FEATURE EXTRACTION USING DISTRIBUTION REPRESENTATION FOR COLORIMETRIC SENSOR ARRAYS USED AS EXPLOSIVES DETECTORS

Tommy S. Alstrøm[a], Raviv Raich[b], Natalie V. Kostesha[c], Jan Larsen[a]

[a]Dept. of Informatics and Mathematical Modeling, Technical University of Denmark
Richard Petersens Plads 321, 2800 Kgs. Lyngby, Denmark
{tsal,jl}@imm.dtu.dk

[b]School of Electrical Engineering and Computer Science
Oregon State University, Corvallis, OR 97331-5501
raich@eecs.oregonstate.edu

[c]Dept. of Micro- and Nanotechnology, Technical University of Denmark
ørsteds Plads 345 East, DK-2800, Kgs. Lyngby, Denmark
natalie.kostesha@nanotech.dtu.dk

ABSTRACT

We present a colorimetric sensor array which is able to detect explosives such as DNT, TNT, HMX, RDX and TATP and identifying volatile organic compounds in the presence of water vapor in air. To analyze colorimetric sensors with statistical methods, a suitable representation of sensory readings is required. We present a new approach of extracting features from a colorimetric sensor array based on a color distribution representation. For each sensor in the array, we construct a K–nearest neighbor classifier based on the Hellinger distances between color distribution of a test compound and the color distribution of all the training compounds. The performance of this set of classifiers are benchmarked against a set of K–nearest neighbor classifiers that is based on traditional feature representation (e.g., mean or global mode). The suggested approach of using the entire distribution outperforms the traditional approaches which use a single feature.

***Index Terms*—** Hellinger distance, chemo–selective compounds, explosives detection, feature extraction, K–nearest neighbor classification

We acknowledge the support from the Danish Agency for Science and Technology's, Program Commission on Nanoscience Biotechnology and IT (NABIIT). Case number: 2106-07-0031 - Miniaturized sensors for explosives detection in air. Further we acknowledge Mikkel Schmidt, Dept. of Informatics and Mathematical Modeling, Technical University of Denmark and Ryota Tomioka, Dept. of Mathematical Informatics, University of Tokyo for invaluable and insightfull discussions.

1. INTRODUCTION

Over the past decade, explosives have been a preferred tool for terrorists, yet there is no satisfactory mobile and portable solution to detect explosives. To detect a variety of military and industrial explosives easily, new technologies must be developed. There are several application areas for explosives sensors, such as anti-terrorism (screening luggage and mail packages, checking suspects and mass transit systems), demining and environmental monitoring of hazardous compounds.

Sensors must not only easily detect a variety of hidden explosives, they must also be able to detect illegal chemicals and products of the explosives industry. Further requirements are that the sensing device should be portable, rapid, highly sensitive, specific (minimize false alarms), and inexpensive [1].

Over the past years a number of detection methods have been developed and successfully applied in explosives detectors. These include, but are not limited to, gas chromatography, Raman spectrometry, mass spectrometry, ion mobility spectrometry and colorimetric sensors. Suslick *et al*. described the application of the colorimetric sensor array for detecting volatile organic compounds in the gas phase [2, 3] as well as for identifying different organic compounds in the liquid phase [4, 5]. In our project we develop a colorimetric sensor array that can be useful in detecting and identifying explosives such as TNT, DNT, HMX, RDX and TATP [6, 7]. The colorimetric sensor is a fascinating technique for distinguishing different chemical compounds belonging to various classes, like amines, cyanides, alcohols, arenes, ketones, aldehydes and acids in the parts-per-million (ppm) and parts-per-billion (ppb) ranges [3, 8, 9]. In our research we use a com-

pletely different class of chemo–selective compound, which has already shown excellent results for detecting TNT. This type of colorimetric sensor could be successfully applied in national security and defense [10, 11].

