DISCOVERING SOUND CONCEPTS AND ACOUSTIC RELATIONS IN TEXT

Anurag Kumar, Bhiksha Raj, Ndapandula Nakashole

School of Computer Science,
Carnegie Mellon University
Pittsburgh, PA, USA - 15213

alnu@andrew.cmu.edu, bhiksha@cs.cmu.edu,ndapa@cs.cmu.edu

ABSTRACT

In this paper we describe approaches for discovering acoustic concepts and relations in text. The first major goal is to be able to identify text phrases which contain a notion of audibility and can be termed as a sound or an acoustic concept. We also propose a method to define an acoustic scene through a set of sound concepts. We use pattern matching and parts of speech tags to generate sound concepts from large scale text corpora. We use dependency parsing and LSTM recurrent neural network to predict a set of sound concepts for a given acoustic scene. These methods are not only helpful in creating an acoustic knowledge base but in the future can also directly help acoustic event and scene detection research.

Index Terms— Sound Concepts, Audio Event Detection, Acoustic Scenes, Acoustic Relations

1. INTRODUCTION

Analyzing non-speech content has been gaining a lot of attention in the audio community. Such non-speech audio content plays an important role in understanding the environment around us. Successful detection of acoustic events and scenes is critical for several applications. One of the most prominent applications is content based retrieval of multimedia recordings [1] [2], where the audio component of multimedia carries a significant amount of information. Other well known applications of automated analysis of audio data are: audio based surveillance [3], human computer interaction [4], classification of bird species [5], context recognition system [6].

The primary focus in automated machine understanding of non-speech content of audio has been on successful detection and classification of audio events and scenes. Several methods have been proposed for audio event and scene detection [7–11]. In the most recent DCASE challenge [1], deep neural network methods dominated performance for audio events whereas factor analysis and Non Negative Matrix Factorization methods were found to be more promising for acoustic scenes. Moreover, due to the limited availability of labeled data and the time consuming and expensive process of manual annotations, there have been attempts to learn event detectors from weakly labeled data as well [12] [13]. These methods rely on weak labels which can be automatically obtained for audio data on web using the associated metadata such as tags, and titles.

One major limitation in most of the current literature on audio content analysis is the limited vocabulary of audio events. In almost all cases, the analysis is done on a very small set of $5 - 20$ acoustic events. Clearly, this is very small from for several of the applications we have mentioned. More importantly, mere detection and classification of audio events does not lead to a comprehensive understanding of acoustic concepts. If we look at the analogous problem in the field of computer vision, one can notice that object detection in images has been scaled to thousands of visual object categories [14]. Moreover, these thousands of categories are organized into a hierarchical structure which allows higher level semantic analysis. Visual concept ontologies [15] have been proposed for reducing dependence on text based retrieval of images. The *Never Ending Image Learner* (NEIL) project [16] not only detects thousands of visual objects and scenes in images but has also learned a variety of commonsense knowledge visual relationships such as *Umbrella looks similar to Ferris wheel* or scene-object relation such as *Monitor is found in Control room*. This allows it to provide visual knowledge to knowledge bases such as Never Ending Language Learner (NELL) [17]. Another architecture EventNet [18] is tailored towards multimedia events. It organizes 500 multimedia events using over 4000 visual concepts. Clearly, these knowledge relations and ontologies are crucial for semantic search of multimedia data on web.

A similar architecture is desirable for sounds as well which is not only aware of a large number of acoustic concepts but can also draw higher level semantic information and relations about sounds. For example, the system should be aware that *honking*, *beeping* and *engine running* can be related through a common source *car*. Even more important are scene-sounds acoustic relations such as, an acoustic scene *Park* consists of sounds event *children laughter*. Some works have tried to relate sounds through hand crafted sound taxonomies [19] [20] [21] [22]. Taxonomies for environmental sounds has been of particular interest in these works. Even in the specific context of environmental sounds there is no clear consensus on building such taxonomies [19]. Different approaches have been applied in different cases and in most cases it is based on subjective opinions. Moreover, in several of these urban taxonomies, a large part of the taxonomy is made up of broad categories and the number of low-level acoustic concepts is once again very small. This limits their utility for both accumulating sound related knowledge as well as for audio event and scene detection research. In this paper, we take a step towards large scale understanding of sound by addressing some of these issues. The motivation is to develop methods which can automatically catalog sounds and generate other commonsense knowledge about sounds. Although not a component of this paper, this can definitely aid in audio event and scene detection tasks.

