What to do first: the initial behavior in a multi-sensory household object recognition and categorization system

Haojun Guan, Weipeng He and Jianwei Zhang
TAMS group, Department of Informatiks
University of Hamburg
Vogt-Koelln-Strasse 30
Email: guan,2he,zhang@informatik.uni-hamburg.de

***Abstract*—Numbers of studies in Biological area have already shown that our human beings have the amazing ability to quickly extract the physical properties of an object from the sound it made and recognize at least the material of it [1][2]. The interactive exploratory behaviours are used as a "question" which "ask" to the object from human, and the feedback, maybe a feeling or a sound, as the "answer". Such an intelligent system can also be developed based on the research above, the ability of recognize and category objects is also necessary for the intelligent system and robot especially service robot in the near future. In this paper, a novel framework of a multi-sensory household object recognition and categorization system is introduced. Both auditory and visual sensory are used in this novel system, the detail of the part of auditory is presented in detail, and the result of the experiments to decide which interactive behaviour is most suitable act as the initial behaviour is also shown.**

I. Introduction

Learning to classify objects to different categories is a milestone in the history of human evolution, by which people can explore the world. It is a complicated task which need the cooperation of all the sense of human. Human beings use interactive exploratory behaviours to understand the world, such as by lifting an object the weight can be know, by scratching the roughness of the object can be found and by pressing the object the hardness can be detected. Auditory data is as crucial as visual information because from sound we can find something which we can not get from visual information [3]. Such a cooperative approach is also an important and necessary for an intelligent system.

There are three main limitations of the currently exist approaches in the area of object recognition and categorization. First, most of these approaches are using only visual data, which means they are rely completely on the 2D image, 3D point cloud or laser scan data [4][5][6][7]. These approaches can reach very high recognition and classification accuracy by giving a clear view of an object. In the other hand, a study in psychology has already point out that some other properties (eg.,material, roughness, hardness, weight, etc.) of objects can only be detected by other modalities, like auditory, instead of visual sensory modality. For example, it is not possible or very hard to distinguish between a white porcelain mug and a mug made of white plastic by only using visual sensory, but it is a simple task for auditory sensory. The second main limitation of currently approaches is only few approach use the natural or physical sound which produced by the object itself [8]. Others use human as a supervisor to train the classifier by telling the system what exactly the object is, which means the problem transform from recognize the sound produced by the object to speech recognition [9]. The last limitation is all of these approaches can be divided into two phases which are data collection phase and data processing phase. They all directly perform all the behaviours on the object in the same phase, and processing all the feedback all together to find the result. Two main problems may arise because of this, which is also the motivations of the framework be presented in this paper. The first one is the behaviour *drop* may cause the broken of some breakable objects. The second problem is that, during the whole data collection phase some pair of behaviour and object may produce a silent sound feedback which means nothing. For example, if the behaviour *grasp* be performed on a plush toy, the auditory sensor can not detect any sound during the whole process. As a research report has pointed out that as the number of sensor using and data collecting increase in a whole research process, the ratio of data usage decrease. It means a

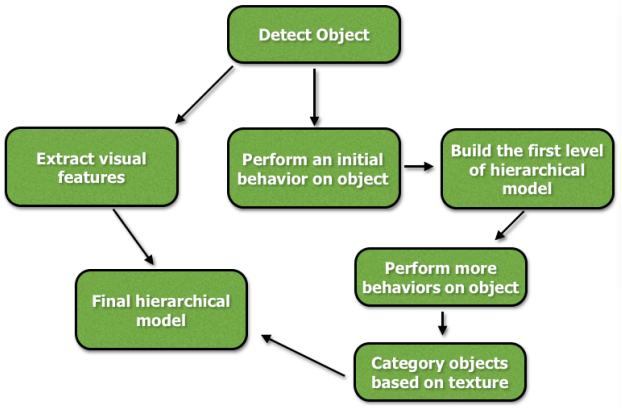

Fig. 1. An overview of the system framework. After the object be detected, only one initial behaviour is performed on the object to detect whether the object is breakable (glass or porcelain in this case) or not. The whole data set is separated into two subset based on the result of the first behaviour, two different behaviours will then be performed respectively to the two set of objects. A hierarchical modal is built based on the result of the two extra behaviours which is the material type of each object. The visual features are extracted at the same time after detecting the object, and these features are used to help building a more detailed hierarchical model with the auditory sensory result.

part of data are useless, they are not helping to get better performance or high accuracy, but cost the same computation resource and time as useful data.

