2012 IEEE RO-MAN: The 21st IEEE International Symposium on
Robot and Human Interactive Communication.
September 9-13, 2012. Paris, France.

Audio-Based Human Activity Recognition Using Non-Markovian Ensemble Voting

Johannes A. Stork Luciano Spinello Jens Silva Kai O. Arras

Abstract— **Human activity recognition is a key component for socially enabled robots to effectively and naturally interact with humans. In this paper we exploit the fact that many human activities produce characteristic sounds from which a robot can infer the corresponding actions. We propose a novel recognition approach called Non-Markovian Ensemble Voting (NEV) able to classify multiple human activities in an on-line fashion without the need for silence detection or audio stream segmentation. Moreover, the method can deal with activities that are extended over undefined periods in time. In a series of experiments in real reverberant environments, we are able to robustly recognize 22 different sounds that correspond to a number of human activities in a bathroom and kitchen context. Our method outperforms several established classification techniques.**

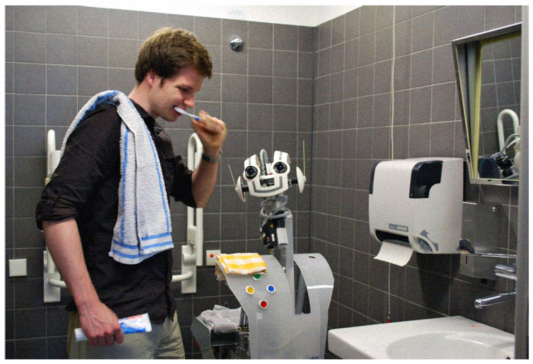

Fig. 1. The capacity to recognize human activities is fundamental to many scenarios including companionship and assistive applications.

I. Introduction

Social robots that share a space with people require the capacity to detect and track humans and recognize their activities. This knowledge is key for effectively integrating robots into people's workflows, as well as natural human-robot interaction in a variety of scenarios (see Fig. 1).

Popular sensory modalities for this task are computer vision and 3D range imaging. Image data provide rich scene information and, by today, allow for accurate body pose estimates even from a single view. Recently, 3D range or RGB-D data have also become popular for human body pose estimation. Body pose can then be used to recognize a large class of human activities. However, these modalities are limited to the field of view of the imaging sensor and not robust over all ranges of environmental conditions. Furthermore, posture information cannot always provide unique evidence about the actions a human is engaged in as very different activities can be carried out in similar body poses. In contrast, the approach we take in this paper employs auditory perception since many human activities produce very characteristic sounds from which a robot can effectively infer the corresponding human actions. Having said this, we do not see audio as a replacement rather than a complement to existing sensory modalities, to be fused for particularly robust activity recognition over wide ranges of conditions.

To date, audio-based human activity recognition has been addressed in the wearable computing community [4, 15], for auditory surveillance systems [5, 9, 3] and multimedia systems [24, 23]. The work of [5] evaluates an audio-based 'context' recognition system for recognizing different indoor and outdoor environments. They evaluate several kinds of audio features used as input of a Hidden Markov Model (HMM). [4] uses a sound recognition framework based on HMMs to recognize context from environmental audio. They make use of this technique to create a wearable platform aware of its audio environment. [15] uses Linear Discriminant Analysis and HMM to process data from body-worn accelerometers and microphones to detect 21 wood workshop activities. [9] addresses the task of acoustic surveillance of events occurring in typical office environments. Several audio features are evaluated and used to classify sound events. [5] aims to recognize several kinds of indoor/outdoor environments from audio streams. They evaluate several audio features and classifiers in 26 different scenes. [3] applies audio recognition to the task of monitoring bathroom activities. They use HMM and Mel Frequency Cepstral Coefficient (MFCC) features to detect bathroom-related sound activities. The works of [24, 23] use audio to classify and segment audio-visual streams in the context of audio visualization and audio indexing.

A large body of work exists in the field of audio-based feature extraction. Mel Frequency Cepstral Coefficients [10] are one of the most robust features in this area. Even though designed for the task of speech recognition they have been used for describing a large number of different sound classes [1, 3, 5] which is why we will also use them in this work. Several authors have used various machine learning techniques to classify sound categories such as AdaBoost [8], Support Vector Machines [7] and Vector Quantization classification [18]. Other works, e.g. [5], make use of HMMs to consider sound categories as sequences of small sound samples. Here, we propose random forests as classifier for this task and present a systematic comparison with the above

All authors were with the Social Robotics Lab, Department of Computer Science, University of Freiburg, Germany. Via e-mail: jastork@kth.se and {spinello,silvaj,arras}@informatik.uni-freiburg.de.

mentioned three alternative approaches.

