Support Vector Machines and Self-Organizing Maps for the recognition of sound events in urban soundscapes

Xavier Valero [a]
Francesc Alías [b]
GTM-Grup de Recerca en Tecnologies Mèdia, La Salle – Universitat Ramon Llull.
Quatre Camins 2, 08022 Barcelona, Spain.

Damiano Oldoni [c]
Dick Botteldooren [d]
Department of Information Technology, Ghent University.
Sint-Pietersnieuwstraat 41, 9000 Gent, Belgium.

Sound event recognition is a crucial aspect of human auditory perception. Hence, it has to be taken into account when it comes to understanding how humans perceive soundscapes. In that context, both unsupervised and supervised learning techniques can be used. On the one hand, this paper takes the latter approach for the recognition of sound events typically encountered in urban environments. Sound signals are described using a set of auditory-based features and then sound event recognition is performed employing multi-class Support Vector Machines. On the other hand, a combined approach including unsupervised learning (specifically, Self-Organizing Maps) for clustering and collecting real world samples and supervised learning for labeling is introduced. Finally, listening tests are also carried out in order to compare the accuracy achieved by the proposed system with the human ability.

1   INTRODUCTION

When it comes to designing an urban soundscape, starting to plan from scratch at an early stage of the project might be preferred. However, in most cases, urban planners and decision makers have to deal with already existing situations that have a predefined architecture and that contain certain pleasant and unpleasant sounds. Thus, their task consist in trying to improve as much as possible the soundscape quality within the given location and context. In these cases, knowing which are the typical neighborhood sounds and the rare sound events that could attract attention is useful information for the soundscape designer.

---

[a] email: xvalero@salleurl.edu
[b] email: falias@salleurl.edu
[c] email: damiano.oldoni@intec.ugent.be
[d] email: dick.botteldooren@intec.ugent.be

In this framework, the role of environmental sound recognition may become especially relevant. This research field aims at creating automated systems able to recognize the sound events occurring in a sonic environment. For this purpose, two different approaches might be considered: supervised and unsupervised learning techniques. The selection of one or another will mainly depend on the available site information, as described in the next paragraphs.

We first consider a scenario in which we know beforehand the sounds that we want to identify and label at a given location. In those cases, it is feasible to employ supervised learning based on sound samples that are collected and labeled manually. In the related literature, several algorithms have been successfully employed, such as Hidden Markov Models[1], Fisher Linear Discriminant[2], K-Nearest Neighbor[3] or Artificial Neural Networks[4, 5].

However, if we consider a scenario in which we do not have sufficient prior knowledge about the occurring sound events, or the sounds we might want to label, the first task is to separate out the sounds from the acoustic scene. For this, it is required to turn to unsupervised learning techniques, which group the data into similarity clusters that provide a representation of the typical sound events occurring. Different clustering algorithms have been used in the environmental sound domain: Markov-Model based clustering[6], non-negative matrix factorization and spectral clustering[7] or co-clustering[8]. Oldoni et al.[9] proposed a specific model for environmental sounds that mapped the acoustical features based on co-occurrence using an extension of the Self-Organizing Map[10]. This methodology allows collecting prototypical samples of the most typical sounds to describe the soundscape at a given location. Verbally labeling the collection of recordings of typical sounds is an important next step because it gives meaning to the sounds and thus allows creating logical families (e.g. road vehicle sounds) and deriving statistics on occurrence.

This work presents a twofold contribution. Firstly, considering a scenario where we know beforehand the most typical sound sources, we test the sound event recognition performance of the supervised Support Vector Machines (SVM), a well known technique in general pattern recognition problems which has also shown a good performance in audio classification tasks[11,12]. Secondly, considering a scenario without previous sound information, a SOM is trained (following the work in Oldoni et al.[9]) and a new automated method for subsequent labeling, based on SVM, is proposed. Finally, by means of a listening test, we validate the proposed method by comparing the output sound labels to those given by human listeners.

The paper is organized as follows. In the next section a brief introduction to SVM theory is presented. Section 3 presents the basics of SOM and its usage to create a compilation of typical sounds which are the sound database used for labeling tasks, followed by a section describing the proposed SOM labeling method. The experimental work and the obtained results are shown in Section 5. The paper ends with presenting the conclusions and future work lines.