To address these limitations, this paper presents the process of how to deal with the collected auditory data and how to decide the suitable initial behaviour for the novel framework of multi-sensory object recognition and categorization system (shown in Fig.1). Due to the motivations described above, only an initial behaviour is performed on the object in the first phase, after analysis the feedback data from the initial behaviour, the whole dataset of objects can be separated into two subset. Two more different behaviours are performed on the object afterwards depends on the result of the first one. A hierarchical model is built based on the results of these two extra behaviours. The framework is evaluated on 20 different household objects with 3 different behaviours (*drop, push and knock*). Both k-Nearest Neighbour (k-NN) and Support Vector Machine (SVM) machine learning algorithms are used to train a classifier in the experiments. The result shows which pair of behaviour and machine learning algorithm is the most suitable act as the initial behaviour.

The remainder of the paper is organized as follows: section 2 introduces some related work; section 3 shows the experimental setup; section 4 describes theory of how to deal with the auditory data; section 5 reports the result of the experiments; and finally in section 6, conclusions are made, and some future research issue are discussed.

II. Related Work

A briefly overview of the area auditory-based object recognition is made in this section. There are few studies work on how can an intelligent system recognize objects by only auditory information. Krotkov *et al.* [10] is one of the earliest person work in this area. The material type of the objects in their experiments can be recognized by interpreting the sound which produced by hitting on objects with 3 robs made of different material. The significance of their work is, their results proved that the spectrogram of sound can be used as a powerful representation for distinguish between objects made of different material (plastic, wood, glass, brass, aluminum in their study). Richmond *et al.* [11] have presented that different material types can be detected from contact sound by modelling the spectrogram of the sounds. Torres-Jara *et al.* [12] have shown that a novel object can also be recognized using the sounds produced when tapping on it. The spectrogram of the novel sound matched to another spectrogram which is already in the training set. Their approach can successfully make prediction for four different objects made of vary materials. More recently, Jivko *et al.* [13] presented an approach which allow a robot platform to recognize 36 different household objects by using both auditory and proprioceptive properties. 5 exploratory behaviours were used to detect these properties, and a relation map of all the objects were also made by using a relational learning method.

These previous studies above make a solid foundation of the further research in this area, but none of them considered the motivations of this paper which has already described before. 20 different objects made of 6 different material type are used for the experiments in this proposing paper, include both breakable and non-breakable objects. A most suitable initial behaviour is decided using the analysed auditory information.

III. Experimental Setup

A. Objects

The set of target objects, $\mathcal{O}$ include 20 different objects (shown in Fig. 2) made of 6 different material type (glass, metal, paper, plastic, porcelain and wood). Within the dataset, there are 10 breakable objects and 10 non-breakable object. All the objects are very common household things which can be easily found in our daily life. They are all selected from the office or home from one of the authors. Most of the objects have the ability to hold liquid inside, but they are all set to be empty.

Fig. 2. The dataset of objects, $\mathcal{O}$ include 20 different household objects made of 6 material type with several colours and patterns. Some of them have texts on them, but the texts are not used to help the recognition and categorization work.

B. Behaviours

The set of interactively exploratory behaviours, $\mathcal{B}$ consists of three behaviours: *knock, push and drop*. All of these three behaviours are used to produce sound by interacting with objects. Except the behaviour *drop*, both behaviours *knock* and *push* are the candidates of the initial behaviour in the experiments which are proposed in this paper.

C. Sound dataset and Sound Recording

Due to the contact sounds produced by behaviours and objects are quite diverse, so many factors can lead to very different result, there does not exist a public dataset in this field. It means the dataset used to evaluate the system should be built by the authors themselves.

A recorder pen made by Sanyo with a built-in stereo microphone is used to record the sound during the execution of each behaviour. Each sound clip lasts 7-10 seconds, and stores in the form of MP3. By considering that some objects

are not suitable for the behaviour *drop*, there are in total 10×3+10×2=50 pairs of behaviour and object are collected by the MP3 recorder. All the sound clips are captured in a quiet environment without any noise. 2 sound sequences are extracted from each sound clips, so in total there are 50×2=100 sequences be collected.

