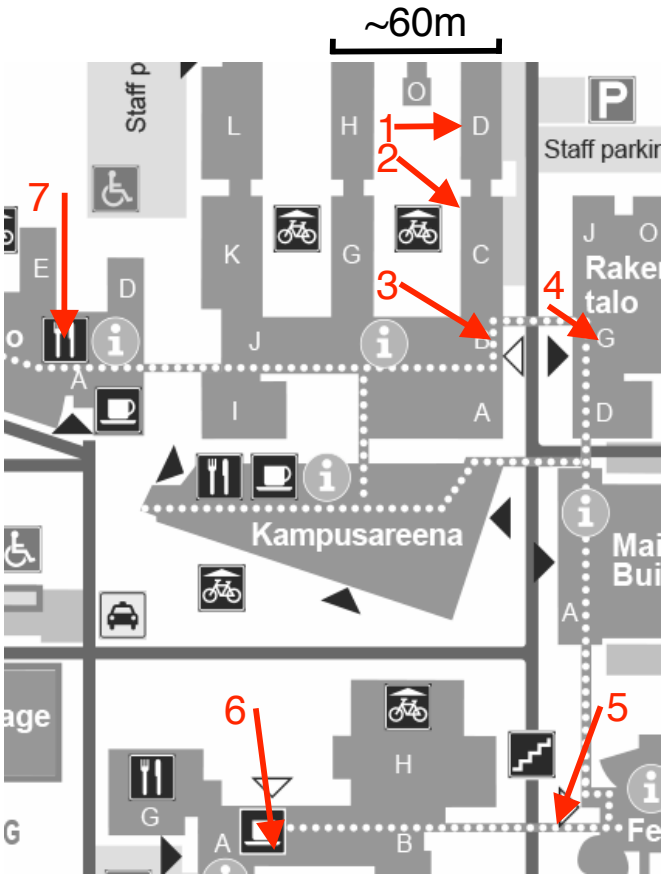

Fig. 1. Representation of the campus with approximate positions of all seven locations.

and test set. In a first experiment both sets contained measurements from each room in empty and crowded conditions. In this experiment classification accuracies of close to 100% were achieved for all tested distance measures, which was in line with the findings in [3]. However, for the experiments presented in this section, in order to evaluate the potential of each distance measure, the training contained only measurements from empty rooms and the test set contained only data from crowded rooms (see Table I).

A total of 67 distance measures, including the Euclidean distance as reference distance measure, were studied. The aim was to minimise misclassification error and evaluation time (aka search time).

A. NN using Euclidean distance

First it was studied if PCA reduces the misclassification error when using the Euclidean distance as distance measure. Fig. 2 shows the misclassification errors when using standardised IMS data (see [3] for details), as reference, and various subsets of principal components. All tests yielded similar misclassification errors, except for using only the first principal component. This indicates that data from the

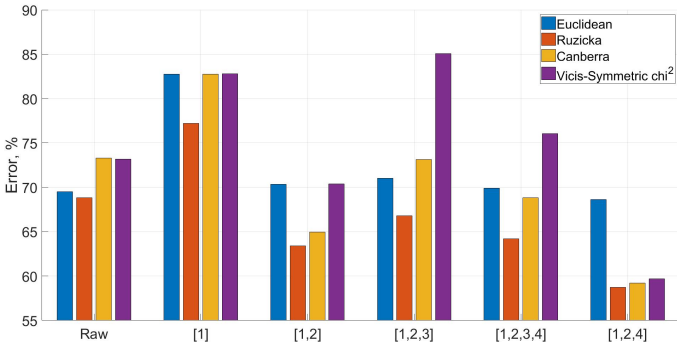

Fig. 2. Errors of room's correct identification. The x axes labels are 'raw' and list of used PCA components. Bars are listed in order: Euclidean, Ruzicka, Canberra and Vicis-Symmentric χ^2.

IMS electrodes was indeed correlated and that the evaluation time could be reduced significantly by transforming training and test data using PCA, because only two to four principal components are enough to achieve similar performance than when using the 14-dimensional IMS data. Therefore, in the following test localisation using IMS data was compared with localisation using two sets of principal components: $\{1, 2, 3, 4\}$ and $\{1, 2, 4\}$.

The large misclassification rates are in line with the results from [3], and might be caused by the differences in the environmental conditions in which training and test data were collected. As mentioned in Section II, the IMS measurement depends on the mobility of the ions in the air sample, which in turn depends on humidity and temperature, but also barometric pressure, and air currents [10, pp. 250 ff.].

B. NN using different distances

The goal of the second test was to find the best distance measures with respect to classification error and evaluation time. In total 67 distance measures were studied. Due to space constraints, here only results from three Pareto optimal distance measures are shown and compared with the results of Euclidean distance. The 67 distance measures are described in [11] and [12], and the full list of distances can be found in [13].