2   SUPPORT VECTOR MACHINES

Support Vector Machines (SVM) is a supervised learning method largely used for classification problems [11, 12, 13]. Considering a binary separation problem, the basis of the SVM is mapping the input samples into a high dimensional space and finding the hyperplane that optimally separates the two classes. The optimal separating hyperplane is chosen following the criteria of maximizing the distance to the closest training instance. Hereafter the basis of SVM theory is briefly presented. For a deeper discussion, we refer the reader to Cristianini and Shawe-Taylor [13].

Let $x_i \in X \subseteq R^n$ be the input feature vector and $y_i \in Y = \{1, -1\}$ the target of a binary classification, where R^n denotes the n-dimensional real space. Suppose a training set $S = \{(x_1,$

y_1), (x_2, y_2), …. (x_l, y_l)} $\subseteq$ (X x Y)^L , where L is the number of examples. Considering a linear classification case, the separating hyperplane can be written as:

$$f(x) = \langle w \cdot x \rangle + b \tag{1}$$

where w is the weight vector orthogonal to the hyperplane and b is the bias. The decision rule given by *sgn(f(x))* divides the input space into two parts. Several hyperplanes might be able to perform the input space division matching the training set S. However, the SVM theory seeks the hyperplane that maximizes the separation to the closest sample (i.e., margin). The optimal hyperplanes are set in such a way that the margin is to 1 (see Figure 1).

Quite often, non-linearly separable problems will be faced. Then, non-linear kernel functions should be used. These functions map the input feature space X to another high-dimensional feature space F. This process can greatly simplify the classification task, since the samples nonlinearly separable in X may be linearly separated in F. The most common kernel functions are the following:

- Polynomial:
$$K(x, y) = \left(\langle x, y \rangle + 1 \right)^d \tag{2}$$

- Gaussian Radial Basis Function:
$$K(x, y) = \exp\left(\frac{-(x-y)}{2\sigma^2} \right)^2 \tag{3}$$

Where d is the polynomial degree and σ^2 is the variance of the Gaussian function.

Another important issue to adapt SVM to real-world problems is the need to generalize the binary separation problems (i.e., recognition of two different classes, sound events in this work) to a multiclass separation (recognition of n different classes or sound categories). Several strategies can be followed, such as the *one.vs all* or the *one vs. one*[12].

3   SELF ORGANIZING MAPS AND ACOUSTIC SUMMARIES

The Self-Organizing Map (SOM) is an unsupervised trained neural network, typically described as a tool for visualizing high-dimensional data. Based on topographic mapping principles, SOM takes inspiration from the observation in the human sensory cortex of many topologically organized regions (see Kohonen[10] for a detailed overview and references), fundamental for sensory processing[14]. Tonotopic maps have been found in the auditory sensory cortex of primates[15,16] and humans[17,18,19]. Retinotopic and somatotopic maps have been discovered in primates and human cortex. Although these topologically organized structures are mainly genetically determined, some sensory projections show a certain degree of plasticity and are able to modify their dimensions and their structure due to experience or specific traumatic events[20]. Moreover, postnatal self-organizing processes occur in other more abstract maps in several area of the brain[10].

In this paper the most used structure of SOM is employed: a two dimensional grid of units or nodes $m_i = (m_x; m_y) \in R^2$, each of which representing a reference vector s_i in the n-dimensional input space R^n. In this paper such space corresponds to a high-dimensional space of acoustical features as in Oldoni et al. [9, 21]. These features are measures for intensity, spectral and temporal modulation using a center-surround mechanism in order to mimic the receptive fields in the

auditory cortex at a low computational cost. At each time step t an input sound feature vector $r(t)$ $\in R^n$ is calculated and the best matching unit (BMU) $m_c(t)$ of the SOM is found, defined as the unit m_c whose reference vector s_c is the nearest to $r(t)$:

$$c(t) = \mathrm{argmin}_i \left\| r(t) - s_i(t) \right\| \tag{4}$$

The training step is then performed, defined as follows:

$$s_i(t+1) = s_i(t) + h_{c(t),i}\left(r(t) - s(t) \right) \tag{5}$$

The reference vectors of the BMU and of its neighbors are adapted at each time step. The definition and the degree of neighborhood is defined by a so-called neighborhood function $h_{c(t)i}$, a smoothing kernel defined on the two-dimensional lattice of units. For convergence, the function $h_{c(t)i}\rightarrow 0$, for t→∞. After vastly iterating the training algorithm as formulated in Eqn. (4) and (5), the reference vectors of the SOM are a discrete non linear and topographically ordered 2D projection of the frequency distribution of the input data. After training, the number of SOM units encoding, by means of their reference vectors, a certain region of the feature space depends on the frequency distribution of the input feature vectors.