In the next section, the detail of the auditory data processing is described, include the machine learning algorithms which are used in the system as well.

IV. Theoretical Model

A. From sounds to sequences

Each sound is collected in the form of MP3, it is not possible to compute the similarity of neither two MP3 sound clips nor the raw sound wave data (shown in Fig. 4), so it should be transformed to sequences which are suitable for a machine learning algorithm to computer the similarity between them. To address this problem, a Self-Organizing Map is used to present each sound as a sequence,S_i. To achieve such a representation, firstly the recorded raw sound data are resampled to 8000 Hz, and then short-time Fourier transform (STFT) is used to extract features from each sound. The short-time Fourier transform is computed for each sound with Hann windows of size 256 samples and with shift size of 128. The magnitudes of the spectrograms (4 samples of the spectrograms are shown in Fig. 5) is 256 / 2 + 1 = 129 dimensional, are then taken as the input feature representation for the SOM. The whole flowchart of the process is shown in Figure 3.

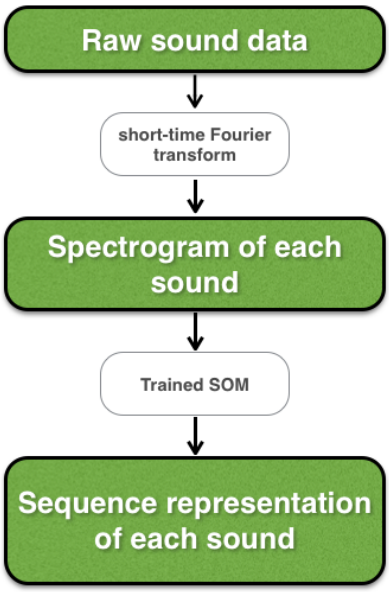

Fig. 3. There are 3 main phases in the whole process: raw sound data record and collect; sound features extract by short-time Fourier transform and present the sound sequences by using SOM. The raw sound data is collected in the first phase, after the second phase the raw sound data have been transformed to spectrogram, and the result of the whole process is sequences representation of each sound.

Let G_i be a spectrogram of a sound, such G_i can be seen as a set of column vectors, $G_i = [v_1^i, v_2^i ... v_j^i]$, where each v_j^i is a 129 dimensional column vector in a spectrogram at time point j. Let $\mathcal{G} = \{G_i\}_{i=1}^{N}$ be the whole collection of all the spectrograms, a dataset of column vectors can be sampled from $\mathcal{G}$ as the input data and the training set for the two dimensional *Self-Organizing Map*. The size of the SOM is set to be 20 by 20, in total of 400 nodes. Figure 6 gives a overview of the SOM training process.

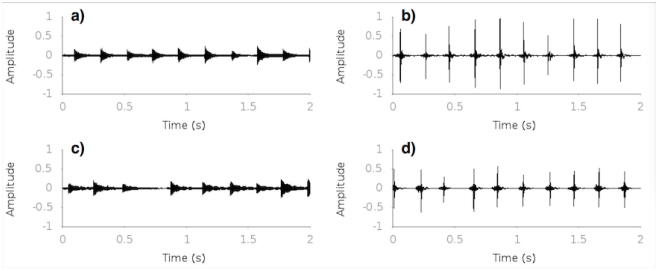

Fig. 4. The raw sound wave data of objects 1 ,5, 7, 9, when the behaviour *knock* is performed on them.The horizontal dimension denotes time(s), while the vertical axis denotes amplitude. a) the wave data of object 1, *Porcelain Bowl*; b) is object 5, *Paper Box*; c) is object 7, *Glass*; d) present the object 9; *Metal can Coke*. It can be easily detected that the figures a) and c) are more similar and b) and d) should be in the same group, which is the objective fact.