First, we try to address the problems of acoustic concept vocabulary by proposing methods for automated discovery of potential sound concepts by applying natural language processing and machine learning techniques on a large corpus of text. For automated discovery of sound concepts we propose a simple yet effective unsupervised approach based on part of speech (POS) patterns. We

[1] http://www.cs.tut.fi/sgn/arg/dcase2016/

follow up on this step by proposing a word embedding based supervised method for classifying a given text phrase into sound phrase (concept) or non sound phrase (non-sound concept). This supervised method allows us to identify the notion of audibility in any given text phrase. Acoustic relations such as acoustic scene- sound concepts, sounds and sources are equally important for understanding sounds. For acoustic relations, we propose a method for automatically describing an acoustic scene or environment through sound concepts. Sound concepts and acoustic scenes are related through dependency paths and then an LSTM neural network [23] is used to predict whether a sound concept is found in an acoustic scene or not. Although, in this work we looked into the specific case of *scene-concept* relations, our method can be extended to other forms of relations as well, for example *concept-source* relations. Learning scene-concept relations can be extremely helpful in creating sound ontologies. To the best of our knowledge, this is the first work on text based understanding of sound concepts and acoustic relations leading to large scale acoustic knowledge. The rest of the paper is as follows.

In Section 2 we describe our methods for discovering sound concepts, in Section 3 we describe the dependency path and LSTM based approach for *scene-concept* relations. We describe our subjective and objective evaluation for the proposed methods in Section 4 and finally we conclude in Section 5.

2. SOUND CONCEPTS

Automated discovery of text phrases which can be designated as sound concepts is a very difficult problem. First, whether a text phrase has a notion of sound or audibility in it can be very subjective and dependent on the context in which it is used. There are unigram and bigram phrases such as *music, laughter, glass breaking, woman screaming* etc. which on its own gives an impression of sound or audibility in it. However, in several cases a direct belief of sound might not be apparent from the text phrase on its own but it is either a source of or is directly related to salient acoustic phenomena which is well understood in commonsense human knowledge. Examples include phrases such as *helicopter, birds, dog, car engine* . These acoustic concepts appear in several audio event databases and it is expected to have capabilities to detect these events. Hence, any large list of sound concepts should include such phrases. Automated discovery of sound concepts (phrases) becomes difficult due to this subjective and contextual way of expressing sound concepts. We propose a simple yet very effective method of obtaining "audible" text phrases or sound concepts from large text corpora. We also propose a supervised method of classifying a given text phrase as sound concept or non-sound concept by incorporating syntactic and semantic content of phrase through word embeddings.

2.1. Unsupervised Sound Concept Discovery

We introduce an unsupervised method for discovering sound concepts in text. Our method is based on the idea that there are patterns in specific forms that are primarily used to express sound concepts in language. These patterns can help in identifying sound concepts.

We begin with a single pattern: "sound(s) of <Y>" where Y is any phrase, we allow Y to be up to 4 words long. We then look for occurrences of the pattern in a large corpus. In our experiments, we used the English part of ClueWeb09[2] From this we obtain a large collection of occurrences such as: "sound of *honking cars*", "sound of *gunshots*".

However, this step produces a significant amount of noise. We therefore treat its output as *candidate sound concepts* and introduce a minimally-supervised method for pruning noise from this collection. First, we generalize candidate concepts by replacing mentions with their part of speech tags, as follows:

$$\text{sound of honking cars} = \text{sound of VBG NN}$$
$$\text{sound of gunshots} = \text{sound of NNS}$$

where the part of speech (POS) tag VBG denotes verbs in the gerund form, NN, and NNS denote singular and plural nouns, respectively [3]. The POS generalized concepts reduce the data size to about only 20 unique patterns. Since the POS patterns are so few, we can use them to filter out noisy concepts with little effort. The key to filtering is that not all POS patterns express valid concepts. We can eliminate all but 6 of the POS patterns. For example, the pattern "sound of JJ (adjective)" does not express sound concepts. All candidate concepts that match the 6 valid POS patterns are retained and the rest are discarded. The full list of valid POS patterns with examples is shown in Table 1. The patterns in Table 1 produced a total of $116,729$ unique sound mentions from the corpus.