This training, purely based on frequency of occurrence, is followed by a specific training called continuous selective learning[9]. Human learning is, in fact, not based only on frequency of occurrence of given sensory stimuli; contrarily, factors as attention play an important role. This second training phase promotes the learning of sounds that could potentially attract attention due to their saliency and novelty, while disregarding the other sounds (details on saliency calculation can be found in De Coensel and Botteldooren[22]).

The reference vectors of the SOM units can be seen as representative abstract sound prototypes, which can be translated into hearable sound samples by means of a sound recording session (details in Oldoni et al.[9]). The set of collected sound excerpts is called the "acoustic summary" of the given soundscape[9].

4   AUTOMATED SOM LABELING

In previous works[9], an acoustic summary was collected finding sounds whose sound feature vectors were as similar as possible to the reference vectors of the SOM units (see Figure 2). Each sound sample of the acoustic summary is linked to one and only one SOM unit. Each sound sample could be manually labeled by an expert listener, thus involving listening one by one to every sound fragment. This process has many drawbacks: it is complex, it requires a lot of time and attention from the expert listener and it is certainly unfeasible for being implemented in a soundscape analyzer tool.

This paper presents an alternative method which notably simplifies the process and does not require the constant participation of an expert listener. The method is based on SVM to automatically label the SOM nodes (see Figure 3). The SVM is formerly trained using the SOM node vectors which inherit the labels given by an expert listener to the correspondent sound sample. The use of the SOM node vectors as input data is based on the assumption that the SOM nodes preserve the original signal feature space[9] (in this case, the features related to loudness, amplitude modulation and frequency modulation of the sound signal as in Section 3). Thus, the process of collecting, parameterizing and listening to short sound samples is avoided: the only required data are one or more formerly labeled SOMs (from other time periods or other locations) in order to train the SVM.

5   EXPERIMENTAL WORK

5.1. Sound database and labeled corpus

Two acoustic summaries, related to the units of two trained SOMs, have been extracted from two different recording sessions of approximately 10 hours long each. The two SOMs have been trained on sound feature vectors calculated from continuous input data collected during three weeks in October and November 2011 respectively from the same location. The recording sessions followed the SOM training periods. The acoustic summaries were composed of 2369 and 2892 samples respectively, i.e. 68% and 83% of the total 3500 SOM nodes.

An expert listener (a researcher specialized on environmental acoustics) listened to the 5seconds long sound samples composing the acoustic summaries and observed that the most common sound events could be referred to the following classes: *bird*, *chatting people*, *car*, *truck*, *motorbike/scooter*, *tram* and *background noise* [a].

The same listener selected the sounds belonging to these classes and classified them. Two sets were then created: the first one was composed of 1046 sound fragments whilst the second set was composed of 1206 sound fragments, i.e. 44% and 42% of the total number of samples composing the acoustic summaries.

5.2   Supervised learning

First, we consider the scenario in which sound labeled data is available and, thus, supervised learning techniques can be applied. Specifically, it was aimed to test SVM performance on environmental sound event recognition. Sound feature vectors related to the sound samples were calculated, as explained in Section 3. A subsequent Principal Component Analysis was applied to reduce their dimensionality[4] and make it suitable for SVM training. The SVM employed a Radial Basis Function Kernel, which was empirically selected among other kernels. A *one vs. all* strategy was followed to face the multiclass problem, given its lower complexity when compared to other strategies[6].

