Deep Learning and Bayesian Networks for Labelling User Activity Context Through Acoustic Signals

Francisco J. Rodríguez Lera[1(✉)], Francisco Martín Rico[2], and Vicente Matellán[3]

[1] AI Robolab, University of Luxembourg, Luxembourg, Luxembourg
francisco.lera@uni.lu
[2] Universidad Rey Juan Carlos, Madrid, Spain
[3] Robotics Group, Universidad de León, León, Spain

Abstract. Context awareness in autonomous robots is usually performed combining localization information, objects identification, human interaction and time of the day. We think that gathering environmental sounds we can improve context recognition. With that purpose, we have designed, developed and tested an Environment Recognition Component (ERC) that provides an extra input to our Context-Awareness Component (CAC) and increases the rate of labeling correctly users' activities. First element, the Environment Recognition Component (ERC) uses convolutional neural networks to classify acoustic signals and providing information to the Context-Awareness Component (CAC) which infers the user activity using a hierarchical Bayesian network. The work described in this paper evaluates the results of the labeling process in two HRI scenarios: robot and user sharing room and robot, and when the human and the robot are in different rooms. The results showed better accuracy when the ERC uses acoustic signals.

1 Introduction

In order to produce natural responses to human behaviors, a robot should understand user's context. In this way to recognize and label the user activity is a cornerstone in HRI [21].

Activity context identification enhances the overall performance of the deliberative system [12] and favors a natural robot-user experience. Adding the ability to identify the context to autonomous robots will also help them to understand the environment where it inhabits and be aware of the situations that happens around it.

It is possible to define two procedures to infer and label the user activity. On the one hand, direct procedure, for instance a dialog system on the robot (through conversations or gestures) or using a software application connected to robot.

This work was partially supported by Spanish Ministry of Economy and Competitivity under grant TIN2016-76515-R.

J.M. Ferrández Vicente et al. (Eds.): IWINAC 2017, Part II, LNCS 10338, pp. 213–222, 2017.
DOI: 10.1007/978-3-319-59773-7_22

On the other hand, indirect procedure, that are based on inference approaches using sensors or wearables devices [21]. Nevertheless, these procedures assume that robots and humans share the physical space, but robots working in long time missions, as for instance home assistance, do not have to.

This challenge scenario could be faced through the deployment of new sensors in the robot platform or in the environment, notwithstanding, this research proposes a solution based on gathering environmental sounds using the robot's microphone, due to the microphone range surpass occupied room.

To the best of our knowledge, environmental acoustic information is mainly used for two tasks in the interaction with humans: automatic speech recognition (ASR), and environmental sound recognition (ESR). Both provides an important set of inputs for the decision taking of an autonomous robots. ASR is usually processed using Hidden Markov Models [19] and efficient programming search techniques [8]. ESR has been less faced in the literature [3], and although there are well defined taxonomies [14], they have not been extended and refined as ASR.

We present here a method for improving the activity-context labeling system recognizing acoustic sounds of the environment. The first contribution of this research is the performance improvement of a generic Context Recognition Component (CRC) based on localization, perception and timers caused by the use of the Environment Recognition Component (ERC).

The second contribution is the use of a convolutional neural network for classify the sound detected by the robot. First works on Convolutional Neural Networks (CNN) date back to the early 1980s [5], but nowadays are receiving a great attention [10]. CNN have been widely used for visual recognition contexts, and also successfully applied in music analysis [4], speech [2] and our domain, domestic sound classification [18].

The remainder of this paper is organized as follows. Section 2 presents the proposed framework integrated on an generic hybrid architecture. Section 3 shows the experiment setup and the description of the experiments to test it. Section 4 presents the discussion about the overall experiments. Finally conclusions and future work are presented in Sect. 5.

2 Context-Awareness Framework

In order to achieve the goals of this research we need to propose a framework able to be integrated in any control architecture. In that manner we propose a hybrid approach (Reactive-Deliberative) based on a motivational principles, this means that the decisions are not taking only with sensor information, but also internal motivations as battery status or robot role.