A colorimetric sensor array consists of a number of chemo–selective compounds of various colors that will undergo a color change when subjected to an environment or a target substance, hereafter denoted an *analyte*. These chemo–selective compounds, which are typically called *dyes* are digitalized. Currently we use a flatbed scanner. One dye consists of several hundred pixels, but classically a dye is considered to have only one color, which is commonly found by calculating the mean or global mode pixel value [12]. We hypothesize that the complete distribution of color pixel value may contain additional information that can improve classification accuracy relative the information associated with a single pixel value such as the mean.

In this paper, we present a new method for representation and analyzing of the output of a colorimetric sensor array using the complete color distribution. To classify a given analyte, we propose a K-NN approach which uses the Hellinger distance between color distributions as a metric. By comparing this with a K-NN that use of a single feature such as the mean or global mode we are able to demonstrate significant improvement in accuracy.

2. COLORIMETRIC SENSORS

The colorimetric sensor array consists of a number of chemo–selective compounds immobilized onto silica gel resulting in circular spots (Fig. 1A). Each individual spot was approximately 3 mm in diameter with the total size of the sensor array of approximately 2.5 cm $\times$ 4.0 cm.

The dataset used in this paper has been discussed in detail in earlier work [12] but is summarized here for completeness. The sensor array has been exposed to analytes belonging to the various chemical families – 9 families in total, making it a multi-class dataset. The chemical families are: acids (45), alcohols (27), amines (42), arenes (14), environment (28), explosives (56), inorganic explosives (14), ketones (13) and thiols (14). The number in the parenthesis denotes the number of examples measured for the class in question, bringing to total number of examples to 253.

Data acquisition

Once the images of the sensor arrays have been digitalized, feature extraction is employed, typically using the mean pixel value. In order for the mean to be a robust measure of color change, the pixels of a dye have to be normally distributed (or at least have a symmetric distribution with one mode) and relatively free from outliers. As can be seen in Fig. 1 this may not always be the case. From a chemical point of view we know that a dye should only have one color, as the dye is homogeneous and exposed to a homogeneous vapor. However, noise is induced from: the scanner, the enhanced temperature for explosive detection, external light, and roughness of the surface. Some of these effects can be handled easily. The high temperature often results in a ring near the perimeter of the dyes (the coffee stain effect) and this area of the dyes is unreliable. In order to accommodate this effect, a smaller area of a dye is used for feature extraction, corresponding to 2/3 of the dye radius. To handle the other noise effects that cause pixel outliers, we have in earlier work suggested that the global mode is the most robust single value statistic compared to the mean, mode or median [12]. The global mode finds the most frequent pixel value occurring in a dye and as such is guaranteed to calculate a pixel value that exist in the given dye.

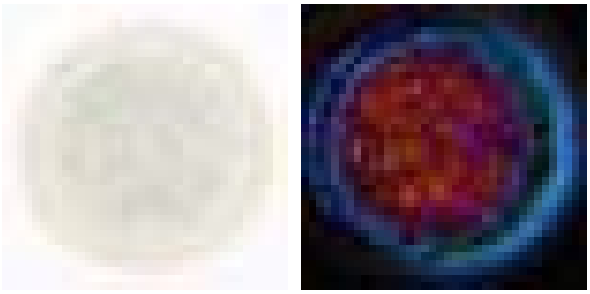

Fig. 1. An example of a specific dye of colorimetric sensor array exposed to the explosive analyte RDX. A: the sensor before exposure. B: the enhanced difference image.

Histogram features

In addition to the mean and global mode features used to characterize the color change response, we consider in this context the bag-of-words representation for multiple instance examples. The i'th example (dye) is represented by $X_i = \{x_{i1}, \ldots, x_{in_i}\}$, where x_{ij} is the j'th three–dimensional difference RGB pixel value between control and exposed, and n_i is the number of pixels considered for the representation of the i'th example. For several classifiers a notion of distance between examples is a key component. To construct a distance between two examples in the bag-of-words representation, we propose to represent each multi-instance example with a distribution and use the Hellinger distance as a metric between two examples. The motivation behind this approach is that differences between distributions, which are not directly measurable through the mean (or other moments), can still be detected. This approach was demonstrated to be effective in several application areas, e.g., disease classification using flow cytometry [13] and document classification [14].

