Exploiting Acoustic Source Localization for Context Classification in Smart Environments

Christian Kleine-Cosack, Marius H. Hennecke,
Szilárd Vajda, and Gernot A. Fink

Robotic Research Institute, TU Dortmund University,
Dortmund, Germany

Abstract. Smart environments rely on context classification in order to be able to support users in their daily lives. Therefore, measurements provided by sensors distributed throughout the environment are analyzed. A main drawback of the solutions proposed so far is that the type of sensors and their placement often needs to be specifically adjusted to the problem addressed. Instead, we propose to perform context classification based on the analysis of acoustic events, which can be observed using arrays of microphones. Consequently, the sensor setup can be kept rather general and a wide range of contexts can be discriminated. In an experimental evaluation within a smart conference room we demonstrate the advantages of our new approach.

Keywords: Context classification, audio localization, smart rooms

1 Introduction

With the advent of pervasive computing technologies, so-called smart spaces have been introduced to support humans during their activities of daily living (ADL). Varieties of sensors are embedded into the people's homes and working environments constantly monitoring activities and surrounding conditions. To allow for practical application, the sensor setups must be chosen with care, respecting the specific requirements of the scenario. In the domestic domain, e.g., privacy issues are one key concern, while the flexibility of the sensor setup is another for smart spaces which are used in a wide range of application. Typical use cases for the latter are smart conference rooms and working environments, which support the human users during meetings and presentations or assist in collaborative working tasks (e.g. [16]).

A fundamental prerequisite for a number of applications related to smart spaces is the precise knowledge of the environment's context. In a smart conference room, e.g., intelligent controlling of light sources depends on knowledge about the current situation. While the luminosity must be kept below a certain level in the projection area during presentation, the opposite is desired for discussions and collaborative tasks. On a more abstract level, an important prerequisite for pro-active behavior of the smart environment is the knowledge about the current situation.

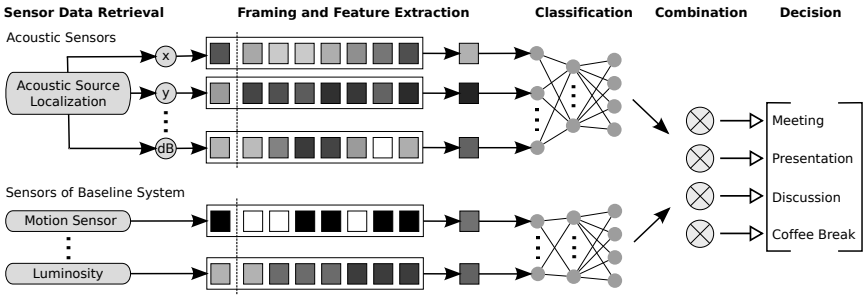


Fig. 1: System overview.

In this paper we propose a real-time context classification system based on acoustic source localization (ASL). As smart conference rooms are usually equipped with multiple microphones, acoustic events within the environment can be constantly monitored. The configuration of the microphone setup is chosen w.r.t. the characteristics of this sensor type only, not considering the specific experimental setting. This generic setup allows the classification of a multitude of situations, without the need to adopt the sensors' configuration to the latter. Hence, it provides high flexibility in practical applications. For reasons of comparison and to further improve the performance, a baseline system utilizing environmental sensors (i.e. motion, air condition, luminosity and door status sensors) is used and integrated into the classification process.

For classification, the signal energy as well as the spatial position of acoustic events are calculated for each time frame and statistical features are derived, i.e. mean and variance. Using a sliding window approach the latter are accumulated over time, and the resulting vector is fed into a Multi-Layer Perceptron (MLP) based classification system. An overview of the system structure is given in Fig. 1 illustrating the ASL based classification, the baseline system, and the integration of the two. Direct measurements of the environmental sensors are used in the feature extraction step whereas an acoustic source localization is used beforehand for the acoustic sensors. The sensor types are treated separately in means of classification and a final combination step results in a context decision.

The effectiveness of the given approach is demonstrated in an experimental evaluation in a smart conference room. In the given scenario the results show improved classification rates compared to the baseline system. Moreover, we show that the two can be combined to improve the overall performance. The proposed system is the first that allows context classification in smart environments based on the acoustic analysis of human interaction patterns. The sensor configuration is completely independent of the situation to classify in a specific scenario, hence, it provides high flexibility in practical applications. Moreover, only statistical data is used for the classification of context. In consequence, the presented approach can also be applied to scenarios in which privacy requirements must be met. The presented system is capable of real-time context classification, a key requirement in practical applications.