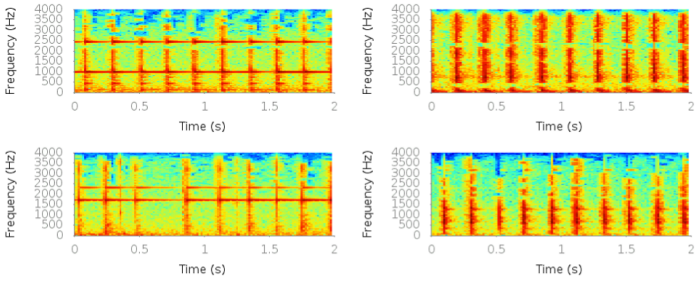

Fig. 5. The spectrogram of objects 1 ,5, 7, 9 calculated by short-time Fourier transform, when the behaviour *knock* is performed on them. The horizontal dimension denotes time(s), while the vertical axis denotes frequency(Hz). a) the wave data of object 1, *Porcelain Bowl*; b) is object 5, *Paper Box*; c) is object 7, *Glass*; d) present the object 9; *Metal can Coke*. It can be easily detected that the figures a) and c) are more similar and b) and d) should be in the same group, which is the objective fact.

After training the *Self-Organizing Map* by $\mathcal{G}$, each spectrogram, G_i can be mapped to a sequence of nodes, S_i, in SOM by mapping each column vector, v_j^i, of the spectrogram, G_i, with each node,s_j^i, in the map. Thus, each sound clip is mapped and represented by a sequence of nodes, $S_i = s_1^i s_2^i ... s_j^i$.

A machine learning algorithm with a similarity function is used in this approach, to calculate how similar two sequences, S_i and S_j, are. A similarity function, $NW(S_i, S_j)$ is defined in the propose approach. The Needleman-Wunsch algorithm [14] is used to measure the similarity between two sequences, while this algorithm is usually used to align protein and nucleotide sequences in bioinformatics.

B. Collection of represented data

The set of exploratory interactively behaviours, $\mathcal{B}$=[*knock, push, drop*], two or three behaviours in $\mathcal{B}$ are performed on each object, depends on whether it is breakable or not. For each pair of object and behaviour, (O_i, B_i), where $O_i \in \mathcal{O}$ is the object which is used in this trial, and $B_i \in \mathcal{B}$ is the

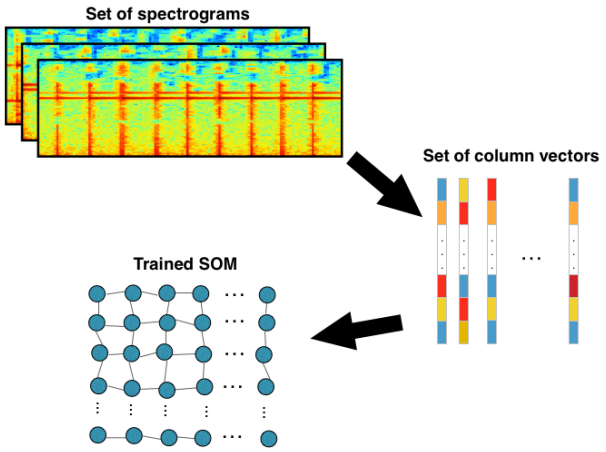

Fig. 6. The overview of training *Self-Organizing Map*: After the raw sound data are transformed to spectrograms, the column vectors are extracted from each spectrogram, and used as the input data to train the SOM by mapping each vector to a node in the *Self-Organizing Map*.

executed behaviour, 5 trials are performed on the object. In total (10×3+10×2)×5=250 interaction trials (according to some of objects in $\mathcal{O}$ are not suitable for the behaviour *drop*) are recorded with the MP3 recorder, and then two sound clips are extracted for each interaction trial. Due to only the behaviours *knock* and *push* are the candidates of the initial behaviour, so there are 200×2=400 sound clips be used is the experiments in this paper. During the i^{th} trial, the auditory data is present in a triple form(O_i, B_i, S_i), where $S_i = s_1^i s_2^i ... s_{l^i}^i$ is the sequence of nodes in *Self-Organizing Map* used to present the produced sound. Use one world to concluded the form (O_i, B_i, S_i) is, it represent the sound sequence S_i which is detected and recorded when the behaviour B_i is performed on the object O_i.

The task of the system is to learn a suitable model which can estimate the object whether it is breakable or not, by giving a sound sequence S_i. A machine learning algorithm which can calculated the similarity between two sequences S_i and S_j is needed to help the system to achieve the goal. The estimation results should be presented in the form, $Pr(O_i = o)$, where o is one of the object class label which is already stored in the machine learning algorithm and O_i is the input pending object. Both k-Nearest Neighbour (k-NN) and Support Vector Machine (SVM) are used to solve this problem. The next subsection describes the way of using them.

