Audio-Based Human Activity Recognition with Robots

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

Social Robotics Lab, University of Freiburg, Germany
{stork,silvaj,spinello,arras}@informatik.uni-freiburg.de

Abstract. Human activity recognition is a key skill for socially enabled robots to effectively and naturally interact with humans. In this paper we exploit the fact that most human activities produce very characteristic sounds from which a robot can infer the corresponding actions. We propose a novel classification approach called Non-Markovian Ensemble Voting (NEV) that is able to recognize human activities in an on-line fashion, does not rely on any audio stream segmentation and can deal with variable-length activities. In the experiments, we are able to robustly recognize 22 human activities in a bathroom and kitchen context.

1 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. Audio appears to be a well suited sensory modality for this task since many human activities produce very characteristic sounds from which a robot can infer the corresponding human actions.

Audio-based human activity recognition has been addressed mainly in the wearable computing community, e.g. for auditory surveillance and multimedia systems [7, 2, 3, 5]. In robotics, the field of robot audition is typically concerned with problems at signal-level such as sound source localization [8], source separation [10], or ego-noise compensation [6]. This paper advances the state of the art in robot audition by addressing the high-level problem of recognizing human activities. In particular, we propose Non-Markovian Ensemble Voting (NEV), a novel on-line and any-time classification technique for sequential data. Unlike our method, previous works rely on audio stream segmentation to recognize human activities [5, 7] usually achieved by silence detection, detection of abrupt feature changes or even manual annotation. Methods that do not rely on segmentation either perform batch processing [3, 11] or classify only short-duration events [4], unlikely to be robust in a robotics context.

2 Audio-based human activity recognition

Non-Markovian Ensemble Voting consists in the three steps explained hereafter.

Feature extraction from raw audio data: The audio stream is subdivided into short-duration frames $\mathbf{f}_i$. A frame spans $40\,ms$ of audio data and overlaps by 87.5% with neighboring frames to ensure a high level of data correlation. For each $\mathbf{f}_i$ at time index $t_{\mathbf{f}_i}$, an MFCC feature descriptor $\mathbf{x}_i$ is computed. MFCC features are widely used for speech and audio processing.

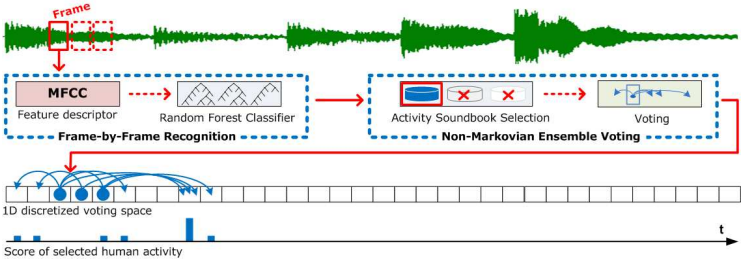

Fig. 1. Non-Markovian Ensemble Voting for human activity recognition.

Frame-by-frame recognition (FBF): Each MFCC feature descriptor $\mathbf{x}_i$ is classified using a Random Forest (RF) [1] classifier $h(\mathbf{x}_i)$ to estimate the activity label y_i. Random Forests are a supervised ensemble classification method that makes use of multiple randomized decision trees to subdivide the feature space. FBF runs in a continuous fashion and computes an estimated activity label $h(\mathbf{x_i}) = y_i$ for each frame $\mathbf{f}_i$ from the set of N classes $y_i \in \{0, 1, ..., N\}$. However, classification is still unreliable at this level due to the natural variability in human actions and signal-level factors such as noise or low SNR.