	Pattern	Example Concept
P1	<X> of (DT) VBG NN(S)	honking cars
P2	<X> of VBG	yelling
P3	<X> of (DT) NN(S) VBG	dogs barking
P4	<X> of (DT) NN(S)	gunshots
P5	<X> of (DT) NN NN(S)	string quartet
P6	<X> of (DT) JJ NN(S)	classical music

Table 1. Patterns for discovering sound concepts in text. VBG is the part of speech tag for verbs in the gerund form, NN for singular nouns (S means plular), DT for determiners, and JJ for adjectives.

2.2. Supervised Classification

The unsupervised discovery of sound phrases (concepts) in the previous section can still contain non-sound phrases. A few examples of such phrases which are clearly not sound concepts but do not get filtered out by the two step process in the previous section are *someone being (NN VBG), price dropping (NN VBG), gaining experience (VBG NN), happy hunters (JJ NNS)*. Hence, to improve upon the unsupervised discovery of potential sound concepts, we propose a supervised method for classifying a text phrase as sound phrase (sound concept) or non sound phrase (non sound concept).

Since bigram phrases are the most dominant and expressive set of sound concepts discovered by the unsupervised method, we focus specifically on bigram phrases. A set of labeled data is required for supervised training of classifiers for text phrase classification. To obtain a completely reliable set of labeled data, we manually inspect a small subset of the sound concepts obtained in the previous section and mark if it is actually a sound concept or not. Note that, in the unsupervised case only 6 POS patterns express valid sound concepts. We use the rest of the POS patterns to create a list of negative examples. We manually inspect and label a small subset of this list as well. Finally, we end up with a total of $\sim$ 6000, sound concept and non-sound concept phrases.

The text phrases need to be appropriately represented by features on which classifiers can be trained. *Word Embeddings* have been found to be very effective in capturing syntactic and semantic similarity between words [24–26] and have shown remarkable success in a variety of semantic tasks [27]. Word embeddings map words into a fixed dimensional vector representation. In this work we use *word2vec* [24] to obtain vector representation for words in

[2] http://lemurproject.org/clueweb09/

[3] POS Abbreviations: https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Table 2. 36 Acoustic environments used in experiments

	Acoustic Environments			
1	Office	Farm	House	Bus
2	Parties	Funeral	Library	Park
3	Street	Parking Lot	Church	Train
4	Airplane	Wedding	Cafe	Cities
5	Campus	Ballgame	Bathroom	Classroom
6	Train Station	School	Parks	Bar
7	Grocery Store	Trucks	Forest	Restaurant
8	Subway	Airport	Arena	Construction
9	Beach	Garden	Stadium	Ranch

text phrases. We use Google News pre-trained embeddings [4] to represent each word by 300 dimensional vectors. We then use two methods for representing each bigram phrase. In the first case, we take the average of the word2vec representation for each word to represent the whole phrase. We refer to this representation as *AWV*. In the second case, we concatenate the vector representation (*CWV*) for each word to obtain a 600 dimensional vector for each phrase. These vectors can then be used for training any classifier. We use Support Vector Machine (SVM) for training the sound and non-sound phrase classifier.

3. ACOUSTIC RELATIONS

In this section we describe an approach to learn relationships in the domain of acoustic world or *acoustic relations*. Acoustic relations can be of different forms such as *acoustic scene - concepts* relations: sound concepts found in an acoustic scene, *source-sound* relations: source of the corresponding sound, co-occurrence relations: sounds which often occur together. In this paper we focus specifically on *scene-concept* relations where our goal is to describe an acoustic scene or environment by a set of sounds which occur in that scene or environment. Information in the form of what types of sounds make up a scene can be extremely helpful in audio scene recognition tasks[1]. Moreover, these relations also provide co-occurrence information about sound concepts. For example, *laughing* and *cheering* often occur together in several acoustic environment. These additional information about sound concepts can be exploited in a variety of applications. From the perspective of semantic analysis in text, we cast this task as a relation classification problem.