C. Machine learning algorithms

1) k-Nearest Neighbour: k-Nearest Neighbour (k-NN) is one of the simplest machine learning algorithm, which is widely used in both the areas of classification and regression. Instead of building a model for the input dataset, k-NN only simply store all the data points and their label. To make the prediction of the input test data point, k-NN finds the *k* nearest neighbours for the input data point, and assigns the input data point to the class most common among its *k* nearest neighbours. In the experiments which are presented in the next section, the problem transforms to, given a test sequence S_i, k-NN finds the the sequences in the training data set with the highest similarity to S_i.

In all the experiments described in the next section, *k* is set to 5, and the normalized global alignment score, $NW(S_i, S_j)$, is used as the similarity function for the k-NN algorithm. An prediction of which class an input data belongs to, $Pr(O_i = o)$, is obtained by counting the labels of the *k* nearest classes of it. For example, one of five nearest neighbour has the object class label *Glass Bottle* and the other four neighbours have the label *Porcelain Mug*, then $Pr(O_1 = Percelain\ Mug) = \frac{4}{5}$ and $Pr(O_1 = Glass\ Bottle) = \frac{1}{5}$. The object O_1 should belong to the class *Porcelain Mug* in this case.

2) Support Vector Machine: Support Vector Machine (SVM) classifier is a supervised learning model with associated learning methods which falls within the family of discriminative models. Let $\mathcal{D}$ be a set of labelled inputs,

$$D = \{(\mathbf{x}_i), y_i) \mid \mathbf{x}_i \in \mathbf{R}^p, y_i \in \{-1, 1\}\}_{i=1}^n \qquad (1)$$

where the y_i is either 1 or -1, indicating which class the point $\mathbf{x}_i$ belongs to. A linear decision function $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} - b$ with the maximum margin should be learned by the Support Vector Machine, which is used to clearly classify the input data points. However, a good linear decision function can not always be found. If the input data set are not linear separable, the input data points have to be mapped to a high-dimensional space, where they can be separated there. A kernel function,$K(X_i, X_j) = \phi(X_i) \bullet \phi(X_j)$ is needed to make the mapping process more efficiently, which can also be considered as a measure of how similar two data points are.

Due to the truth that, SVM is a suitable machine learning algorithm used to solve the binary classification problem, which is exactly match the requirement in this case. In other word, in the experiments which are described in the next section, the machine learning algorithm is used to separate the whole input data set into two subset, which are *breakable* and *non-breakable*. A power function of $NW(S_i, S_j)$, which is the normalized global alignment score, is used as the kernel function, $K(S_i, S_j) = NW(Si, Sj)^p$, where p is set to 5, in this case. During the prediction phase, each trained SVM votes one of the subsets, and all the votes are collected to get a final result.

In the next section, the detail of experiments and result is presented, from the design of the experiments to the result of two groups of them.

V. Experiments and Result

A. Experiments design

To achieve the motivations, experiments for determining the suitable initial behaviour (classify the objects whether they are breakable or not) are needed as the first step of the whole framework and system. This subsection presents how the experiments are designed and organised. As it has been described in previous section, two behaviours (*knock* and *push*) are the candidates of the initial behaviour, and two machine learning algorithms (k-Nearest Neighbour and Support Vector Machine) are used to classify the input data. In total 4 pairs of behaviour and machine learning algorithm, B_i, M_i are tested in this section (*kock+k-NN; knock+SVM; push+k-NN; push+SVM*). The experiments present in this paper can be separate to two parts, in the first series of experiments, 8

sound clips for each object are used to train the classifiers (machine learning algorithm), the last 2 clips for each object is used to evaluate the accuracy of recognition. In the second one, the sound clips from 18 objects are used to train the classifier, and the rest 2 objects are used as the novel object to test the accuracy of categorization of the system. In the first series of experiments, in total 160 sound clips are used to train the classifier, and the rest 40 clips are used to test the accuracy of this model, while the numbers in the second series of experiments are 180 and 20. The performance of the framework is evaluated in term of the percentage of the correct results:

$$Accuracy = \frac{correct\ predictions}{total\ predictions} \times 100\% \qquad (2)$$