Very briefly, our framework is made up by two components levels: the ERC (Environment Recognition Component) and the Context-Awareness Component (CAC). Both components are deployed in the reactive subsystem which is divided in three blocks: data gathering, data preprocessing/fusion, and low-level reasoning. The ERC is a new element in the data gathering layer of the system, it works directly with low-level data from sensors along with perception nodes. The CAC

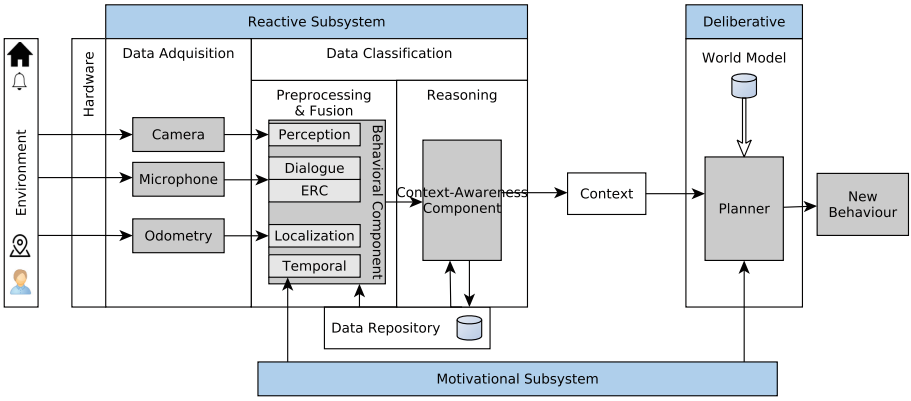

Fig. 1. ERC and CAC components integrated in a generic motivational architecture.

is a preprocessing-data fusion component that uses information from different data acquisition sources. It is deployed along with Natural Language Processing Components or Human recognition Components

Figure 1 shows the input stream associated to low-level sensors: perception, dialogue, ERC (Environment Recognition Component), localization and the timeline of the robot in the environment. Sections below describe in detail the environment-recognition component, in charge of the natural acoustic signal recognition and the CAC.

2.1 Environment Recognition Component

The Environment Recognition Component (ERC in Fig. 1) identifies the sounds perceived by robot microphones. It classifies the environmental sounds using a convolutional deep neural network. It has been developed to identify 14 different relevant sounds, associated to locations or scenarios, grouped into classes:

- General: Door bell, Entry Door, Phone, Door, Silence, Window
- Bathroom: Cistern, tap
- Kitchen: Induction, Fridge, Kettle, Microwave, Oven Alarm

We have used a convolutional neural network to implement this system, using both the sound and its variation. The topology of the neuronal network is shown in Fig. 2. It is composed by these layers:

- The input is a 60×200 matrix. Each of the elements of the matrix is a tuple of the spectrogram value and its variation in time (delta).
- The first convolution ReLU (Rectified Linear Units [6]) layer of 80 filters of shape (57×6) and stride (1×1).
- A max pooling layer of shape (57×6) and stride (1×3).
- We use a dropout layer with probability of 0.5 to reduce over-fitting.

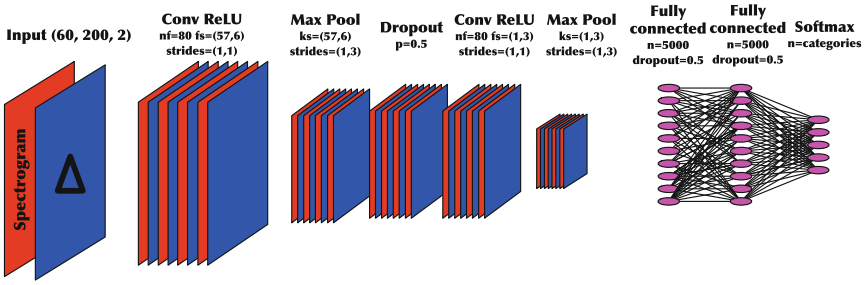

Fig. 2. Convolutional neural network topology

- A second convolution ReLU layer of 80 filters of shape (1×3) and stride (1×1).
- A second max pooling layer of shape (1×3) and stride (1×3).
- A second dropout layer with probability of 0.5.
- Two fully connected hidden layer of 5000 ReLUs each.
- A SoftMax output layer with a neuron for each category of sounds.