First we find all sentences in the ClueWeb corpus that mention at least one of the $116,729$ sound concepts discovered in Section 2, and at least one acoustic environment such as “beach”, “park”, etc. In our experiments, we worked with a total of 36 acoustic environments which we define in Table 2, but our method is generic and can work with any number of environments. Most of acoustic scenes from DCASE [1] scene classification challenge are part of our setup as well. We then apply a dependency parser[5] to any sentence that mentions a sound concept and an acoustic environment. This step produces dependencies that form a directed graph, with words being nodes and dependencies being edges. For example, the sentence: *“The park was filled with the sound of children playing”* , yields the following dependencies:

Table 3. Example Paths for Positive Training Data

prep_along()	*prep_of() sound nsubjpass() heard prep_in()*
prep_of()	*nsubjpass() filled prep_with() sound prep_of()*
prep_of() sound prep_on()	*conj_and() sounds prep_of()*
prep_with() sounds prep_of()	*prep_of() sound prep_to()*
nsubj() alive prep_with() sound prep_of()	*prep_upon()*
prep_of() sounds prep_from()	*prep_of() sounds prep_on()*
prep_of() sound nsubj() came prep_from()	*prep_of() sounds prep_at()*

Table 4. Example Paths for Negative Training Data

conj_and()	*amod()*
poss()	*nn()*
nn() sound prep_of()	*prep_through()*
prep_of()	*appos()*
det()	*conj_and() sound prep_of()*
prep_to()	*prep_of() sound nsubj() filled dobj()*

det(park-2, The-1)
nsubjpass(filled-4, park-2)
auxpass(filled-4, was-3)
root(ROOT-0, filled-4)
det(sound-7, the-6)
nsubj(playing-10, sound-7)
prep_of(sound-7, children-9)
prepc_with(filled-4, playing-10)’

The details of the dependency relations can be found in [28]. Next, we traverse the dependency graph in order to obtain the path between the mention of a sound concept, in this case “children playing”, and the mention of the acoustic environment “park”. Shortest paths between entities have been found to be a good indicator of relationships between entities [29, 30]. We therefore extract the shortest path. In our example, the shortest path labeled with edge and node names is as follows: “nsubjpass() filled prepc_with() sound prep_of()”.

3.1. Training Data

Given the paths, we would like to classify scene-sound pairs into those that express the relationship of interest (SoundFoundInEnvironment) and those that do not. Classifier training would require labeled training data.

To obtain training data, we proceed as follows: We sort the paths by frequency, that is, how often we have seen the path occur with different scene-sound pairs. Among the most frequent paths, we label the paths yes or no, depending on whether they express the relationship of interest. This gives us a way to generate positive and negative examples using the labeled paths. Examples of paths that generate positive training data are shown in Table 3. Examples of paths that generate negative training data are shown in Table 4.

3.2. Classification

We use an LSTM recurrent neural network to learn the scene-sound relationship. Each word w is mapped to a d-dimensional vector $\boldsymbol{v_w} \in \mathbb{R}^d$ through an embedding matrix $\boldsymbol{E} \in \mathbb{R}^{|V| \times d}$, where $|V|$ is the vocabulary size, and each row corresponds to a vector of a word. We initialize the word embeddings with the 300-dimensional Google News pre-trained embeddings[4]. For the dependency relations in the path, we randomly initialize their vector embeddings, and learn them during training.

Path Encoding. To encode the shortest path between a sound concept and an acoustic scene, we use an LSTM recurrent neural networks (RNN) which is capable of learning long range dependencies. While regular RNNs can also learn long dependencies, they tend be biased towards recent inputs in the sequence. LSTMs tackle this limitation with a memory cell and an adaptive gating mechanism that

[4] https://code.google.com/archive/p/word2vec/

[5] https://pypi.python.org/pypi/practnlptools/1.0

Table 5. Analysis of Unsupervised Sound Concept Discovery

	Pattern	# Concept	+ in Top 100 Freq.
P1	<*X*> of (DT) VBG NN(S)	9335	98
P2	<*X*> of VBG	1395	71
P3	<*X*> of (DT) NN(S) VBG	19194	91
P4	<*X*> of (DT) NN(S)	20064	59
P5	<*X*> of (DT) NN NN(S)	26473	93
P6	<*X*> of (DT) JJ NN(S)	40268	49

controls how much of the input to give to the memory cell, and the how much of the previous state to forget [23].

We have a path: $\boldsymbol{p} = \boldsymbol{p}_1, ..., \boldsymbol{p}_p \in \mathbb{R}^d$ and an associated path matrix $\boldsymbol{P} \in \mathbb{R}^{p \times d}$, where each row corresponds to the embedding vector of the word in that position.