The field of robot audition is typically concerned with problems at signal-level including sound source localization [16], sound source separation [20], echo cancellation [19], or ego-noise compensation [11]. These problems are relevant to make audio a robust sensory cue over a wide range of conditions. Unlike these works, we address a high-level recognition problems using audio, a little explored area apart from speech. We will neglect signal-level factors at this point and focus on the introduction of the recognition method to demonstrate that audio can be highly useful sensory modality for human activity recognition.

To this end, we propose a novel audio-based recognition technique, Non-Markovian Ensemble Voting (NEV), an on-line classification approach able to predict events in the past and future, filling gaps of missing information. Previous works rely on audio stream segmentation to recognize human activities [9, 24, 15, 1, 18, 23] usually achieved through silence detection, detection of abrupt feature changes or even manual annotation. Our method does not require any segmentation and is able to compute an estimate in an on-line fashion. Methods that do not rely on segmentation either make use of batch processing [5, 25] assuming a minimum time duration of activities, or classify only short-duration audio features [8, 10] which is unlikely to be a robust approach in the noisy conditions robots typically encounter. Our approach integrates information over time to come up with a prediction result and is able to refine its estimate with more incoming information over time. Our approach has been inspired by visual codebook approaches [13, 17] and by the work of Wang [22]. In fact, our technique can be interpreted as a generalization of the latter, also known as *Shazam*'s recognition algorithm. The approach recognizes audio snippets in a large database of songs that makes use of a hash-based voting. Our technique instead is able to generalize several categories of sounds and deal with multiple classification errors in the voting phase.

Clearly, identical human activities can differ largely in their durations. Approaches based on sequential models (e.g. HMM [4, 3]) or based on short-duration audio features [8, 10] cannot generalize well over such variable-length activities. The reason is that first-order Markovian methods such as HMMs rely on a limited duration model that assumes exponentially distributed duration probabilities of each state. Our approach, in contrast, is able to estimate human activities with different durations with great flexibility. No parameters have to be changed or tuned.

Finally, we evaluate our method extensively by comparing four different classifiers for MFCC features and two different high-level classification methods using standard performance indices.

II. Audio-based Human Activity Recognition

In this section we present Non-Markovian Ensemble Voting (NEV), an audio-based recognition algorithm able to classify activities of different duration from a continuous audio stream. The procedure consists in three separate steps

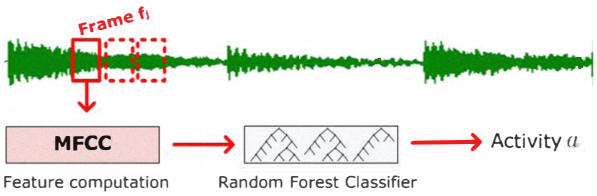

Fig. 2. Frame-by-frame recognition (FBF). MFCC feature descriptors are computed in short-duration audio frames $\mathbf{f}_i$ and then classified using a learned Random Forest (RF) classifier. The output is a predicted activity label a.

feature extraction, frame-by-frame recognition, and Non-Markovian ensemble voting.

A. Feature extraction from raw audio data

The audio stream is subdivided into short-duration segments called frames. A frame $\mathbf{f}_i$ collects 40 ms of audio data. Consecutive frames are designed to overlap by 87.5% of their duration to ensure a high level of data correlation. For each $\mathbf{f}_i$ at time index $t_{\mathbf{f}_i}$, an MFCC feature descriptor $\mathbf{x}_i \in \mathbb{R}^P$ is computed.