2 Related Work

A large volume of literature on situation analysis in smart environments exists, which can be split into activity recognition (AR) and context classification. While the former focuses on the analysis of ADL, the intention of the latter and the proposed approach is to recognize the overall context of the environment, i.e. the current, long-term situation of a smart room.

2.1 Situation Analysis in Smart Environments

The literature covers AR either based on computer vision [15] or by applying pervasive computing approaches [1]. Whereas AR systems are usually based on statistical modeling frameworks, such approaches can rarely be found in case of context classification.

The evaluation of ontologies represents the methodology of choice instead. An ontology represents a well-defined, expert-made set of rules, which describes situations in certain environments by means, e.g., of (fixed) combinations of sensor events. For the conceptual and technical implementation of ontologies different semantics and toolkits are used [6,14]. Also, semi-automatically constructed ontologies are described that allow to split the context model into a handcrafted and an automatically derived part [12]. Whereas the first part covers more general, domain-related aspects, the latter focuses on specialties of the particular application. Still, prior expert knowledge remains an integral part of these rule based approaches. In consequence, ontologies provide a methodology for context classification in scenarios which can be described by a set of simple rules. Such prerequisite can be met in case of sensor setups tailored towards a specific situation, in which a predefined combination or sequence of sensor events determines the current situation. In case of continuous, rather noisy and widespread sensor events; however, the application of ontologies is limited and more likely to fail.

Situational context classification using statistical models and common sensors is hardly covered by related work. Using binary motion sensors in [18] topological sensor networks are evaluated for this task. Focusing on energy saving in [5] multi-modal environmental data (monitoring air quality, water consumption, motion etc.) are analyzed using a hidden Markov model (HMM). In [8] measurements from custom-made binary state change sensors are analyzed using an HMM and Conditional Random Fields. A similar sensor network is considered in [17] for tracking and activity recognition using a particle filter. However, all these approaches rely on a multitude of sensors which must be installed and configured precisely to cover possible human activities. Hence, expert knowledge and a deep understanding of the users' behavior is mandatory. In contrast, acoustic sensors can be installed without such specific knowledge, enhancing flexibility for practical applications.

2.2 Acoustic Source Localization

The problem of localizing the spatial position of an acoustic event is known as acoustic source localization (ASL). Several techniques for solving this task

using multiple acoustic sensors, i.e. microphones, are known. The main challenges for an ASL system in a real world environment are noise and reverberation. Possible noise sources are fans of electrical equipment, foot fall sounds or any other putative unwanted sound events. Robustness against these factors are of utmost importance, in order to obtain reliable results from an ASL system.

Time delay estimation based algorithms [3] first determine the delay of a signal between two or more sensors and combine those time-difference of arrivals (TDOA) in a second stage to form a source location. Not taking an explicit room reverberation model into account, the generalized cross-correlation (GCC) [10] provides a simple and efficient approach. Unlike the GCC method adaptive eigenvalue decomposition (AED) [2] models reverberation explicitly, by blindly estimating the acoustic channels impulse responses. Its higher computational costs are justified by a more robust TDOA estimation in reverberant environments. Each TDOA restricts the acoustic source to lie on one sheet of a two-sheeted hyperboloid and hence the second stage needs to find the intersection of all TDOA parameterized hyperboloids (e.g. [7]).

Steered response power (SRP) based ASL schemes use beamforming techniques to steer a passive sensor array to different locations. The position with the highest beamformer output power is then assumed to be the source location. Hence, SRP algorithms avoid the two stage approach of the TDOA methods. Calculating the SRP for all possible locations requires a high computational effort. Nevertheless, DiBiase et al. [4] showed that the SRP of a simple delay-and-sum beamformer is equivalent to a spatial combination of all pair-wise GCCs, which leads to a computationally tractable ASL algorithm.

3 Context Classification using ASL

Today, there is a trend to equip smart conference rooms with multiple microphones for teleconferencing applications or meeting recordings, which facilitates the extraction of positions of interacting human speakers. In the near future, distributed ad-hoc microphone arrays, e.g. the combination of mobile phones and other embedded devices available, might fulfill this task. In the following, we present our system for context classification in a smart environment which is based on ASL information.

3.1 Acoustic Source Localization

Steered response power based ASL methods have shown good results under moderate noise and reverberation levels. Due to this robustness and their additional simplicity, we employ such an SRP scheme—namely the SRP-PHAT [4]—and use the positions of the localized acoustic sources for context classification. Here, we will shortly revisit SRP-PHAT.