The network topology is similar to successful works, like [16], with some differences:

- A silence category which improves the classification results.
- Using a more appropriate duration in the input sound clips, attending to the characteristics of the sound and the robot operation.
- Implementation in TensorFlow [1], which let us to try a great variety of learning algorithms.
- Integration of the trained net in ROS nodes [17] to on board evaluation of sounds while robot operation.

The audio files from database are divided into clips of 2.95 s (200 frames). We think that this is enough to classify domestic sounds whose main characteristic is the monotonous repetition with different intervals, alarms or telephone tones.

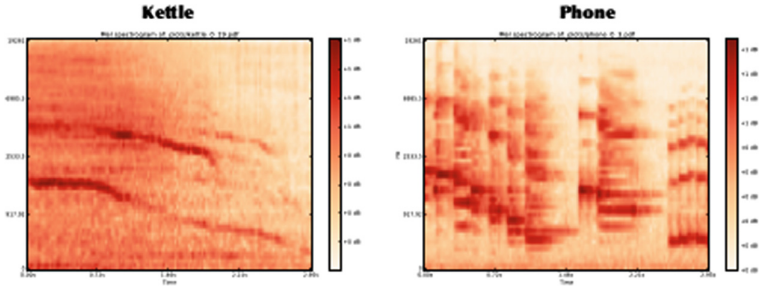

Fig. 3. Examples of normalized and log-scaled spectrograms.

These segments are processed to extract the input patterns for training and evaluating the net:

- Log-scaled mel-spectrograms *spec*, resampled to 22050 Hz and normalized with window size of 1024, hop length of 512 and 60 mel-bands, using the *librosa*[1] implementation. Figure 3 shows examples of the spectrogram of some of the categorized objects.
- The variation in time of this spectrogram $\Delta = \frac{\partial spec}{\partial t}$, computed with default settings.

2.2 Context Awareness Component (CAC)

Many researchers have faced the context awareness inference based on logic based models [13], ontologies [20] or probabilistic approaches [22]. We will formalize it using Bayesian methods [9,15], in particular, our inference system is supported on a Bayesian network approach.

A Bayesian network (BN) is a probabilistic directed acyclic graph generated from a group of random variables and their dependencies. Nodes (random variables) which are connected by arcs (conditional dependencies) compose it.

The definition of the variables of our BN is based on the American Occupational Therapy Association, Inc. (AOTA)[2]. We have determined three hierarchical layers for our BN: (1) class activities, (2) activities, and (3) Observations. These three levels of abstraction allowed us to identify and formalize the elements involved during the daily user activity, the notation used is:

- Observations: represent the information acquired by the robot. We denoted these nodes as; $O = \{o_1, o_2, \ldots, o_n\}$ where the n is the total number of observation defined in our system.
- Activities: These nodes identify the daily activities made by the users. For instance, meal preparation (cooking) or health management (medication control). Each activity is defined by a subset of observations: We denote this as: $P(A|O) = P(a_i|o_1, o_2, o_3, \ldots, o_n)$.
- Class Activities: These nodes identify the class of activity used in the system. In our case we used the eight AOTA class activity definition (ADLs, IADLs, rest,..).

The system works as follows: The robot processes a set of observations. Each activity has set of observations associated which different levels of probability, that identifies the user activity context. With this information we calculate the conditional distribution, the joint probability of all the nodes in our proposal is defined by:

$$P(ClassActivity, Activity, Observation) = \\ P(Observation) \cdot P(Activity|Observation) \cdot P(ClassActivity|Activity, Observation)$$

[1] librosa: v0.3.1 library by B. McFee et al., doi:10.5281/zenodo.12714.
[2] https://www.aota.org/.

At this point we have a set of contexts with different level of probabilities. This information is then used in the behavioral or the deliberative level to take decisions or generate new robot behaviors.

3 Experimental Validation

In the experiments, we wanted to measure the likelihood of positively labeling the user activity context using just classical methods based on Localization, Dialogue and Time of day (LDT) versus the use of environmental sound recognition in addition to the classical methods.

The robot assumes a place at home (kitchen or living room). In two cases, the user is talking to the robot, in a third one there is no user speaking with the robot. In each case, we trigger three of our previously defined acoustic signal during the dialogue scene between user and robot. If ERC has recognized the signal the context subsystem infers the context. In order to analyze the performance of the CAC using the LDT procedure we also performs the same test without using acoustic signals.