Non-Markovian Ensemble Voting (NEV): The training procedure of NEV consists in building a *soundbook* for each human activity. This procedure is in spirit similar to the generation of a visual bag-of-words dictionary [9]. A training set for a human activity a is composed by M audio samples $\mathcal{A}_i$, $i = 1, .., M$ of that activity. The samples are equal in length and thus have the same number of frames F. MFCC features are computed for each frame of each audio sample to form the set of vectors in feature space $\mathcal{X}_i = \left\{\mathbf{x}_i^1, \mathbf{x}_i^2, ..., \mathbf{x}_i^F\right\}$ for all i. For each frame $\mathbf{f}_j$ in each audio sample $\mathcal{A}_i$, a temporal displacement Δt_i^j between the frame and the center point in time of the audio sample is associated to $\mathbf{x}_i^j$. This displacement is called *vote*. In order to generalize features and votes from the same activities, a clustering step is performed. All features for all audio samples $\mathcal{X}_1, ..., \mathcal{X}_M$ are clustered using k-means. The cluster centroids $\hat{\mathbf{x}}_1, ..., \hat{\mathbf{x}}_K$ are called *soundbook entries*. To each entry, a set of vote centroids $\hat{\mathcal{V}}_k$ is associated. Together they make up the soundbook of a given human activity a, $\mathcal{S}_a = \{(\hat{\mathbf{x}}_1, \hat{\mathcal{V}}_1), (\hat{\mathbf{x}}_2, \hat{\mathcal{V}}_2), ..., (\hat{\mathbf{x}}_K, \hat{\mathcal{V}}_K)\}$. Finally, the learned model $\mathcal{S}$ is a set of soundbooks $\mathcal{S} = \{\mathcal{S}_{a_1}, \mathcal{S}_{a_2}, ..., \mathcal{S}_{a_N}\}$ for the activity classes of interest.

The NEV recognition phase consists in a per-frame voting process (Fig. 1). For each incoming frame $\mathbf{f}_i$, the MFCC features $\mathbf{x}$ are computed and classified in the FBF recognition stage. The label y is used to select the activity soundbook $\mathcal{S}_a$ from $\mathcal{S}$. Then, $\mathbf{x}$ is matched with all soundbook entries $\hat{\mathbf{x}}_k$ using a L_2 distance criterion and a k-NN strategy within a ball of radius ϵ. The number of matched entries is denoted as L. The associated votes $\hat{\mathcal{V}}$ of all matched entries are used to cast a set of votes forwards and backwards in time. The voting space is the time axis, discretized into a high-resolution histogram for each class.

The score in each histogram bin is then computed as the accumulation of all V votes from all L matched entries from all frames from the incoming audio stream indexed by i, $\sigma(t) = \sum_i \sum_{j=1}^{L_i} \sum_{k=1}^{V_j} \delta_{i,k}(t)\, w_j$ with $\delta_{i,k}(t)$ being a selector variable that is 1 if $t = t_{\mathbf{f}_i} + \Delta t_k$ and 0 otherwise and $w = 1/L$ being a weight that reflects the ambiguity with which $\mathbf{x}$ was matched. $\sigma(t)$, if normalized, can

ID: Activity	ID: Activity	ID: Activity
1: No human activity	9: Sorting flatwares	17: Flushing a toilet
2: Opening a food bag	10: Using a food processor	18: Brushing teeth
3: Mixing with a blender	11: Using a hairdryer	19: Using a vacuum cleaner
4: Pouring cereals into bowl	12: Microwave (mw) cooking	20: Using a washing machine
5: Eating cereals	13: Switching off a mw oven	21: Boiling water
6: Pouring water into a cup	14: Closing a mw oven door	22: Using the water tap
7: Using a dishwasher	15: Sorting dishes	
8: Shaving w. electric razor	16: Stirring water in a cup	

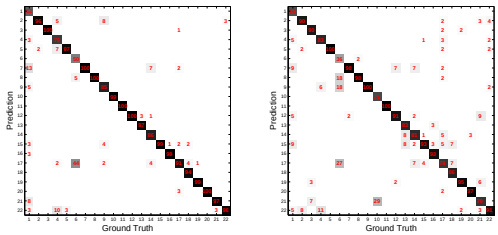

Fig. 2. The activities considered in the experiments (**top**) and the resulting confusion matrices for the NEV approach (**bottom left**) and the BOS approach (**bottom right**).

be interpreted as the likelihood of detecting an activity, $\sigma_a(t) \propto p(y = a|\mathbf{x})$. Long-duration activities lead to high value plateaus in the voting space, short activities such as door closing produce isolated peaks. To select the winning activity among the N voting spaces, we perform a non-maxima suppression across all classes. The method's name lends itself from the acausal evidence accumulation from votes that are cast forward and backwards in time.

3 Experiments

In the experiments we consider 22 classes of human activities listed in Fig. 2. Data have been collected using a consumer-level dynamic cardioid microphone mounted on a mobile robot in reverberant real-world environments with several sources of ambient noise. We assume that human activities occur sequentially, one at a time, and keep the robot still to avoid non-stationary ego-noise.