With those settings, the SVM was trained using the corpus collected in October 2011 and labeled by the expert listener and tested using the set collected in November 2011. From the 1206 test sound files labeled by the expert listener, 983 (81.5%) were correctly recognized by the SVM. The confusion matrix among the different classes was calculated so as to understand in which cases the SVM failed to recognize the sound events. As detailed in Table 1, *background noise*, *birds* and *cars* were the sounds attaining the highest accuracy, with rates beyond the 90% of correctly recognized sound events. The accuracy decreased (around 60%) when it came to recognize *truck* and *motorbike/scooter* events. The confusions between the three road vehicle sounds were the cause for that decrease, as also noticed in Valero and Alías[4]. Finally, it could be observed that p*eople talking* presented the lowest accuracy. That sound category was confused either with *background noise* (in sound samples where the people where far away from the microphone) or with *cars* (in sound samples where those were far away but also simultaneously present).

---

[a] The term "background noise" refers to low sound events where no specific sound source can be clearly recognized.

5.3 Self Organizing Maps labeling

A SOM was constructed based on sound information collected during November 2011. As shown in Fig. 4a, several clusters can be observed. The SOM labeled by the expert listener, taken as the reference in this work, is shown in Fig. 4b. It can be noticed that not all the SOM nodes have a label (i.e., nodes not colored): some less frequent sound categories were not considered (church bells, different kind of alarming sounds as horns etc.), neither the mixtures of co-occurring sounds.

The proposed automated SOM labeling method, as explained in Section 4, was next tested. To train the SVM, another SOM labeled by the expert listener was taken. This trained SOM contained sound information collected from the same location but in a different period, specifically October 2011. As observed in Fig. 4b-c, the proposed SVM automated method provides a SOM labeling quite similar to the one given by the expert (917 matching labeled SOM units, 76% of the1206 units labeled by the expert listener).Thus, the results suggest that the proposed SVM labeling method is able to reproduce with a high degree of accuracy the human SOM labeling when sufficient data are available.

5.4 Listening tests - Non expert labeling

Two different tests were carried out to refer the accuracy obtained by the two approaches (i.e., supervised learning and SOM labeling) to human ability. The first set of listening tests was conducted to compare human performance to that obtained by the supervised learning approach (using SVM). A total of 14 persons, including both experts and non experts on acoustics, were asked to classify 60 sound events randomly selected from the testing set (see Section 5.1). The averaged recognition rate obtained by the 14 participants is 78.3%, which is slightly lower than the 81.5% obtained by the system. This result is important because it means that the SVM algorithm is comparable to human labeling capabilities.

The second test consisted on labeling the whole SOM used for testing by one of the previous 14 participants, hereafter referred as the non-expert listener. This test was much more demanding: 2892 sound events had to be labeled, in front of the 60 of the previous test. Observing the labeling provided by the non-expert listener (see Fig 4d), some differences may be found when compared to the SOM labeled by the expert (Fig 4b). The labels belonging to road vehicle categories (*car, truck* and *motorbike/scooter*) seem to be slightly more mixed in the case of the non-expert listener. Also its perception of *background noise* is different, reflected on the bigger cluster of labels referred to that sound category. Summing up all the categories, the non-expert gave a higher amount of labels than the expert (1543 and 1206, respectively). All these results confirm a natural human variability in distinguishing and tagging sounds. It is observed that the labeling deviation between human listeners is slightly larger than the deviation between an expert human listener and the proposed automated method, making it an interesting solution for automating the labeling without losing precision.

6   CONCLUSIONS

This paper has gathered two different approaches to tackle the recognition of environmental sound events, a key issue to understand urban soundscapes composition. Firstly, SVM (a supervised learning algorithm) has been tested. Despite facing the recognition of noisy data, the

performance of SVM is noticeable, achieving an accuracy rate higher than 80%, which is comparable to the human performance shown in the listening tests.

Secondly, a SOM has been constructed with sound data from the same location. After a specific unsupervised training phase, the SOM has learned both the typical sounds and the sounds that stand out composing the given soundscape. This way a set of sounds can be selected for labeling. In order to understand the obtained clusters, a SOM labeling method based on SVM classification has been proposed. The method, which is totally automatic, could be implemented in future real time applications and advanced soundscape analyzer tools. By means of listening tests, it has been shown that the labeling deviation of the system compared to the expert listener labeling is slightly smaller than the deviation found between human listeners.