Assuming an underlying probability density function f_i

such that $x_{ij} \sim f_i$ for $j = 1, 2, \ldots, n_i$, one can associate X_i with the following kernel density estimate

$$f_i(x) = \frac{1}{n_i} \sum_{j=1}^{n_i} K(x - x_{ij})$$

where $K(x) = 1/(2\pi\sigma^2)^{d/2} \exp(-||x||^2/2\sigma^2)$, $d = 3$ in our case. Recall that given two PDFs f_i and f_k, the squared Hellinger distance between the two distributions is given by

$$d_H(f_i, f_k)^2 = \int \left(\sqrt{f_i(x)} - \sqrt{f_k(x)}\right)^2 dx$$

i.e., the Euclidean distance between the square-root of the PDFs. Note that the squared Hellinger distance can be computed using the following equivalent formula: $d_H(f_i, f_k)^2 = 2 - 2\int \sqrt{f_i(x) f_k(x)} dx$. For computational simplicity, we consider the following equivalent alternative:

$$d_H(f_i, f_k)^2 = 2 - 2(E_{f_i}[\sqrt{T(x)(1 - T(x))}] + E_{f_k}[\sqrt{T(x)(1 - T(x))}])$$

where $T(x) = \frac{f_i(x)}{f_i(x) + f_k(x)}$ and $E_h[\cdot] = \int \cdot h(x) dx$. A sample-based version of this expression can be computed by replacing the expectations with their sample averages and the distributions with their kernel estimates,

$$E_{f_i}[\sqrt{T(x)(1 - T(x))}] \approx \frac{1}{n_i} \sum_{j=1}^{n_i} \sqrt{T(x_{ij})(1 - T(x_{ij}))}$$

Naturally, the distance calculation can be directly applied to a K-NN classifier. This approach can be considered an alternative to a set distance between two collections instances.

Moreover, this approach allows for a feature vector construction. Consider a new example X associated with PDF f. The feature vector for this example can be constructed as $\phi(X) = [d_H(f, f_1), d_H(f, f_2), \ldots, d_H(f, f_N)]^T$ where N is the number of training examples. Note that this feature vector has a fixed size, independent of the number of instances (pixels) in its bag-of-words representations. This representation can be applied to a variety of classifiers. For example, in SVM [15] the classifier can be of the form $\text{sgn}\langle w, \phi(X)\rangle$. In many cases, the SVM solution results in a sparse vector w for which the non-zero entries correspond to support vectors. In our setup, the Hellinger distance to key multi-instance examples will determine the output of the classifier.

3. METHODS AND RESULTS

Despite its simplicity, K-NN is an effective classification technique [15] which works as follows. When testing an unknown data point, the Euclidean distances for all known points are calculated. The classes of the closest K points are then identified and the unknown point is classified using majority voting of these known points.

		Dye rank		
Class	Method	1st	2nd	3rd
Acids	Mean	**1.2**	**2.4**	4.3
Acids	GMode	2.4	**2.4**	3.6
Acids	Hellinger	1.6	**2.4**	**2.8**
Alcohols	Mean	**7.5**	**8.3**	**8.3**
Alcohols	GMode	8.3	8.7	8.7
Alcohols	Hellinger	7.9	**8.3**	8.7
Amines	Mean	7.1	7.1	7.1
Amines	GMode	7.1	7.1	7.5
Amines	Hellinger	**6.3**	**6.7**	**6.7**
Explosives	Mean	2.8	3.2	4.3
Explosives	GMode	3.2	4.7	5.9
Explosives	Hellinger	**1.2**	**2.0**	**2.8**
Thiol	Mean	0.8	5.1	5.1
Thiol	GMode	0.8	3.6	4.7
Thiol	Hellinger	**0.4**	**3.2**	**4.0**

Table 1. The error rate of the 3 best performing dyes for each feature extraction method. The numbers are reported as % leave–one–out classification error.