All experiments with raw data and with principal components $\{1, 2, 3, 4\}$ and $\{1, 2, 4\}$ yielded different sets of Pareto optimal distance measures. The lowest misclassification errors were obtained using principal components 1, 2 and 4, as can be seen in Fig. 2. Therefore, the three Pareto optimal distance measures based on the test with principal components 1, 2 and 4 were chosen. Euclidean distance was used for comparison. The formulas of the three Pareto optimal distance measures and Euclidean are shown in Table II. The evaluation time of squared Euclidean distance is slightly less than Euclidean evaluation time with the same error. The Euclidean distance is used for simplicity.

Fig. 2 shows the misclassification errors of the three Pareto optimal distance measures and of the Euclidean distance. Based only on the errors the Ruzicka distance is the best

TABLE II
DISTANCE MEASURES

id	*name*	*d(P,Q)*	*Source*
1.	Euclidean, $p = 2$, $r = 2$	$(\sum \lvert P_i - Q_i \rvert^p)^{1/r}$	[11, chapter 17.2]
25.	Ruzicka, aka Soergel, Tanimoto	$\sum \lvert P_i - Q_i \rvert / \sum \max(P_i, Q_i)$	[11, chapter 17.1]
30.	Canberra	$\sum(\lvert P_i - Q_i \rvert / \lvert P_i \rvert + \lvert Q_i \rvert)$	[11, chapter 17.1]
60.	Vicis-Symmetric χ^2	$\sum \frac{(P_i - Q_i)^2}{\min(P_i, Q_i)^2}$	[12]

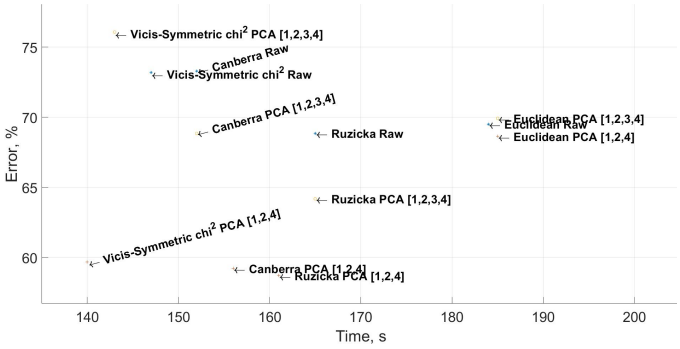

Fig. 3. Pareto optimal with PCA components $\{1, 2, 3, 4\}$, $\{1, 2, 4\}$ and Raw data.

choice. However, for practical application also the evaluation time has to be taken into account. Therefore, Fig. 3 shows misclassification errors and evaluation times of the three Pareto optimal measures and Euclidean distance when using IMS data, first four principal components, and principal components 1, 2 and 4. The figure shows that Ruzicka with principal components 1, 2 and 4 yielded the lowest misclassification rate but not the fastest evaluation time, which was achieved by Canberra distance using also principal components 1, 2, and 4. These two combinations of distance measures and choice of principal components are therefore Pareto optimal. The Euclidean distance showed middling accuracy and the worst evaluation time independent on the choice of data.

An interesting observation from Fig. 2 was that the misclassification errors for all four distance measures increased when using the third principal component in addition to the first two components. Therefore, the third component was removed for the last test (see rightmost group of bars in Fig. 2), which yielded lower misclassification rates than using the first four principal components.

Finally, Fig. 4, Fig. 5, Fig. 6 and Fig. 7 show the confusion matrices for Euclidean, Ruzicka, Canberra and Vicis-Symmetric χ^2 distances when using data on principal components 1, 2 and 4. Over all seven rooms Ruzicka

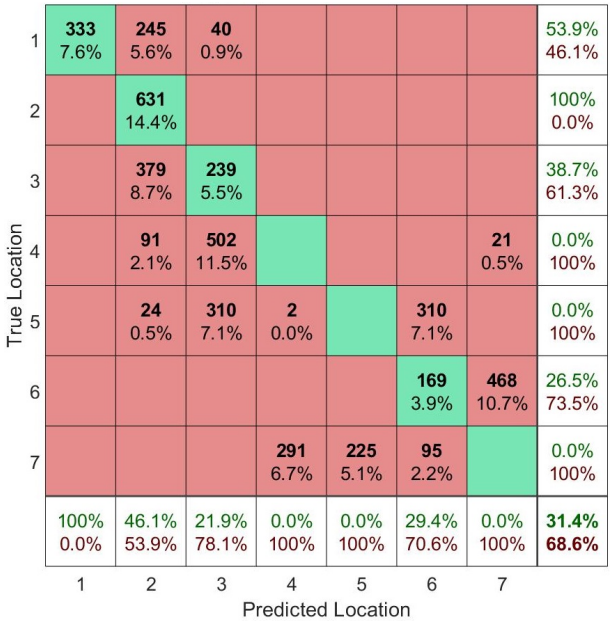

Fig. 4. Confusion matrix using PCA components $\{1, 2, 4\}$ and Euclidean distance. The number of observations and the percentage of the total number of observations are shown in each cell.