B. Frame-by-frame recognition (FBF)

Each MFCC feature descriptor $\mathbf{x}_i \in \mathbb{R}^P$ is classified using a learned Random Forest (RF) [2] classifier $h(\mathbf{x}_i)$ to estimate the activity label a, see Figure 2. A Random Forest classifier is a supervised ensemble classification method that makes use of multiple randomized decision trees to subdivide the feature space. In our case, the RF classifier is trained also with a *background* class, that is a class that contains sounds unrelated to human activities. It is worth to mention that all sounds of a human activity not included in the training set are classified as background sounds. The process of frame-by-frame audio classification (FBF) runs in a continuous fashion and computes an estimated human activity a for each frame $\mathbf{f}_i$:

$$h(\mathbf{x_i}) = a \qquad \text{for all} \quad a = \{0, 1, ..., N\} \tag{1}$$

where N is the number of activities considered. Frame-level classification is a key property for the on-line capability of the proposed approach because it provides an instantly available recognition result with little delay to the current time. At the per-frame level, classification is still unreliable which is due to both, the natural variability in human actions and signal-level factors such as background- or ego-noise, low signal-to-noise ratio, or reverberation effects. All factors are likely to affect the audio content in a short-duration frame and can cause misclassification. This motivates the third stage, described hereafter.

C. Non-Markovian ensemble voting recognition

Non-Markovian ensemble voting (NEV) is an *on-line* method, robust to short-term noise that builds upon the classification output of the previous stage to recognize multiple activities of variable length that can occur in any moment in time. The training procedure of NEV consists in building a *soundbook* for each human activity. This

ID	Activity	# trn/tst	Experimental conditions
1	No human activity	19029/17372	Background noise, various sounds
2	Opening a food bag	15771/15774	Opening/shaking three kind of bags
3	Mixing with a blender	11843/11837	One blender in a small kitchen
4	Pouring cornflakes into a bowl	7101/7057	Three different bowls
5	Eating cornflakes	8626/8210	Heavy crunching, close distance
6	Pouring water into a cup	4338/4525	Several cups
7	Using a dishwasher (humming)	17357/17568	Two different dishwashers
8	Shaving with electric razor	15911/16476	Three different electric razors
9	Sorting flatwares	7903/7916	Sorting flatware into two different boxes
10	Using a food processor	7123/6682	One food processor at different speeds
11	Using a hairdryer	13017/13026	Hairdryer at different speeds, five models
12	Cooking with a microwave	18192/18201	Three different microwave ovens
13	Switching off a microwave oven	2368/2333	End chime of three different ovens
14	Closing a microwave oven door	8411/8388	Pushing the lock/unlock button
15	Sorting dishes	19288/19425	Dishes of different size and shape
16	Stirring water in a cup	11860/11504	Several kinds of cups
17	Flushing a toilet	11808/12261	Four different toiled brands
18	Brushing teeth	6095/5608	Brushing, recorded at close distance
19	Using a vacuum cleaner	15356/1573	Three different models
20	Using a washing machine	13405/13014	Three different operating modes
21	Boiling water	12628-13033	Electrical boiler, three different models
22	Using the water tap	38120/22469	Water splashing from a faucet into a sink

TABLE I

LIST OF HUMAN ACTIVITIES. THE CENTRAL COLUMN GIVES THE NUMBER OF MFCC TRAINING (# TRN) AND TESTING (# TST) SAMPLES. THE RIGHTMOST COLUMN DESCRIBES THE VARYING CONDITIONS DURING DATA COLLECTION OF THE ACTIVITIES.