B. Experiments on recognition

In the first experiment, in total 20×8=160 sound sequences, S_i, are used to train either k-Nearest Neighbour or Support Vector Machine classifier for each behaviour, and the rest 40 sequences are used to evaluate the recognition accuracy for each behaviour. For the k-Nearest Neighbour classifier, the accuracy is computed in the following way: Collect the results of all the classifiers for each object O_i, because of *k* is already set to 5 for all the experiments in this paper, in total 10 trained classifiers vote for the prediction of O_i, for example, three fifth of the first sequence S_{i1} nearest neighbours have the class label *breakable*, and four fifth of the second sequence S_{i2} nearest neighbours have the class label *breakable*, then $Pr(O_i = breakable) = \frac{3}{5} + \frac{4}{5} = \frac{7}{10} = 70\%$. After collecting all the data, the average final accuracy for a pair of behaviour and machine learning algorithm, (B_i, M_i), can be calculated by,

$$Average\ accuracy = \frac{\sum_{i=1}^{20} Pr(O_i)}{20} \qquad (3)$$

For the Support Vector Machine, the method is similar, but instead of only the 5 nearest neighbours have the permission to vote for the prediction, all the trained classifiers are used to make the final prediction.

TABLE I shows the results of all the 4 pairs (*kock+k-NN; knock+SVM; push+k-NN; push+SVM*) in the first series of experiments. Both k-NN and SVM perform a good performance in this series of experiments, where the best choice is the pair *konck+k-NN* with the accuracy 98.5% and the pair *push+SVM* gets the lowest accuracy (92.74%), which is also a decent result. The accuracy for each individual object is also shown in this table, most of them reach a high accuracy (above 95%), but there are still some pairs of object and behaviour, (O_i, B_i), receive a low accuracy, such as *push+SVM* for O_{10}, it is only 75.63% and for O_{14} the results for all four pairs do not reach a very high accuracy, which the highest is the pair of *knock+SVM* with the result of 94.38%, it may cause by the sound clip itself. In one word to conclude the results of this series of experiments is, all the four groups show a quite good performance, where the pair of *knock+ k-NN* is most suitable for the initial behaviour in this kind of task.

C. Experiments on accuracy of categorization

In the second experiment, all the sound clips of 18 objects are used to train the classifier, so in total there are 18×10=180 sound sequences, S_i, and the rest 20 sequences from two objects are used to evaluate the accuracy of categorization. After training all the classifiers, the result of algorithm k-NN is computed in the following way: Collect the nearest neighbours for all 10 sound sequences, S_i, of one object, so in total 10×5=50 nearest neighbours have the permission to vote the class label for an individual object, and the final accuracy of one pair of learning algorithm and behaviour,(B_i, M_i), can be computed by equation (3). For the SVM classifier, as the same method in the first experiment, instead of using 50 classifiers to vote the result, all the trained classifiers are used to make the prediction for one individual object.

TABLE I. OBJECT RECOGNITION RESULTS

Objects	*knock*		*push*	
	k-NN	SVM	k-NN	SVM
1	100%	98.13%	100%	94.38%
2	100%	97.17%	100%	98.13%
3	100%	99.38%	100%	97.5%
4	90%	96.88%	100%	98.13%
5	100%	86.25%	100%	95.63%
6	100%	98.75%	80%	93.75%
7	100%	75.63%	100%	97.5%
8	100%	95.63%	100%	99.38%
9	100%	96.25%	100%	95.63%
10	100%	80.63%	80%	75.63%
11	100%	93.75%	80%	98.13%
12	100%	99.38%	100%	94.38%
13	100%	97.5%	100%	80%
14	80%	94.38%	90%	80.63%
15	100%	95%	100%	97.17%
16	100%	100%	100%	96.25%
17	100%	80%	100%	96.25%
18	100%	100%	100%	99.38%
19	100%	95.63%	100%	86.25%
20	100%	93.75%	100%	80.63%
Average	98.5%	93.71%	96.5%	92.74%

TABLE II shows the categorization results of the 4 pairs of algorithm and behaviour, (M_i, B_i). All the results is not so good compare to the results in the first experiment, where the highest accuracy is achieved by the pair *push+k-NN* with the percentage 83% and the lowest result is 49.2% which is made by the pair *knock+k-NN*. The results for every individual object are also shown in this table, which present the accuracy is vary depending on the performance of the sound sequence. For example, the object O_{13}, both *k-NN* and *SVM* achieve quite low results (38% for *k-NN* and 52.78% for *SVM*) with the behaviour *knock* for this object, while the results for the behaviour *push* is acceptable (90% for *k-NN* and 80.56% for *SVM*). Currently the pair of *push+k-NN* is the most suitable pair act as the initial behaviour, but it need some more improvement on performance.