The LSTM mention encoder generates the path encoding, $\boldsymbol{v_p}$, as follows:

$$\begin{aligned} \boldsymbol{h}_i &= LSTM(\boldsymbol{v}_{p_i}, \boldsymbol{h}_{i-1}, \boldsymbol{c}_{i-1}), i = 1, \ldots, p \qquad (1) \\ \boldsymbol{v}_p &= \boldsymbol{h}_i : i = p \end{aligned}$$

The LSTM encodes the word at timestep $i = t$ in the path using the word embedding vector $\boldsymbol{v}_{p_t}$, the previous output $\boldsymbol{h}_{t-1}$, and the previous state of the LSTM cell $\boldsymbol{c}_{t-1}$. The output $\boldsymbol{h}_t$ is computed using the four main elements in the LSTM cell: an input gate $\boldsymbol{i}_t$, a forget gate $\boldsymbol{f}_t$, an output gate $\boldsymbol{o}_t$, a memory cell $\boldsymbol{c}_t$ with a self-recurrent connection. The cell takes as input a d-dimensional input vector for word $\boldsymbol{x}_t = \boldsymbol{p}_i$, the previous hidden state $\boldsymbol{h}_{t-1}$, and the memory cell $\boldsymbol{c}_{t-1}$. It calculates the new vectors using the following equations:

$$\begin{aligned} \boldsymbol{i}_t &= \sigma\left(\boldsymbol{W}_{xi}\boldsymbol{x}_t + \boldsymbol{U}_{hi}h_{t-1} + \boldsymbol{b}_i\right), \qquad (2) \\ \boldsymbol{f}_t &= \sigma\left(\boldsymbol{W}_{xf}\boldsymbol{x}_t + \boldsymbol{U}_{hf}\boldsymbol{h}_{t-1} + \boldsymbol{b}_f\right), \\ \boldsymbol{o}_t &= \sigma\left(\boldsymbol{W}_{xo}\boldsymbol{x}_t + \boldsymbol{U}_{ho}\boldsymbol{h}_{t-1} + \boldsymbol{b}_o\right), \\ \boldsymbol{u}_t &= \tanh\left(\boldsymbol{W}_{xu}\boldsymbol{x}_t + \boldsymbol{U}_{hu}\boldsymbol{h}_{t-1} + \boldsymbol{b}_u\right), \\ \boldsymbol{c}_t &= \boldsymbol{i}_t \odot \boldsymbol{u}_t + \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1}, \\ \boldsymbol{h}_t &= \boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t), \end{aligned}$$

where σ is the sigmoid function, $\odot$ is element-wise multiplication, the W and U parameters are weight matrices, and the b parameters are bias vectors.

Prediction. From path encoding $\boldsymbol{v}_p$, we compute the output of the neural network, a distribution over the positive and negative labels. The output for each path is decoded by a linear layer and a *softmax* layer into probabilities over the two labels. Therefore, the prediction $\boldsymbol{d}_r$

$$\boldsymbol{d}_r = \mathrm{softmax}(\boldsymbol{W}_r \cdot \boldsymbol{v}_p) \qquad (3)$$

where $\mathrm{softmax}(z_i) = e^{z_i} / \sum_j e^{z_j}$.

4. ANALYSIS AND EVALUATION

In this Section we analyze and evaluate our proposed methods. The complete list of sound concepts discovered by our method is available on this [31] webpage. We had six unique POS patterns for filtering candidate sound concepts. The total number of sound concepts corresponding to each POS pattern is shown in Table 5. Since, the total number of sound concepts discovered is fairly large, manually inspecting it to identify actual sound concepts among the discovered ones is very difficult. For each concept discovered by our method we maintain a count of total number of times that sound concept occurred in the text corpus. We select the *Top 100* for manual inspection and identify concepts which can actually be labeled as a *sound concept*. We do it for each POS pattern. The number of positive (+) hits in this Top 100 most frequent concepts is shown in Table 5.