Several opportunities for future work still exist. Firstly, enhancing sound signal parameterization by calculating features with narrower windows to make the system more sensitive to sound events typically short and highly frequency modulated, like *speech*. Secondly, testing the proposed labeling method with sound data collected in different locations and comparing it to labels given by more listeners. Finally and most importantly, improving the way in which vagueness in labeling by human listeners is handled.

7 ACKNOWLEDGEMENTS

8 REFERENCES

1. C. Couvreur, V. Fontaine, P. Gaunard and C.G. Mubikangiey, "Automatic classification of environmental noise events by Hidden Markov Models", *Applied Acoustics*, **54**(3), (1998).
2. M. Sobreira Seoane, A. Rodriguez Molares and J.L Alba Castro, "Automatic classification of traffic noise", *Proc. Acoustics'08*, (2008).
3. X. Valero and F. Alías, "Applicability of MPEG-7 low level descriptors to environmental sound source recognition", *Proc. Euroregio*, (2010).
4. A.J. Torija and D.P. Ruiz, "ANN-based model to identify noticed sound events. A tool for inclusion in actions plans against environmental noise in urban environments", *Proc. Forum Acusticum'11*, (2011).
5. X. Valero and F. Alías, "Automatic recognition of environmental sound sources", *Proc. Tecniacustica'11,* (2011).
6. K. Lee, D.P.W. Ellis and A.C. Loui, "Detecting local semantic concepts in environmental sounds using Markov Model based clustering*", Proc. ICASSP,* (2010).
7. J. Xue, G. Wichern, H. Thornburg and A. Spanias, "Fast query by example of environmental sounds via robust and efficient cluster-based indexing", *Proc. ICASSP,* (2008).
8. R. Cai, L. Lu, and A. Hanjalic, "Co-clustering for auditory scene categorization", *IEEE Transactions on Multimedia*, **10**(4), (2008).
9. D. Oldoni, B. De Coensel, M. Boes, T. Van Renterghem, S. Dauwe, B. De Baets and D. Botteldooren, "Soundscape analysis by means of neural network-based acoustic summary", *Proc. Internoise 2011,* (2011).
10. T. Kohonen, *Self-Organizing Maps, 3rd Edition*, Springer-Verlag (2001).

11. A. Rabaoui, M. Davy, S. Rossignol and N. Ellouze, "Using One-class SVMs and Wavelets for audio surveillance", *IEEE. Transactions Information Forensics and Security*, **3**(4), (2008).

12. C-C. Ling, S-H. Chen, T-K. Truong and Y. Chang, "Audio classification and categorization based on Wavelets and Support Vector Machine", *IEEE. Transactions on Speech and Audio Signal Processing*, **13**(5), (2005).

13. J. Cristianini and J. Shawe-Taylor, "An introduction to Support Vector Machines and other kernel-based learning methods", *Cambridge University Press*, (2000).

14. J.H. Kaas, "Topographic maps are fundamental to sensory processing", *Brain research bulletin*, **44**(2), (1997).

15. A. Morel and J. H. Kaas, "Subdivisions and connections of auditory cortex in owl monkeys", *J. Comp. Neurol.*, **318**, (1992).

16. C.I. Petkov, C. Kayser, M. Augath, and N.K. Logothetis, "Functional imaging reveals numerous fields in the monkey auditory cortex", *PLoS Biol.*, **4**, (2006).

17. T.M. Talavage, P.J. Ledden, R.R. Benson, B.R. Rosen, and J.R. Melcher, "Frequency-dependent responses exhibited by multiple regions in human auditory cortex", *Hearing Res.*, **150**, (2000).

18. T.M. Talavage, M.I. Sereno, J.R. Melcher, P.J. Ledden, B.R. Rosen, and A.M. Dale, "Tonotopic organization in human auditory cortex revealed by progressions of frequency sensitivity", *J. Neurophysiol.*, **91**, (2004).

19. C. Humphries, E. Liebenthal, and J.R. Binder, "Tonotopic organization of human auditory cortex", *NeuroImage* **50**, (2010).

20. R. Hunt and N. Berman, "Visual projection to the optic tecta after partial extirpation of the embryonic eye", *J. Comp. Neurol*, **162**, (1975).