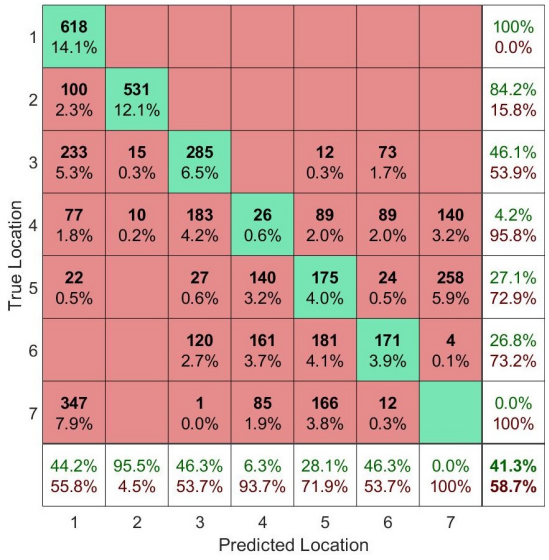

Fig. 5. Confusion matrix using PCA components $\{1, 2, 4\}$ and Ruzicka distance. The number of observations and the percentage of the total number of observations are shown in each cell.

distance improves the classification accuracy by ≈10% while using only ≈88% of the evaluation time compared to the Euclidean distance. The Canberra distance improved the classification accuracy by ≈9% while using only ≈84% of the evaluation time compared to the Euclidean distance. The Vicis-Symmetric χ^2 distance improved the classification accuracy by ≈8.9% while using only ≈75% of the evaluation time compared to the Euclidean distance.

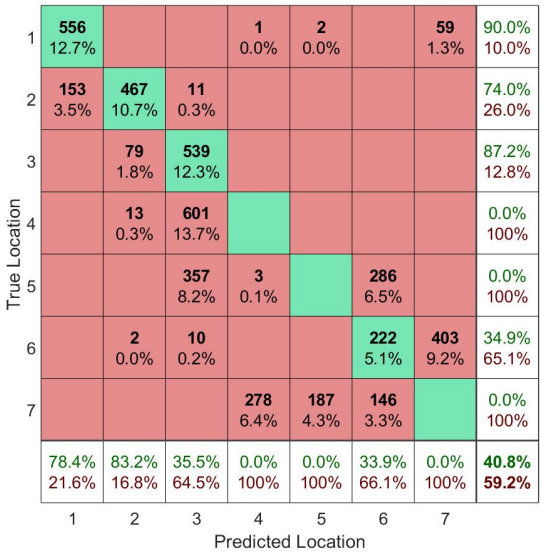

Fig. 6. Confusion matrix using PCA components $\{1, 2, 4\}$ and Canberra distance. The number of observations and the percentage of the total number of observations are shown in each cell.

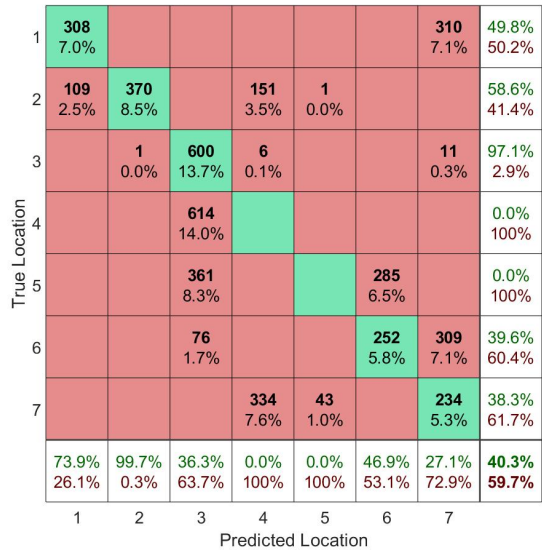

Fig. 7. Confusion matrix using PCA components $\{1, 2, 4\}$ and Vicis-Symmetric χ^2 distance. The number of observations and the percentage of the total number of observations are shown in each cell.

VI. Conclusion and Outlook

Localisation based on measurements from an electronic nose has only recently attracted some attention. In the first, to the authors' knowledge, paper dealing with eNose-based localisation KNN classifiers using Euclidean distance as distance measure were tested [3]. In this paper, 66 alternative distance measures were tested on the data from [3] using a NN classifier (i.e. a KNN with $K = 1$). Furthermore, principal component analysis was applied to the data.