Table 6. Accuracy of Supervised Classification

	Fold 1	Fold 2	Fold 3	Fold 4	Avg
AWV	87.03	89.05	87.84	89.77	88.42
CWV	90.00	89.32	91.87	90.30	90.37

Table 7. Examples of Environment (Scene)-sounds relations discovered by our method

Environment	**Sounds**
Forest	Birds Singing, Breaking Twigs, Cooing, Falling Water
Restaurant	Jazz, Laughter, People Talking, Music Drifting
Airport	Planes Flying, Plane Engines, Aircraft, Intercoms
Park	Laughing, Police Siren, Birds Chirping, Footsteps
Ranch	Horses, Gunfire, Tapping Water, Bulldozers
Church	Children Laughing, Church Bells, Singing, Applause
Beach	Waves Crashing, Waves Lapping, Surf Hitting
Construction	Hammering, Jackhammers, Engines, Blasting
Street	Sirens, Men Shouting, Honking Cars, Cheering
Bar	Piano Playing, Laughter, Clinking Glasses, Cheering

We note that 3 POS pattern have more than 90 positives. The lowest is for <JJ NN(S) >. However, note that the frequency of occurrence of a discovered concept has nothing to do with it being *truly* a sound concept. Discovered concepts such as *sobbing voices, siren breaking, cheering crewmen* contain an impression of sound and are positive examples of sound concepts but occur very few times in the text corpus. Hence, the purpose of column 2 in Table 5 is to show how well our method did on phrases which occurred frequently in our process.

As described in Section 2.2, we created a list of sound concepts (positive) and non-sound concepts (negative) bigram phrases. The total number of positive examples is 3189 and total number of negative examples is 2758. We randomly divide this data into 4 folds. 3 folds are used for training and then the trained model is tested on left out fold. The experiment is done all 4 ways. Linear SVMs are trained on both *AWV* features and *CWV* features. The accuracies for both feature representation are shown in Table 6. Concatenated word2vec features gives slightly better performance compared to averaged word2vec features. An average accuracy of more than 90% is achieved which shows that our supervised classifier is highly reliable in classifying a text phrase as sound or non-sound phrase.

Table 7 shows a few examples of *scene-concept* relations found by the system. Some unusual findings are *Rifle Shots* in *Library*, *Chirping Birds* in *Library*. The full list of sound concepts discovered for each acoustic scene or environment is available on this [31] webpage. A subjective analysis of all discovered relations shows that for most of the relations discovered are meaningful in the sense that the sound concept is actually found in that acoustic environment.

5. CONCLUSIONS

In this paper we presented methods for text based understanding of sounds and acoustic relations. We proposed a method for automated discovery of sound concepts using a large text corpus. It discovered over $100,000$ sound concepts and to the best of our knowledge no such other exhaustive list of sound concepts exists in current literature. We found among the discovered concepts, those corresponding to POS patterns in form of <VBG NN(S) >and <NN(S) VBG >are in general very reliable. This is clearly expected as a large number of sounds are related to its sources through some action. We also proposed a simple word embedding based method for learning supervised classification of text phrases in sound or non-sound phrases (concepts). This supervised method achieved an accuracy of over 90%. Although, the total number of examples considered in supervised classification experiments is not very large (~ 6000), it does validate our proposed word embedding based approach. An important aspect of any knowledge base about sounds would be to relate different sounds. In this work we took on the specific case of scene-concept relations where we try to find out the sound concepts which may occur in an acoustic scene or environment. This is helpful in defining an acoustic scene by the sounds which occur in that scene and hence it can be exploited in acoustic scene recognition tasks. Moreover, it also allows us to relate (co-occurrence) of sound concepts through common scenes. Other meta level inferences can also be drawn from such acoustic relations. We continue to investigate in this direction.

6. REFERENCES

[1] Shoou-I Yu, Lu Jiang, Zexi Mao, Xiaojun Chang, Xingzhong Du, Chuang Gan, Zhenzhong Lan, Zhongwen Xu, Xuanchong Li, Yang Cai, et al., “Informedia@ trecvid 2014 med and mer,” in *NIST TRECVID Video Retrieval Evaluation Workshop*, 2014.

[2] Flora Amato, Luca Greco, Fabio Persia, Silvestro Roberto Poccia, and Aniello De Santo, “Content-based multimedia retrieval,” in *Data Management in Pervasive Systems*, pp. 291–310. Springer, 2015.

[3] Pradeep K Atrey, Namunu C Maddage, and Mohan S Kankanhalli, “Audio based event detection for multimedia surveillance,” in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*. IEEE, 2006, vol. 5.

[4] Maxime Janvier, Xavier Alameda-Pineda, Laurent Girinz, and Radu Horaud, “Sound-event recognition with a companion humanoid,” in *2012 12th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2012)*. IEEE, 2012, pp. 104–111.