We apply a K-NN classifier to each dye for each feature extraction technique in a 1 vs all setting. From earlier work [12] it was shown that the sensor is proficient in detecting acids, alcohols, amines, explosives and thiols so these are the classes for which we train classifiers. In order to carry out both model selection and estimation of the generalization error, double-cross validation using *leave–one–out* is performed. Our scheme result in a total of 155 classifiers per feature extraction method (31 dyes $\times$ 5 classes).

To establish if the Hellinger method produces greater or smaller classification error rate relative to the mean and global mode we examine wins, ties, and losses. To determine ties we use significance testing following McNemar significance test [16] using $\alpha = 0.01$ due to the amount of hypotheses we test. For Hellinger vs mean we find that Hellinger has eight wins, 146 ties and one loss. For Hellinger vs global mode we find Hellinger better nine times and global mode two times and 144 ties. Of out the twenty significant results we have an positive false discovery rate (pFDR) of of 0.10, that is, we expect that two of the significant results where declared significant by error [17]. Table 1 shows how the feature extraction methods compare against each other when we choose the three best dyes for each case.

We also apply K-NN classifiers in a multi-class setting resulting in a total of 31 classifiers per feature extraction method, one classifier per dye. Fig. 2 shows the classification error for each of the method ordered by classification error using the Hellinger method. Performing the same hypothesis test idiom as before, we find that the Hellinger method was significantly better in 14 cases out of 62 (better than the mean and global mode in seven cases respectively, not always for

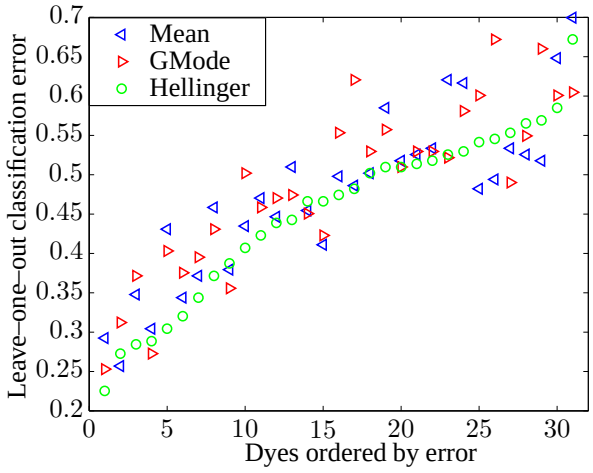

Fig. 2. Classification error for the feature extraction methods when k-NN is used to quantify the errors.

the same dyes) and worse in zero cases. The pFDR is 0.02 in the multi-class setting.

4. CONCLUSION

Despite the variability in the color reading of a given compound using one sensor, traditional methods consider representing the entire reading using a single value. To account for this variability, we proposed a complete distribution representation. To classify using the distribution representation, we adopted the Hellinger distance-based K-NN algorithm. To evaluate the potential benefit of using the complete distribution as opposed to the mean only for example, we compared single feature vector representation with the full distribution representation. We showed that the distribution representation with a Hellinger K-NN approach is either equal or better than the single vector representation with a Euclidean K-NN approach. The evidence for Hellinger being the better method is especially strong in the multi-class setting where it was significantly better in 23% of the cases.

References

[1] M. S. Schmidt, N. Kostesha, F. Bosco, J. K. Olsen, C. Johnsen, K. A. Nielsen, J. O. Jeppesen, T. S. Alstrøm, J. Larsen, T. Thundat, M. H. Jakobsen, and A. Boisen, "Xsense - a miniaturised multi-sensor platform for explosives detection," in *Proceedings of SPIE*, 2011.