In the first experiment we compare NEV with a bag-of-sounds (BOS) approach. BOS is the audio-based counterpart of the bag-of-words method, widely used in information retrieval and object recognition [9]. Using BOS, a human activity is represented as an unordered collection of soundbook entries whose occurrences are accumulated in bins of a histogram. The entries are also obtained as centroids from k-means clustering. Being points in the space of soundbook entries, histograms are recognized following a one-vs-all linear SVM strategy.

NEV and BOS have been trained with $k = 50$ soundbook entries. The resulting confusion matrices are shown in Fig. 2. By ranking the matrices, NEV is more accurate than BOS. The experiment shows the contribution of the voting process that yields an ordering of soundbook matches. BOS disregards this ordering. Note also that BOS has no any-time 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. Summarizing, NEV classifies 91% of the 22 human activities with a high detection rate of 76% to 100%.

In the second experiment, we evaluate the system's ability to recognize human activities continuously. A user is asked to perform unscripted kitchen-related

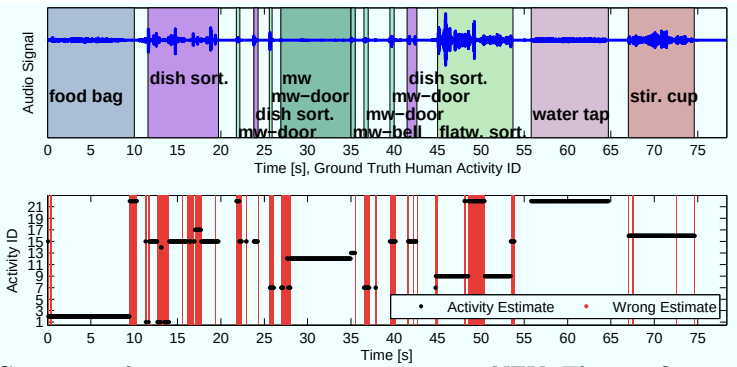

Fig. 3. Continuous human activity recognition using NEV. The top figure shows the (manually added) ground truth labels, the bottom figure shows the recognition result: black dots indicates the estimated activity ID, white areas indicate correct classification. Our system achieves 85.8% correct recognition rate of the user's activities.

activities. No new training is performed. NEV achieves a very high accuracy of 85.8% correctly predicted audio frames demonstrating the ability of the approach to perform well under realistic conditions (Fig. 3).

Acknowledgment This work has been supported by DFG contract num. SFB/TR-8.

References

1. Breiman, L.: Random forests. Mach. Learn. 45, 5–32 (2001)
2. Chen, J., Kam, A.H., Zhang, J., Liu, N., Shue, L.: Bathroom activity monitoring based on sound. In: Pervasive (2005)
3. Eronen, A.J., Peltonen, V.T., Tuomi, J.T., Klapuri, A.P., Fagerlund, S., Sorsa, T., Lorho, G., Huopaniemi, J.: Audio-based context recognition. IEEE Transactions on Audio, Speech, and Language Processing 14(1), 321–329 (2006)
4. Guo, G., Zhang, H.J., Li, S.Z.: Boosting for content-based audio classification and retrieval: An evaluation. In: IEEE Conf. on Multimedia and Expo (2001)
5. Härmä, A., McKinney, M., Skowronek, J.: Automatic surveillance of the acoustic activity in our living environment. In: IEEE Conf. on Multimedia and Expo (2005)
6. Ince, G., Nakadai, K., Rodemann, T., Hasegawa, Y., Tsujino, H., Imura, J.: A hybrid framework for ego noise cancellation of a robot. In: Int. Conf. on Robotics & Automation (2010)
7. Lukowicz, P., Ward, J.A., Junker, H., Stäger, M., Tröster, G., Atrash, A., Starner, T.: Recognizing workshop activity using body worn microphones and accelerometers. In: Pervasive Computing, Lecture Notes in Comp. Sci., vol. 3001 (2004)
8. Nakadai, K., Matsuura, D., Okuno, H.G., Kitano, H.: Applying scattering theory to robot audition system: Robust sound source localization and extraction. In: IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (2003)
9. Nowak, E., Jurie, F., Triggs, B.: Sampling strategies for bag-of-features image classification. In: Eur. Conf. on Comp. Vis. (2006)
10. Valin, J.M., Rouat, J., Michaud, F.: Enhanced robot audition based on microphone array source separation with post-filter. In: IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (2004)
11. Zhu, Y., Ming, Z., Huang, Q.: Automatic audio genre classification based on support vector machine. In: 3rd Int. Conf. on Natural Comp. (2007)