21. D. Oldoni, B. De Coensel, M. Rademaker, B. De Baets, and D. Botteldooren, "Context-dependent environmental sound monitoring using SOM coupled with LEGION", *Proc. IEEE IJCNN* , (2010).

22. B. De Coensel and D. Botteldooren, "A model of saliency-based auditory attention to environmental sound", *Proc. ICA*, (2010).

23. A. Ultsch, "Self organized feature maps for monitoring and knowledge acquisition of a chemical process", *Proc. ICANN*, **93**, (1993).

Table 1 – Confusion matrix obtained with Support Vector Machines. The most frequent confusions are colored in red.

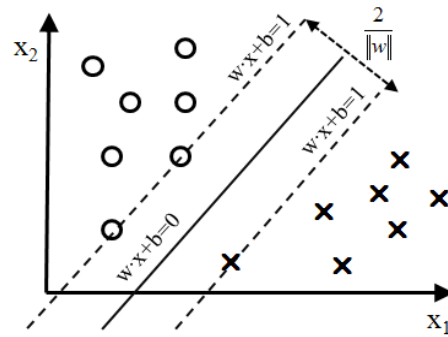|		Background noise	Bird	People talking	Car	Motorbike/Scooter	Truck	Tram
				TARGET				
OUTPUT	Background noise	91.1	1.7	**16.9**	0.3	1.0		0.3
	Bird		98.3	1.7		1.0		0.3
	People talking	8.9		33.9	1.3	1.0		1.5
	Car			**32.2**	94.7	10.7	**20.1**	2.5
	Motorbike/Scooter			13.6	1.8	60.2	11.4	7.1
	Truck				1.0	**26.2**	61.1	4.6
	Tram			1.7	1.0		7.4	83.7

Fig. 1 – Optimal separation hyperplane obtained with Support Vector Machine algorithm.
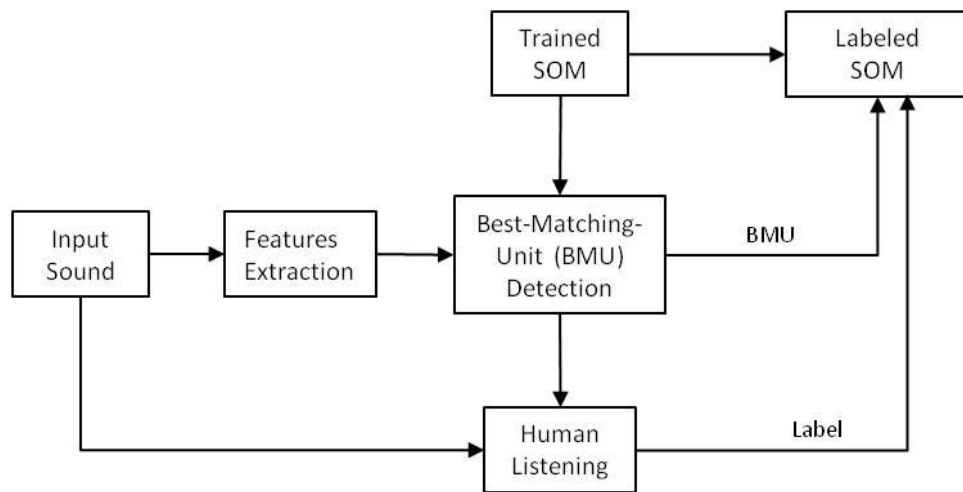

Fig. 2 –After a SOM has been trained on the soundscape from a given location, its units are manually labeled by an human listener based on sounds recorded from the same location.
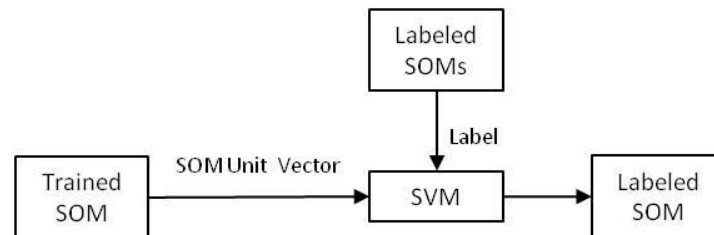