[2] K. S. Suslick, N. A. Rakow, and A. Sen, "Colorimetric sensor arrays for molecular recognition," *Tetrahedron*, vol. 60, no. 49, pp. 11133–38, Nov. 2004.

[3] N. Rakow, A. Sen, M. C. Janzen, J. B. Ponder, and K. S. Suslick, "Molecular recognition and discrimination of amines with a colorimetric array.," *Angewandte Chemie (International ed. in English)*, vol. 44, no. 29, pp. 4528–32, July 2005.

[4] C. Zhang and K. S. Suslick, "A colorimetric sensor array for organics in water.," *Journal of the American Chemical Society*, vol. 127, no. 33, pp. 11548–9, Aug. 2005.

[5] C. Zhang, D. P. Bailey, and K. S. Suslick, "Colorimetric sensor arrays for the analysis of beers: a feasibility study.," *Journal of agricultural and food chemistry*, vol. 54, no. 14, pp. 4925–31, July 2006.

[6] N. V. Kostesha, T. S. Alstrøm, C. Johnsen, K. A. Nielsen, J. O. Jeppesen, J. Larsen, M. H. Jakobsen, and A. Boisen, "Development of the colorimetric sensor array for detection of explosives and volatile organic compounds in air," in *Proceedings of SPIE*, Apr. 2010, vol. 7673, pp. 76730I–76730I–9.

[7] N. V. Kostesha, T. S. Alstrøm, C. Johnsen, K. A. Nielsen, J. O. Jeppesen, J. Larsen, A. Boisen, and M. H. Jakobsen, "Multi-colorimetric sensor array for detection of explosives in gas and liquid phase," in *Proceedings of SPIE*, 2011.

[8] S. H. Lim, L. Feng, J. W. Kemling, C. J. Musto, and K. S. Suslick, "An Optoelectronic Nose for Detection of Toxic Gases.," *Nature chemistry*, vol. 1, pp. 562–567, Sept. 2009.

[9] C. Zhang and K. S. Suslick, "Colorimetric sensor array for soft drink analysis.," *Journal of agricultural and food chemistry*, vol. 55, no. 2, pp. 237–42, Jan. 2007.

[10] D. Kim, V. M. Lynch, K. Nielsen, C. Johnsen, J. O. Jeppesen, and J. L. Sessler, "A chloride-anion insensitive colorimetric chemosensor for trinitrobenzene and picric acid," *Analytical and bioanalytical chemistry*, vol. 395, no. 2, pp. 393–400, Sept. 2009.

[11] J. S. Park, F. Le Derf, C. M. Bejger, V. M. Lynch, J. L. Sessler, K. Nielsen, C. Johnsen, and J. O. Jeppesen, "Positive Homotropic Allosteric Receptors for Neutral Guests," *A European Journal*, vol. 16, no. 3, pp. 848–54, Jan. 2010.

[12] T. S. Alstrøm, J. Larsen, N. V. Kostesha, M. H. Jakobsen, and A. Boisen, "Data representation and feature selection for colorimetric sensor arrays used as explosives detectors," in *IEEE International Workshop on Machine Learning for Signal Processing*, Sept. 2011.

[13] K. M. Carter, R. Raich, W. G. Finn, and A. O. Hero, "Information preserving component analysis: Data projections for flow cytometry analysis," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 3, no. 1, pp. 148–158, 2009.

[14] K. M. Carter, R. Raich, W. G. Finn, and A. O. Hero, "Fine: Fisher information nonparametric embedding," *IEEE transactions on pattern analysis and machine intelligence*, pp. 2093–2098, 2009.

[15] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[16] Quinn McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, pp. 153–157, 1947.

[17] John D. Storey, "A direct approach to false discovery rates," *Journal of the Royal Statistical Society: Series B*, vol. 64, no. 3, pp. 479–498, 2002.