3.1 ERC

As we presented in Sect. 2.1, the dataset is composed by sound clips belonging to 14 categories of domestic sounds: door bell, cistern, tap, induction plaque, fridge, entry phone bell, kettle, phone, entry door bell, microwave alarm, door closing, window closing, silence and oven alarm. The clips of each category have different lengths, in a range of [55–371] seconds. To generate the segments of the input data, we have split the whole clip into 88.5% overlapping segments of 2.95 s, with a step of 1 s.

The dataset used for the training phase is balanced, so we use 52 segments (the size of the smaller category) of each category, 95% for training and 5% for validation.

As we previously mentioned, we used TensorFlow framework to train and evaluate the network. For the training process, we used a Stochastic Gradient Descent algorithm with a learning rate of 0.002 a learning rate decay of 0.96 with a decay step of 1000. The training took 14 h in a i7-4960HQ CPU 8 cores @ 2.60 GHz and 16 GB RAM computer for 300 epochs. The result is a net with an accuracy of 86%. The prediction result for the entire dataset is:

```
DB  CI  TA  IP  FR  EP  KE  PH  ED  MW  DO  WI  SI  OV
Door Bell [ 63  5   0   2   0   3   0   0   0   0   0   0   0   0  ]
Cistern   [ 3   162 0   0   0   0   0   0   0   1   0   0   0   0  ]
Tap       [ 3   9   555 3   0   0   62  0   1   0   10  5   2   0  ]
Induc Pla [ 0   0   0   301 0   0   0   0   0   0   0   0   0   0  ]
Fridge    [ 0   0   0   0   99  0   0   0   0   3   1   0   0   0  ]
Entry Pho [ 2   0   0   0   3   95  0   0   0   0   0   0   0   0  ]
Kettle    [ 0   0   0   16  1   0   183 0   0   0   0   0   0   0  ]
Phone     [ 0   0   0   0   0   0   0   104 0   0   0   0   0   0  ]
Entry Door[ 5   26  0   0   3   0   0   11  192 24  9   3   0   0  ]
Microwave [ 0   0   0   22  228 0   0   0   0   558 1   0   0   60 ]
Door      [ 0   0   0   0   0   0   0   0   0   0   224 0   0   0  ]
Window    [ 0   0   0   0   0   0   0   0   0   0   14  130 0   0  ]
Silence   [ 0   0   0   0   0   0   0   0   0   0   0   0   371 0  ]
Oven Alarm[ 0   0   0   0   0   0   0   0   0   0   10  0   3   184]
```

3.2 CAC

Having in mind @home competitions as RoboCup or ERL, we have been tried to formalize a set of scenarios where the dialogue, localization, time of day and acoustic sounds are involved. The proposal divides the tests in six scenarios: (1) and (4) the robot and human stayed at the same location and they have a conversation about an activity which can be performed in their location; (2) and (3) the robot and human stayed at the same place and they have a conversation about an activity which can be performed in other location at home; and (5) and (6) the robot and the human do not have a conversation nor share location context.

Under these six scenarios, the characteristics of the LDT+ Acoustic sounds are:

(a) Dialogue: we reduced the dialogue possibilities to just two, one related to cook something and one related with an upcoming visit.
(b) Localization: robot and user could stay in two places, kitchen and living room, or each one in one place.
(c) Time of day: we fixed the time at 12:00 PM.
(d) Acoustic sounds: we used the signals previously defined and recognized in the ERC section.

We defined three activity contexts: Meal (M) preparation context, that is the probabilities to be cooking something are high. Emergency context (E), the circumstances present a context where something is going wrong, so the user has to make a decision; and Social Interaction (I) context, meaning that the probabilities of interaction with other human are high.

We have used Elvira [11] to evaluate the inferences. Table 1 outlines the results. We have set a threshold to label the context. We have defined a base limit of 50%, under that value we do not recognize the case.

Table 1. Context classification results.