The hypothesis was proved with the experiments. It is possible to improve indoor localisation accuracy using Aroma Fingerprints, in comparison with [3]. The Pareto optimal distance measures are found and shown in the experiments.

The results of PCA showed that the measurement channels from the ion mobility spectrometry eNose are correlated, which means that localisation should rely on PCA-transformed data to remove these correlations and at the same improve the evaluation time of search algorithm. Lowest misclassification errors were achieved when using first, second and fourth principal components. The analysis of the 67 distance measures using this subset of principal components yielded three Pareto optimal distance measures. These three measures, Ruzicka, Canberra and Vicis Symmetric χ^2 achieved better evaluation speed and accuracy than Euclidean distance. The three distance measures are non-dominated points (Pareto-optimal). Ruzicka $[1, 2, 4]$ has lower error than Canberra $[1, 2, 4]$ but higher evaluation time. Canberra $[1, 2, 4]$ has lower error than Vicis Symmetric χ^2 $[1, 2, 4]$ but higher evaluation time. All three distance measures dominate Euclidean distance in all experiments and present the Pareto-optimal points, the Pareto Front.

The experiments were repeated for some distances with KNN classifiers using $K = \{3, 5, 7, 9\}$. However, the change in K had no significant influence on the classification accuracy because for all tested Ks all nearest neighbours were from the same class (aka room).

More experiments with PCA will be done in the future. For example, the reason for the beneficial effect of removing the third principal component on classification accuracy will be studied in more detail. In addition, the effect of other similarity measures (see e.g. [11]) will be studied. Furthermore, alternative classification algorithms will be tested and compared with KNN-type classifiers. Finally, adding filtering and smoothing to eNose-based localisation will be investigated.

References

[1] P. Davidson and R. Piché, *A survey of selected indoor positioning methods for smartphones*, IEEE Communications Surveys & Tutorials, vol. 19, no. 2, pp. 1347–1370, Secondquarter 2017.

[2] L. Mainetti, L. Patrono, and I. Sergi, *A survey on indoor positioning systems*, in 2014 22nd International Conference on Software, Telecommunications and Computer Networks (SoftCOM). Split, Croatia, September 2014, pp. 111–120.

[3] P. Müller, J. Lekkala, S. Ali-Löytty and R. Piché, *Indoor Localisation using Aroma Fingerprints: A First Sniff*, in 14th Workshop on Positioning, Navigation and Communications (WPNC). Bremen, Germany, October 2017

[4] S. Kiani, S. Minaei, and M. Ghasemi-Varnamkhasti, *Application of electronic nose systems for assessing quality of medicinal and aromatic plant products: A review*,Journal of Applied Research on Medicinal and Aromatic Plants, vol. 3, no. 1, pp. 1–9, March 2016.

[5] A. D. Wilson and M. Baietto, *Applications and advances in electronic-nose technologies*, Sensors, vol. 9, no. 7, pp. 5099–5148, 2009.

[6] [Online]. Available: https://www.environics.fi/product/chempro100i/

[7] D. Zamora and M. Blanco, *Improving the efficiency of ion mobility spectrometry analyses by using multivariate calibration*, Analytica Chimica Acta, vol. 726, pp. 50–56, 2012.

[8] Owlstone nanotech. Available: http://info.owlstonenanotech.com/rs/owlstone/images/FAIMS%20Whitepaper.pdf .

[9] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Wiley-Interscience, 2001.

[10] G. A. Eiceman, Z. Karpas, and H. H. Hill Jr., *Ion Mobility Spectrometry*, 3rd ed. CRC Press, 2014.
[11] M. M. Deza and E. Deza, *Encyclopaedia of Distances*, Springer, 2009.
[12] S. H. Cha, *Comprehensive survey on distance/similarity measures between probability density*, International Journal of Mathematics and Methods in Applied Sciences, vol. 1, no. 4, pp. 300–307, 2007.
[13] [Online] G. Minaev, R. Piché, A.Visa, *Distance measures for classification of numerical features*, http://www.tut.fi/~piche/misc/distanceMeasures.pdf
[14] V. H. Bennetts, A. J. Lilienthal, P. P. Neumann, and M. Trincavelli, *Mobile robots for localizing gas emission sources on landfill sites: is bio-inspiration the way to go?*, Frontiers in Neuroengineering, vol. 4, pp. 1-12, January 2012.
[15] A. Loutfi, *Odour recognition using electronic noses in robotic and intelligent systems*, PhD thesis, Institutionen för teknik, Örebro universitet, Örebro, Swedish, 2006.