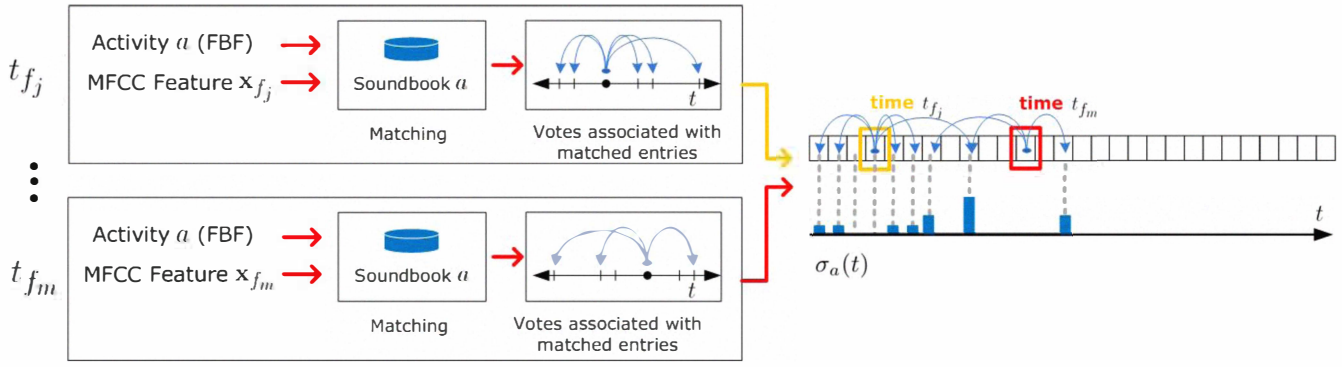

Fig. 3. Non-Markovian ensemble voting recognition (NEV). The figure exemplifies the NEV pipeline for two audio frames at times t_{f_i} and t_{f_m} that vote for the same activity class a. For clarity, this figure illustrates only the voting process, exemplified with a single activity. The output of the per-frame classification stage is a predicted activity label used to select the corresponding soundbooks that are then compared to the MFCC feature descriptor. The votes associated with the matched entries accumulate in the 1D voting space and give the score distribution of activity a.

procedure is in spirit similar to the generation of a visual bag-of-words dictionary [17].

NEV Soundbook generation

This section presents the procedure to learn a *soundbook* for one activity. A training set for a human activity is composed by M training samples $\mathcal{A}_i$, $i = 1, .., M$ of that activity. Training samples are audio files of short duration. The samples are assumed to have the same duration and are subdivided into F number of frames. For each audio sample $\mathcal{A}_i$, we compute F MFCC feature descriptors to form a set of vectors $\mathcal{X}_i = \{\mathbf{x}_i^1, \mathbf{x}_i^2, ..., \mathbf{x}_i^F\}$. Then, for each frame $\mathbf{f}_j$, a temporal displacement Δt_i^j between the frame and the center point in time of the audio sample is computed and associated to $\mathbf{x}_i^j$. This displacement is called *vote*. In order to group features and votes associated to similar sounds, a clustering step is performed. The MFCC feature vectors computed on all the audio samples $\mathcal{X}_1, ..., \mathcal{X}_M$ are clustered in K clusters by using k-means [12]. The cluster centroids $\hat{\mathbf{x}}_1, ..., \hat{\mathbf{x}}_K$ are called *soundbook entries*. Each entry $\hat{\mathbf{x}}_k$ corresponding to the kth cluster is associated to a set of votes $\hat{\mathcal{V}}_k$. Together they make up the soundbook of a given human activity a

$$\mathcal{S}_a = \left\{(\hat{\mathbf{x}}_1, \hat{\mathcal{V}}_1), (\hat{\mathbf{x}}_2, \hat{\mathcal{V}}_2), ..., (\hat{\mathbf{x}}_K, \hat{\mathcal{V}}_K)\right\} \quad (2)$$

Note that each soundbook entry is a generalization of all sounds contained in the cluster. Another important aspect of our approach is how the vote distribution of activity classes is described. A vote set $\hat{\mathcal{V}}_k$ is a sample-based representation not constrained by parametric or predefined distribution models and is thus able to represent arbitrary distributions with varying number of modes. Codebook approaches have been proven to work well in other domains such as visual object recognition [13, 17].

NEV Activity recognition

The working principle of NEV is that human activities are recognized by the score maxima that occurs when many votes vote in a *consistent* manner. This becomes clear by considering the test phase. The NEV recognition phase consists in a per-frame voting process. As soon as a frame is available from the audio stream, its MFCC feature vector $\mathbf{x}$ is computed and the FBF recognition stage predicts its class according to equation (1). The resulting label a is used to select the soundbook $\mathcal{S}_a$ from $\mathcal{S}$. Then, $\mathbf{x}$ is compared with all entries in $\mathcal{S}_a$, $\hat{\mathbf{x}}_1, ..., \hat{\mathbf{x}}_K$ using a L_2 distance criterion: the L nearest neighbors that are distant less than a fixed threshold ϵ are considered valid matches. The set of votes $\hat{\mathcal{V}}$ associated to all matched soundbook entries are used to cast votes forwards and backwards in time. The number of votes is denoted as V. The voting space is the time axis, discretized into a high-resolution histogram for each activity class. Votes are accumulated in the histogram bins using a weighting model that accounts for the degree of ambiguity with which $\mathbf{x}$ was matched. Votes receive a high weight if L is small and a low weight otherwise. This is implemented by an inverse weight function $w = 1/L$. A schematic illustration of the voting procedure is presented in Figure 3.