To summary this section, the first series of experiments is quite successful, where all the four groups get good performance, but there are still some points can be improved in the future work, such as the sound clips of O_{14} should be checked clearly or even re-recording. For the novel object categorization experiments, the results still not satisfied and have a large space for improvement.

VI. CONCLUSION

This paper presents a novel framework of multi-sensory object recognition and categorization system. The decision of the initial behaviour is also described in detail, which is one of the most important step in this system. To make the choice, each sound is firstly recorded by a MP3 recorder, and then be transformed from a high-dimensional sound spectrogram to

TABLE II. OBJECT CATEGORIZATION RESULTS

Objects	*knock*		*push*	
	k-NN	SVM	k-NN	SVM
1	46%	57.78%	90%	83.89%
2	48%	57.22%	88%	79.44%
3	52%	50%	60%	75.56%
4	54%	51.57%	78%	52.78%
5	56%	55%	76%	73.89%
6	38%	51.11%	88%	63.89%
7	48%	48.89%	86%	75%
8	52%	52.78%	84%	76.67%
9	54%	55.56%	76%	67.22%
10	78%	46.67%	68%	75.63%
11	32%	46.11%	88%	67.22%
12	36%	45%	94%	83.33%
13	38%	52.78%	90%	80.56%
14	40%	43.89%	66%	83.89%
15	42%	52.22%	86%	80%
16	48%	42.78%	84%	73.89%
17	56%	40%	96%	82.22%
18	54%	51.67%	88%	73.33%
19	52%	37.22%	78%	86.25%
20	60%	56.67%	96%	81.67%
Average	49.2%	49.75%	83%	75.81%

a low-dimensional sequence by Self-Organizing Map(SOM). Either learning algorithm k-Nearest Neighbour(k-NN) or Support Vector Machine(SVM) is used to calculate the similarity of the sequence between objects. The framework is evaluated using 20 household objects made of 6 different material type and 2 candidates of the initial behaviour: *push* and *knock*, an extra behaviour *drop* is also used to collected more auditory information from the objects.

The results showed that the pair *knock+k-NN* can reach the highest accuracy (99%)for the sound clip recognition, while the pair *push+k-NN* is most suitable (with the accuracy of 83%) for the novel object categorization. It also showed that the other 3 pairs of algorithms in the first experiment have a good performance too. The model for the novel objects categorization should be improved in the future, some more input data are needed to train the classifier.

There are several possible future work. First, as the description in section 3, it said that there is no public dataset in this field of work, a public dataset of the behaviour, objects and sound can be built. Second, the performance of the whole framework can be improved by verifying every most suitable pairs of behaviour and machine learning algorithm in each step. Also, some more experiments can be focus on the diversity of a same object with or without liquid in it. The last but not the least is the presented framework can be demonstrate on a service robot to help it understand the environment around it much easier.

ACKNOWLEDGMENT

This work is supported and founded by the DFG German Research Foundation (grant #1247) International Research Training Group CINACS (Cross-modal Interactions in Natural and Artificial Cognitive Systems).

REFERENCES

[1] W. W. Gaver, *"What in the world do we hear? An ecological approach to auditory event perception"*, Ecological Psychology, vol. 5, pp. 129, 1993

[2] M. Grassi, *"Do we hear size or sound? Balls dropped on plates"*, Perception and Psychophysics, vol. 67, no. 2, pp. 274284, 2005.

[3] D. Norman, *"The Design of Everyday Things. Doubleday"*, 1988.

[4] R.B. Rusu, Z.C. Marton, N. Blodow, M. Dolha, and M. Beetz, *"Towards 3D point cloud based object maps for household environments"*, Robotics and Autonomous Systems, vol. 56, no. 11, pp. 927 941, 2008.

[5] S. Srinivasa et al., *"HERB: a home exploring robotic butler"*, Autonomous Robots, vol. 28, no. 1, pp. 520, 2010.