Fig. 3 – Proposed automated SOM labeling method using SVM.The units of the trained SOM are labelled by means of SVM, which has been trained using (one or more) formerly labeled SOMs
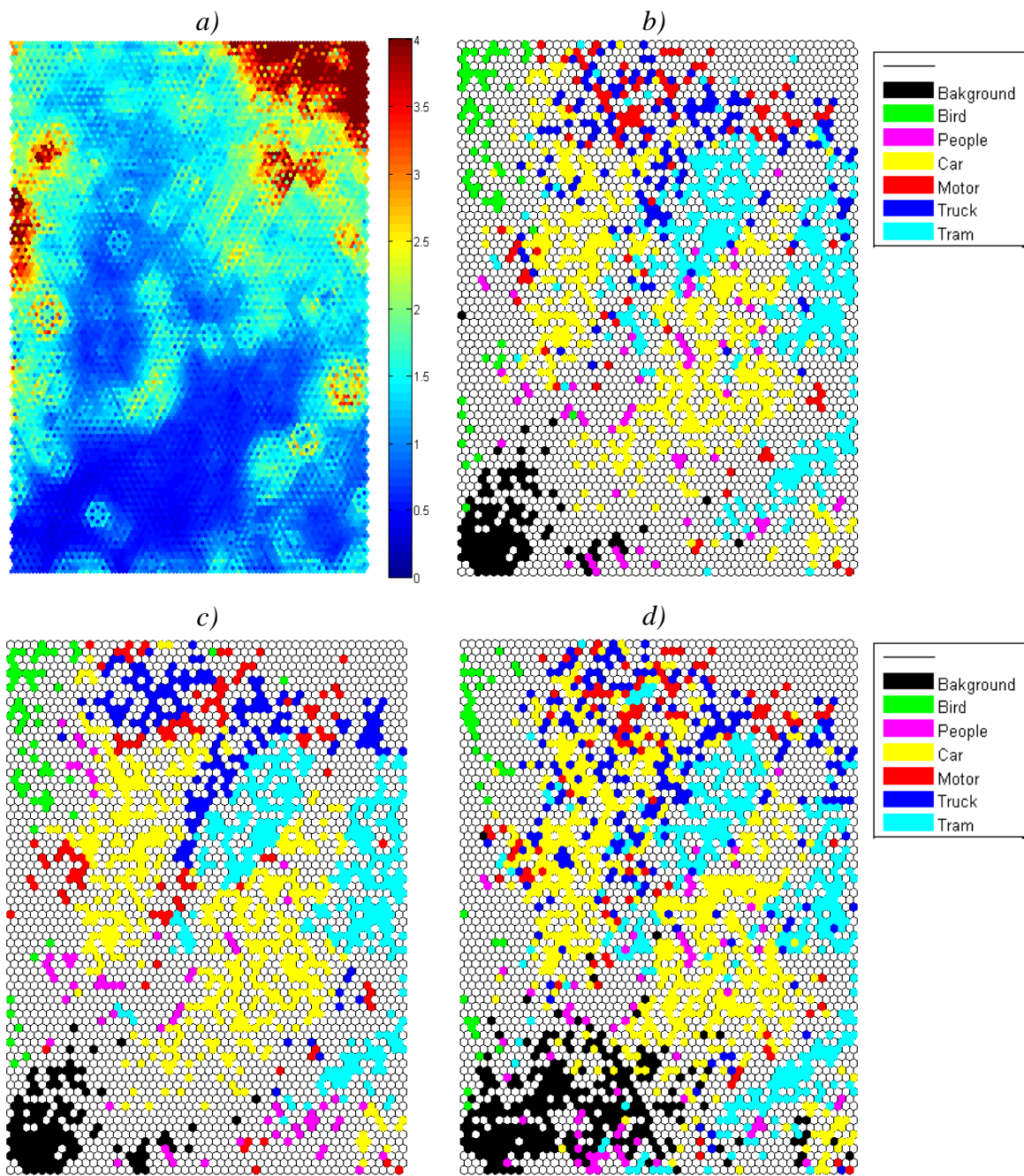
Fig. 4– a) U-matrix[23] representation of the trained SOM: the color shows the reciprocal distance among the nearest units of the SOM. In the other figures, SOM labelled by: b) an expert listener; c) the SVM automated method; d) a non-expert listener.