The score in each histogram bin at time index t, for a considered activity a, is then computed as the accumulation of all V cast and weighted votes from all L matched entries from all audio frames indexed by i

$$\sigma(t) = \sum_{i} \sum_{j=1}^{L_i} \sum_{k=1}^{V_j} \delta_{i,k}(t)\, w_j \tag{3}$$

with $\delta_{i,k}(t)$ is the Kronecker delta that is 1 if $t = t_{\mathbf{f}_i} + \Delta t_k$ and 0 otherwise. $\Delta t_k \in \hat{\mathcal{V}}$ are the votes associated with the matched soundbook entries for frame $\mathbf{f}_i$. Equation 3, if normalized, can be interpreted as the likelihood of detecting an activity of type a

$$\sigma_a(t) \propto p(y = a|\mathbf{x}). \tag{4}$$

Large values of $\sigma_a(t)$ for human activity a represent high confidence that a is carried out. Activities that last for extended periods of time lead to high value plateaus in the voting space, whereas short activities such as door closing produce isolated peaks.

In order to select the winning activity, we perform a non-maxima suppression among all the 1D-voting spaces,

$$a(t) = \underset{a=\{0,1,...,N\}}{\arg\max} \{\, \sigma_a(t) \,\}. \tag{5}$$

By this process NEV smooths the noisy frame-by-frame recognitions by collecting consensus from past and future audio frames. NEV runs in a continuous fashion each time a FBF result becomes available (on-line capability) and it computes estimates that are refined over time. The method's name lends itself from the acausal evidence accumulation from votes that are cast forward and backwards in time.

III. Experiments

Audio data have been collected using a consumer-level dynamic cardioid microphone with integrated wind and pop noise filter. The microphone is mounted on a tripod and pointed towards the source of sound or towards the center of the room. The signal is preamplified by an analog audio mixer and sampled at 44100 Hz via a USB audio interface. All experiments have been conducted using a single microphone in unmodified reverberant real-world environments with several sources of stationary ambient noise (e.g. PC fans whirring).

For computing a MFCC feature descriptor in an audio frame, the signal is loudness-normalized and then used for a Discrete Fourier Transform computation that discards frequencies higher than 8 kHz. Then, a mel-scaled triangular filter bank with 12 filter outputs is used to compute the Inverse Discrete Fourier Transform of the logarithm value of the power spectrum.

Table I lists the 22 human activities considered in this paper. Note that a large number of MFCC training and testing samples (denoted as *#trn* and *#tst*) have been used in the experiments to obtain statistically meaningful results.

The first experiment consists in the comparison of different frame-by-frame activity recognition approaches (FBF). We compare the results of MFCC feature descriptor classification using Random Forests (RF), linear Support Vector Machines (SVM) [21], AdaBoost (AB) [6], and Vector Quantization (VQ) [14]. The SVM classifier has been trained with the stiffness parameter $C = 430$. The AdaBoost classifier has been trained with 50 decision stumps. The Vector Quantization method uses $k = 70$ cluster centers computed with k-means. The Random Forest classifier is trained with 200 decision trees and 50% of the training data (randomly sampled). All parameters have been found through cross-validation.

Frame-by-frame recognition results are shown as the green and red bars in Figure 4. We make use of the f-score metric to ease the ranking between different methods. F-score is a standard classification performance index that considers both precision and recall values. It is defined as $2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$.

Overall, RF achieves higher f-score than all the other methods in 95% of the human activity classes, followed by SVM, VQ, and AB. SVM and VQ obtain similar results. Random Forests and Vector Quantization are both native multi-class classification approaches, they are known to be robust to mislabeling and generalize well in sparse feature spaces. SVM shows good classification capabilities in the complex MFCC feature space thanks to the choice of a linear kernel that avoids overfitting. Adaboost does not perform well because it tends to concentrate many weak classifiers in parts of the space that are difficult to model thereby overfitting locally. In particular, based on the per-frame information, Random Forests classify human activities with a f-score up to 0.95. We have observed that the predominant reason for misclassification is a low signal-to-noise ratio.

The second experiment addressed the third step in our approach and compares the proposed Non-Markovian Ensemble Voting approach (NEV) with a bag-of-sounds (BOS)

approach. BOS is the audio-based counterpart of the bag-of-words method, a widely used method in information retrieval and visual object recognition [17]. Using BOS, a human activity is represented as an unordered collection of soundbook entries whose frequencies are collected in bins of a histogram. The entries are also obtained as centroids from a clustering method (here: k-means). Histograms of different activity classes, represented as points in the space of soundbook entries, are finally predicted following a one-vs-all linear SVM strategy.