	Kitchen (Robot & Human)				Living Room (Robot & Human)			
Env. Signal	(O)	(F)	(D)	(-)	(O)	(F)	(D)	(-)
D:Dinner	*M*(99%)	*M*(97%)	*M*(97%)	*M*(97%)	*M*(98%)	M(15%)	M(15%)	M(15%)
	E(1%)	*E*(50%)	E(5%)	E(1%)	E(24%)	*E*(25%)	E(5%)	E(5%)
12:00pm	I(1%)	I(1%)	*I*(41%)	I(1%)	I(5%)	I(5%)	*I*(45%)	I(5%)
	Scenario 1				Scenario 2			
D:Visit	*M*(95%)	M(5%)	M(5%)	M(5%)	*M*(83%)	M(1%)	M(1%)	M(1%)
	E(24%)	*E*(50%)	E(5%)	E(1%)	E(24%)	E(25%)	E(5%)	E(5%)
12:00pm	*I*(94%)	*I*(94%)	*I*(98%)	*I*(94%)	*I*(95%)	*I*(95%)	*I*(99%)	*I*(95%)
	Scenario 3				Scenario 4			
Dialogue	Kitchen (Robot alone)				Living Room (Robot alone)			
D:(-)	*M*(95%)	M(5%)	M(5%)	M(1%)	*M*(83%)	M(1%)	M(1%)	M(1%)
	E(24%)	*E*(50%)	E(5%)	E(5%)	E(24%)	*E*(25%)	E(5%)	E(5%)
12:00pm	I(1%)	I(1%)	*I*(41%)	I(5%)	I(5%)	I(5%)	*I*(45%)	I(5%)
	Scenario 5				Scenario 6			

4 Discussion

The experiments in controlled scenarios have shown that the system is very reliable. It was able to successfully recognize the context more than 85% of the times even when random noise was added to the ambient.

In summary, we have developed a functional system for recognizing different acoustic signals in real world and we have identified two main issues. First, environmental noises as loud music or people shouting (this situation is common in robotics competitions) contaminates the ambient sound, thus it increases the number of false positives. Second, the microphone model and position in the robot is a key decision, for instance a directional microphone has drawbacks in indoors environments.

The context recognition component showed positive results in the six scenarios proposed (S1–S6) (Table 1). On the one hand, we have those cases where the robot is able to infer the context using LDT: scenarios S1, S3 and S4. It happens even when there is no acoustic signal triggered. The acoustic signal slightly improve the probability (they are depicted as dark gray cells): the oven increases a 2% the context probability in scenario S1; the doorbell increases 4% context probability in scenario in scenarios S3 and S4. In addition, the acoustic signals provided extra information about the human activity in these scenarios (black cells), with a likelihood within our threshold (fifty percent or more).

We have defined and additional case called *valuable information*. These cases are produced when the final probability is within a threshold between 25% and 50%. Even though these cases should not be used directly to the decision making process, it can be used to generate alternative sub-tasks, for instance the robot can ask about this special case.

Scenario S1 under doorbell signal presents this situation (41%), the robot should ask about if the user is cooking because it is going to receive visits, thus it is able to increase or decrease this probability.

On the other hand, we find those cases where the robot is not able to infer the situation using LDT method: scenarios S2, S5, S6. These cases show overall best results through our proposal.

The robot is not able to infer actual or future activity context of the user using LDT on the scenario S2. However, if the oven signal is recognized the robot knows that there is an activity related with meal preparation running. This scenario has two cases with *valuable information*, fridge signal presents a 25% of an emergency and doorbell shows a 45% of a visit context.

Finally, we have the scenarios S5 and S6 where the robot is not able to infer the situation using LDT because it is not sharing location or dialogue. In these scenarios, our proposal presents better results. We have 95% of certainty in S5 and 83% in S6 if the oven signal is recognized. We have 50% of certainty in S5 that there is an emergency context if the fridge signal is recognized. We also have three cases of *valuable information* that although we do not have the certainty about the context, they give to the robot information about what is happen or what will happen at home.

5 Conclusions

We present a two components framework to recognize and label user activity context in indoor environments based on four elements: Localization Dialogue, timers and acoustic signals. Two major contributions are presented in this paper: a component (ERC) for recognizing environmental acoustic signals and a context-awareness component (CAC) that is able to recognize user activities even when the scenario is not shared between robot and user. The ERC uses a deep convolution neural network for the Environment Recognition Component, that provides 87% of accuracy in the recognition of acoustic signals.