[5] Seppo Fagerlund, “Bird species recognition using support vector machines,” *EURASIP Journal on Applied Signal Processing*, vol. 2007, no. 1, pp. 64–64, 2007.

[6] Antti J Eronen, Vesa T Peltonen, Juha T Tuomi, Anssi P Klapuri, Seppo Fagerlund, Timo Sorsa, Gaëtan Lorho, and Jyri Huopaniemi, “Audio-based context recognition,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, pp. 321–329, 2006.

[7] Xiaodan Zhuang, Xi Zhou, Mark A Hasegawa-Johnson, and Thomas S Huang, “Real-world acoustic event detection,” *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1543–1551, 2010.

[8] A Kumar, P Dighe, R Singh, S Chaudhuri, and B Raj, “Audio event detection from acoustic unit occurrence patterns,” in *IEEE ICASSP*, 2012, pp. 489–492.

[9] Pasquale Foggia, Nicolai Petkov, Alessia Saggese, Nicola Strisciuglio, and Mario Vento, “Reliable detection of audio events in highly noisy environments,” *Pattern Recognition Letters*, vol. 65, pp. 22–28, 2015.

[10] Zvi Kons and Orith Toledo-Ronen, “Audio event classification using deep neural networks.,” in *Interspeech*, 2013.

[11] Emre Cakir, Toni Heittola, Heikki Huttunen, and Tuomas Virtanen, “Polyphonic sound event detection using multi label deep neural networks,” in *2015 international joint conference on neural networks (IJCNN)*. IEEE, 2015, pp. 1–7.

[12] Anurag Kumar and Bhiksha Raj, “Audio event detection using weakly labeled data,” in *24th ACM International Conference on Multimedia*. ACM Multimedia, 2016.

[13] Anurag Kumar and Bhiksha Raj, “Weakly supervised scalable audio content analysis,” in *2016 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2016.

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.

[15] VCO, “Visual concept ontology,” `http://http://disa.fi.muni.cz/results/software/visual-concept-ontology/`.

[16] Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta, "NEIL: Extracting Visual Knowledge from Web Data," in *International Conference on Computer Vision (ICCV)*, 2013, `http://www.neil-kb.com/`.

[17] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka Jr, and Tom M Mitchell, "Toward an architecture for never-ending language learning.," in *AAAI*, 2010, vol. 5, p. 3.

[18] Guangnan Ye, Yitong Li, Hongliang Xu, Dong Liu, and Shih-Fu Chang, "Eventnet: A large scale structured concept library for complex event detection in video," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 471–480.

[19] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 1041–1044.

[20] R Murray Schafer, *The soundscape: Our sonic environment and the tuning of the world*, Inner Traditions/Bear & Co, 1993.

[21] AL Brown, Jian Kang, and Truls Gjestland, "Towards standardization in soundscape preference assessment," *Applied Acoustics*, vol. 72, no. 6, pp. 387–392, 2011.

[22] Manon Raimbault and Daniele Dubois, "Urban soundscapes: Experiences and knowledge," *Cities*, vol. 22, no. 5, pp. 339–350, 2005.

[23] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[24] T Mikolov and J Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, 2013.

[25] Jeffrey Pennington, Richard Socher, and Christopher D Manning, "Glove: Global vectors for word representation.," in *EMNLP*, 2014, vol. 14, pp. 1532–43.

[26] Kazuma Hashimoto, Pontus Stenetorp, Makoto Miwa, and Yoshimasa Tsuruoka, "Task-oriented learning of word embeddings for semantic relation classification," *CoNLL 2015*, p. 268, 2015.

[27] Marco Baroni, Georgiana Dinu, and Germán Kruszewski, "Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors.," in *ACL (1)*, 2014, pp. 238–247.

[28] Marie-Catherine De Marneffe and Christopher D Manning, "Stanford typed dependencies manual," Tech. Rep., 2008.

[29] Yan Xu, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, and Zhi Jin, "Classifying relations via long short term memory networks along shortest dependency paths," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, 2015, pp. 1785–1794.

[30] Ndapandula Nakashole, Gerhard Weikum, and Fabian M. Suchanek, "Discovering semantic relations from the web and organizing them with PATTY," *SIGMOD Record*, vol. 42, no. 2, pp. 29–34, 2013.

[31] A Kumar, "Sound concepts and relations," `http://www.cs.cmu.edu/%7Ealnu/SOExpt.htm` Copy and Paste in browser if clicking does not work.