NEV and BOS have been trained with 50 entries in the soundbook for each human activity. By analyzing the f-scores, NEV is more accurate than BOS. The experiment shows the contribution from the voting process that yields an ordering of soundbook matches. BOS disregards the ordering and might further suffer from the fact that a linear separation of the class histograms is too approximative. Note also that BOS has no on-line property. There is no way to extract the start and end of an activity – an information that is readily available as score changes in the NEV approach. Therefore, BOS requires a segmentation of the audio stream to classify the segments. This in turn renders the recognition of actions shorter than a segment a very difficult task. Summarizing, NEV classifies human activities with an average f-score of 0.92, average precision 94% and average recall 91%, see Figure 4-top. NEV clearly outperforms all the other human activity recognition methods presented in this paper.

A detailed f-score comparison for all the techniques is shown in Figure 5, rightmost columns. The NEV approach is largely more accurate than every frame-by-frame classification method (RF,AB,SVM,VQ). NEV is also more accurate than BOS in classifying 77% of the human activities often with large performance margin. In all the other cases, NEV is only marginally outperformed by BOS albeit not requiring any audio stream segmentation.

In the third experiment, we evaluate the system's ability to recognize human activities in a continuous fashion. A user is asked to perform unscripted kitchen-related activities as it would happen in a human-robot interaction scenario. The experiment includes both silence between activities and activities carried on for undefined periods of time. No new training is performed for this experiment. The ground truth labels have been added manually, see Figure 6, top. NEV achieves a very high recognition rate of 85.8% correctly predicted audio frames in the experiment (Figure. 6, bottom) demonstrating the ability of NEV to perform well under realistic conditions.

IV. Conclusions

This paper addresses the problem of audio-based human activity recognition. We propose a novel segmentation-free approach called Non-Markovian Ensemble Voting (NEV), able to robustly classify human activities in an on-line and any-time fashion. The method, inspired by visual codebook approaches, relies on learned soundbooks of activity classes, recognized by score maxima that emerge when votes from short-duration audio frames are cast in a consistent way with respect to the learned model. NEV does not rely on audio stream segmentation and can deal with variable-length activities. We performed three experiments with a set of 22 human activities from a bathroom and kitchen context. NEV outperforms several alternative classification methods (Support Vector Machines, Vector Quantization, AdaBoost, and Bag-of-words) and leads to high recognition rates of more than 85% in a final experiment on continuous activity recognition.

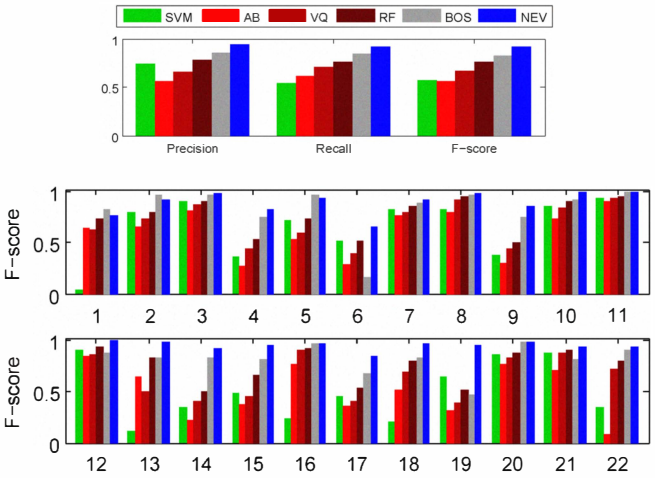

Fig. 4. Overall comparison of the classification methods for the 22 human activity classes. Colored bins depict different recognition methods: per-frame recognition methods (FBF) are shown in shades of red and green, BOS in gray, and the proposed method NEV in blue. **Top**: Overall comparison of average precision, recall and f-score. NEV clearly outperforms all the other recognition methods. **Bottom**: Detailed analysis of the f-score for the 22 classes. NEV outperforms all the other methods or achieves very competitive f-scores.

Future work will analyze how robot-typical signal-level factors such as non-stationary ego-noise from moving joints impact these recognition rates.

Acknowledgment

This work has partly been supported by the German Research Foundation (DFG) under contract number SFB/TR 8.