As other authors have pointed out [7] modeling user contexts may seem unnatural if the context consists of problems with solutions. However, the relevance of this information for getting autonomous robots is beyond doubt as we observed in those cases where human and robot do not share space.

The scalability of this solution depends of previous knowledge of user daily life but not by learning. If we propose a solution by learning, we should care about to store historical context data on runtime. In terms of memory, it would generate an uncontrolled growth of past context information.

As future work we are going to add an automatic learning component into ERC able to incorporate new user's environment acoustic signals. As well as a routine analysis system, to extract daily information to predict future context schedule attending user tasks.

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: large-scale machine learning on heterogeneous systems (2015). Software available at http://tensorflow.org/
2. Abdel-Hamid, O., Mohamed, A., Jiang, H., Penn, G.: Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4277–4280. IEEE (2012)
3. Chachada, S., Kuo, C.C.J.: Environmental sound recognition: a survey. In: Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific, pp. 1–9. IEEE (2013)
4. Dieleman, S., Brakel, P., Schrauwen, B.: Audio-based music classification with a pretrained convolutional network. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 669–674. IEEE (2011)
5. Fukushima, K.: Features for content-based audio retrieval. Biol. Cybern. **36**(4), 193–202 (1980)

6. Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier neural networks. In: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, vol. 15, pp. 315–323 (2011)
7. Göker, A., Myrhaug, H.I.: User context and personalisation. In: Workshop proceedings for the 6th European Conference on Case Based Reasoning (2002)
8. Jiang, H.: Confidence measures for speech recognition: a survey. Speech Commun. **45**(4), 455–470 (2005)
9. Korpipaa, P., Mantyjarvi, J., Kela, J., Keranen, H., Malm, E.J.: Managing context information in mobile devices. IEEE Pervasive Comput. **2**(3), 42–51 (2003)
10. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
11. Lacave, C., Luque, M., Díez, F.J.: Explanation of Bayesian networks and influence diagrams in Elvira. Syst. Man Cybern. Part B: Cybern. IEEE Trans. **37**(4), 952–965 (2007)
12. Liao, L., Fox, D., Kautz, H.: Location-based activity recognition. Adv. Neural Inf. Process. Syst. **18**, 787 (2006)
13. McCarthy, J., Buvac, S.: Formalizing context (expanded notes) (1997)
14. Mitrović, D., Zeppelzauer, M., Breiteneder, C.: Features for content-based audio retrieval. Adv. Comput. **78**, 71–150 (2010)
15. Moore, D.J., Essa, I.A., Hayes, M.H.: Exploiting human actions and object context for recognition tasks. In: The Proceedings of the Seventh IEEE International Conference on Computer Vision, 1999, vol. 1, pp. 80–86. IEEE (1999)
16. Piczac, K.: Enviromental sound classification with convolutional neuronal network. In: Proceedings of the 2015 IEEE International Workshop on Machine Learning for Signal Processing. IEEE (2015)
17. Quigley, M., Conley, K., Gerkey, B.P., Faust, J., Foote, T., Leibs, J., Wheeler, R., Ng, A.Y.: ROS: an open-source robot operating system. In: ICRA Workshop on Open Source Software (2009)
18. Salamon, J., Bello, J.P.: Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Process. Lett. (2017)
19. Trentin, E., Gori, M.: A survey of hybrid ann/hmm models for automatic speech recognition. Neurocomputing **37**(1), 91–126 (2001)
20. Wang, X.H., Zhang, D.Q., Gu, T., Pung, H.K.: Ontology based context modeling and reasoning using OWL. In: Proceedings of the Second IEEE Annual Conference on Pervasive Computing and Communications Workshops, 2004, pp. 18–22. IEEE (2004)
21. Zhu, C., Sheng, W.: Motion-and location-based online human daily activity recognition. Pervasive Mobile Comput. **7**(2), 256–269 (2011)
22. Ziebart, B.D., Maas, A.L., Dey, A.K., Bagnell, J.A.: Navigate like a cabbie: probabilistic reasoning from observed context-aware behavior. In: Proceedings of the 10th International Conference on Ubiquitous Computing, pp. 322–331. ACM (2008)