[6] F. Endres, C. Plagemann, C. amd Stachniss, and W. Burgand, *"Unsupervised discovery of object classes from range data using latent Dirichlet allocation"*, Robotics: Science and Systems, vol. 2, pp. 113120, 2009

[7] J. Zhang, J.W. Zhang, S. Chen,Y. Hu and H. Guan, *"Constructing Dynamic Category Hierarchies for Novel Visual Category DIscovery"*, IEEE International Conference on Intelligent Robots and Systems(IROS), 2012

[8] J. Sinapov and A. Stoychev, *"Object Category Recognition by a Humanoid Robot Using Behavior-Grounded Relational Learning"*, IEEE International Conference on Robotics and Automation(ICRA), 2011

[9] T.Nakamura, T. Nagai and Naoto Iwahashi *"Bag of Multimodal Hierarchical Dirichlet Processes: Model of Complex Conceptual Structure for Intelligent Robots"*, IEEE International Conference on Intelligent Robots and Systems(IROS), 2012

[10] E. Krotkov, R. Klatzky, N. Zumel *"Robotic perception of material: Experiments with shape-invariant acoustic measures of material type"*, Experimental Robotics IV Lecture Notes in Control and Information Sciences Volume 223, 1997, pp 204-211

[11] J. Richinond and D. Pai *"Active Measurement of Contact Sounds"*, IEEE International Conference on Robotics and Automation(ICRA), 2000

[12] E. Torres-Jara, L. Natale and P. Fitzpatrick*"Tapping into Touch"*, Lund University Cognitive Studies, 2005, pp 22-24

[13] J. Sinapov, M. Wiemer and A. Stoychev, *"Interactive Learning of the Acoustic Properties of Household Objects"*, IEEE International Conference on Robotics and Automation(ICRA), 2009

[14] S. Needleman and C. Wunsch, *"A general method applicable to the search for similarities in the amino acid sequence of two proteins"*, J. Mol. Biol., vol. 48, 1970, pp443-453

[15] G. Navarro, *"IA guided tour to approximate string matching"*, ACM Computing Surveys (CSUR) Volume 33 Issue 1, March 2001, pp 31-88

[16] Y.W. Teh, M.I. Jordan, M.J. Beal and D.M. Blei, *"Hierarchical Dirichlet processes"*, Journal of the American Statistical Association, vol.101, 2006, pp.1566-1581

[17] K.R. Canini, M.M. Shashkov, and T.L. Griffiths, *"Modeling transfer learning in human categorization with the hierarchical Dirichlet process"*, International Conference on Machine Learning, June 2010, pp.151158

[18] N. Iwahashi, *"Robots that learn language: Developmental approach to human-machine conversations"*, Symbol Grounding and Beyond, 2006, pp.143-167

[19] B. Kehoe, A. Matsukawa, S. Candido, J. Kuffner and K. Goldberg, *"Cloud-Based Robot Grasping with the Google Object Recognition Engine"*, IEEE International Conference on Robotics and Automation(ICRA), 2013

[20] G. Evangelopoulos, K. Rapantzikos, P. Maragos, Y. Avrithis, A. Potamianos, *"Audiovisual Attention Modeling and Salient Event Detection"*, Multimodal Processing and Interaction Multimedia Systems and Applications Volume 33, 2008, pp 1-21

[21] S. Mcadams, *"Recognition of sound sources and events"*, in "Thinking in Sound: The Cognitive Psychology of Human Audition", Oxford University Press, 1993

[22] J.B. Fritz, M. Elhilali, S.V. David and S.A. Shamma, *"Auditory attention focusing the searchlight on sound"*, Current Opinion in Neurobiology, vol.17, Issue 4, August 2007, pp 437455

[23] D. Lynott and L. Connell, *"Modality exclusivity norms for 423 object properties"*, Behavior Research Methods, vol.41, Issue 2, May 2009, pp 558-564

[24] S. Takamukua, K. Hosodab and M. Asadac, *"Object Category Acquisition by Dynamic Touch"*, Advanced Robotics, vol.22, Issue 10, 2008

[25] J. Sinapov, T. Bergquist, C. Schenck, U. Ohiri, S. Griffith and A. Stoytchev, *"Interactive object recognition using proprioceptive and auditory feedback"*, The International Journal of Robotics Research, vol. 30, Issue 10, September 2011, pp 1250-1262