References

[1] J. Breebaart and M. Mckinney, "Features for audio classification," in *Proc. Philips Symp. of Intel. Algorithms*, 2002.

[2] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.

[3] J. Chen, A. H. Kam, J. Zhang, N. Liu, and L. Shue, "Bathroom activity monitoring based on sound," in *Proc. of the Int. Conf. on Pervasive Computing*, 2005.

[4] B. Clarkson and A. Pentland, "Extracting context from environmental audio," in *Proc. of the IEEE Int. Symposium on Wearable Computers*, 1998.

[5] A. J. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 321–329, 2006.

[6] Y. Freund and R. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," in *Computational Learning Theory (Eurocolt)*, 1995.

[7] G. Guo and S. Li, "Content-based audio classification and retrieval by support vector machines," *IEEE Transactions on Neural Networks*, vol. 14, no. 1, pp. 209 – 215, 2003.

[8] G. Guo, H.-J. Zhang, and S. Z. Li, "Boosting for content-based audio classification and retrieval: An evaluation," in *Proc. of the IEEE Conf. on Multimedia and Expo (ICME)*, 2001.

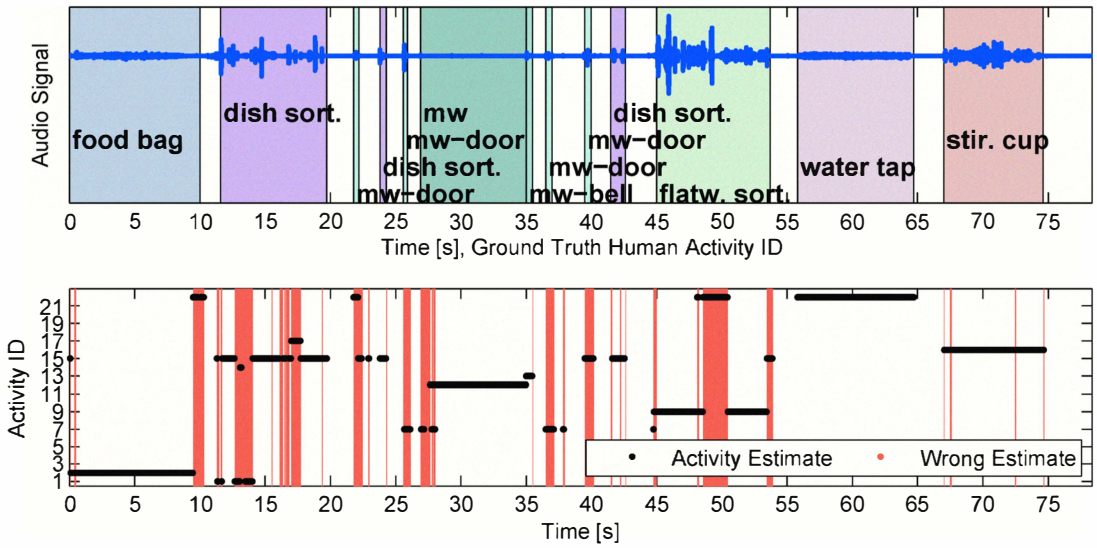

Fig. 6. NEV result for continuous recognition under real-world conditions: a user is asked to perform unscripted kitchen-related activities. **Top**: The recorded audio signal is shown in blue, the activity ground truth with colored bars. **Bottom**: Activities recognized by the NEV approach: black dots indicates the estimated activity ID, white areas indicate correct classification. In this example, our system achieves 85.8% recognition rate.

ID	LSVM	AB.	VQ	RF	BOS	NEV
1	0.0441	0.6393	0.6286	**0.7369**	**0.8367**	0.7667
2	0.7939	0.6580	0.7364	**0.8071**	**0.9629**	0.9261
3	0.9031	0.8198	0.8711	**0.9054**	0.9716	**0.9923**
4	0.3673	0.2659	0.4387	**0.5413**	0.7535	**0.8288**
5	0.7292	0.5429	0.5988	**0.7425**	**0.9655**	0.9324
6	0.5172	0.2843	0.4030	**0.5241**	0.1667	**0.6662**
7	0.8237	0.7662	0.7928	**0.8688**	0.8952	**0.9196**
8	0.8335	0.8021	0.9220	**0.9491**	0.9707	**0.9921**
9	0.3749	0.3005	0.4442	**0.5119**	0.7469	**0.8663**
10	0.8563	0.7374	0.8504	**0.9015**	0.9232	**0.9955**
11	0.9462	0.9145	0.9386	**0.9556**	**1.0000**	0.9992
12	0.9152	0.8432	0.8639	**0.9419**	0.8842	**0.9952**
13	0.1211	0.6386	0.4975	**0.8320**	0.8319	**0.9802**
14	0.3534	0.2280	0.4145	**0.5090**	0.8279	**0.9183**
15	0.4817	0.3772	0.4615	**0.6638**	0.8090	**0.9590**
16	0.2345	0.7660	0.9067	**0.9281**	0.9643	**0.9711**
17	0.4572	0.3652	0.4046	**0.5402**	0.6761	**0.8435**
18	0.2127	0.5165	0.6866	**0.7922**	0.8239	**0.9648**
19	**0.6425**	0.3257	0.3977	0.5210	0.4701	**0.9480**
20	0.8581	0.7728	0.8373	**0.8726**	**0.9846**	0.9842
21	0.8763	0.7014	0.8780	**0.9113**	0.8126	**0.9334**
22	0.3469	0.0804	0.7268	**0.7988**	0.9124	**0.9369**

Fig. 5. Comparison of F-scores for different audio-based human activity recognition methods. The first four columns (LSVM, AB, VQ, RF) are techniques applied for FBF classification. The last two columns (BOS, NEV) are techniques that take into account the information collected in multiple audio frames. The method proposed in this paper (NEV) outperforms all FBF techniques. NEV is also more accurate than BOS in 77% of the human activities and very competitive in all the other cases albeit not requiring any audio stream segmentation.

[9] A. Härmä, M. McKinney, and J. Skowronek, "Automatic surveillance of the acoustic activity in our living environment," in *Proc. of the IEEE Conf. on Multimedia and Expo (ICME)*, 2005.

[10] M. R. Hasan, M. Jamil, M. G. Rabbani, and M. S. Rahman, "Speaker indentification using MEL frequency cepstral coefficients," in *3rd Int. Conf. on Electrical & Computer Engineering ICECE*, 2004.

[11] G. Ince, K. Nakadai, T. Rodemann, Y. Hasegawa, H. Tsujino, and J. Imura, "A hybrid framework for ego noise cancellation of a robot," in *Proc. of the Int. Conf. on Robotics & Automation (ICRA)*, 2010.

[12] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," *IEEE Trans. on Pattern Analysis & Machine Intell.*, vol. 24, pp. 881–892, 2002.

[13] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," in *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*, 2005.

[14] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Trans. on Comm.*, vol. 28, no. 1, pp. 84–95, 1980.

[15] P. Lukowicz, J. A. Ward, H. Junker, M. Stäger, G. Tröster, A. Atrash, and T. Starner, "Recognizing workshop activity using body worn microphones and accelerometers," in *Pervasive Computing*, 2004, pp. 18–32.

[16] K. Nakadai, D. Matsuura, H. G. Okuno, and H. Kitano, "Applying scattering theory to robot audition system: Robust sound source localization and extraction," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2003.

[17] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," in *Eur. Conf. on Comp. Vis. (ECCV)*, 2006.

[18] V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi, and T. Sorsa, "Computational auditory scene recognition," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 2002.

[19] R. Takeda, K. Nakadai, T. Takahashi, K. Komatani, T. Ogata, and H. G. Okuno, "ICA-based efficient blind dereverberation and echo cancellation method for barge-in-able robot audition," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2009.

[20] J.-M. Valin, J. Rouat, and F. Michaud, "Enhanced robot audition based on microphone array source separation with post-filter," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2004.

[21] V. N. Vapnik, *The nature of statistical learning theory.* Springer-Verlag, 1995.

[22] A. Wang, "An industrial strength audio search algorithm," in *Int. Conf. on Music Information Retrieval*, 2003.

[23] J. Zhang, J. L. Whalley, and S. Brooks, "Time mosaics - An image processing approach to audio visualization," in *11th Int. Conf. on Digital Audio Effects*, 2008.

[24] T. Zhang and C.-C. Jay Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 4, pp. 441–457, 2001.

[25] Y. Zhu, Z. Ming, and Q. Huang, "Automatic audio genre classification based on support vector machine," in *Proc. of the Conf. on Neural Information Processing Systems (NIPS)*, 2007.