The TDOA of a source signal between two or more receivers is the basic building block of SRP-PHAT. A common approach for estimating a TDOA $\hat{\tau}_{ij}$

for the two-channel case is the GCC [10]

$$R_{ij}(\tau) = \mathcal{F}^{-1}\{\Psi_{ij}(\omega)X_i(\omega)X_j^*(\omega)\}(\tau) \ . \tag{1}$$

It is the inverse Fourier transformation $\mathcal{F}^{-1}\{\cdot\}(t)$ of the cross-spectrum of the acoustic channel i and j, where $x_i(t) = \mathcal{F}^{-1}\{X_i(\omega)\}(t)$ is the i-th channel's signal. In order to shape the GCC for better TDOA estimation performance the phase transformation (PHAT) [10] $\Psi_{ij}(\omega) = |X_i(\omega)X_j^*(\omega)|^{-1}$ is commonly used. It is motivated by the fact that a pure time delay results in a phase shift and leaves the signals amplitude unchanged. Hence, it is a simple whitening of the cross-spectrum. Despite its solely heuristic nature, PHAT as a pre-filter has shown robustness under moderate reverberation and noise conditions. For simplicity, we omit some technical details, e.g. a block-processing scheme.

The robustness of SRP-PHAT stems from the spatial combination

$$P(\boldsymbol{q}) = \sum_{(i,j)\in\mathcal{P}} R_{ij}(\tau_{ij}(\boldsymbol{q})) \ , \qquad \tau_{ij}(\boldsymbol{q}) = \frac{|\boldsymbol{q}-\boldsymbol{p}_i| - |\boldsymbol{q}-\boldsymbol{p}_j|}{c} \ , \tag{2}$$

of possibly redundant GCCs for pairs $(i,j) \in \mathcal{P} \subseteq \{1,\ldots,M\}^2$ using a total of M microphones. Given a spatial position $\boldsymbol{q}$ and the microphone positions $\boldsymbol{p}_i, \boldsymbol{p}_j$ the TDOA leading to $\boldsymbol{q}$ is calculated via $\tau_{ij}(\boldsymbol{q})$, where $c = 343\,\mathrm{m\,s^{-1}}$ is the speed of sound and assumed constant. Equation (2) defines a pseudo spatial likelihood function (SLF) for a source position $\boldsymbol{q}$. Hence, an estimate for the location of the dominant acoustic source is given as $\hat{\boldsymbol{q}} = \arg\max_{\boldsymbol{q}} P(\boldsymbol{q})$.

3.2 Feature Set

The aforementioned ASL system operates on frames of 300 ms. With a sampling rate of $f_s = 48\,\mathrm{kHz}$ and a frame shift of 150 ms the system estimates up to seven source positions per second. Each estimate includes the three-dimensional location $\hat{\boldsymbol{q}} = (x,y,z)^{\mathrm{T}}$, the SLF value $P(\hat{\boldsymbol{q}})$ and the speech energy of the current signal frame in dB averaged over all acoustic channels. For those five raw ASL measurements the mean and variance over two second windows (14 raw ASL samples) are calculated, leading to ten ASL features in total. The window size is chosen, such that a fair amount of context-relevant speech portions are considered.

Given those ASL features a sliding window procedure is applied, extracting frames of fixed lengths l with overlap $o = 10\,\mathrm{s}$. Elements of the frame are accumulated, resulting in a ten-dimensional vector which, after mean subtraction, is normalized element-wise and fed into the classification system.

3.3 Baseline System

In order to compare our approach for context classification based on ASL, we use a baseline system which utilizes only environmental, non-intrusive sensors [9]. Monitoring the temperature, CO_2 content and humidity of the air, luminosity,

the status of doors and windows, as well as motion in certain regions of the environment, this system has already proven its capabilities for the given task. Moreover, this baseline system is integrated in the ASL based context classification approach to further improve the results.

3.4 Neural Network Classifier

In order to classify the smart environment's context based on audio source data (ASL approach) and sensory information (baseline system), a neural network has been built. Instead of using rule based ontologies which involves human expert knowledge, our system is based on statistical learning and can adapt to arbitrary data. For our experiments a fully connected multi-layer perceptron (MLP) with sigmoid transfer function as activation has been used [11]. For training purposes an error back-propagation mechanism has been employed to minimize a square-error metric.

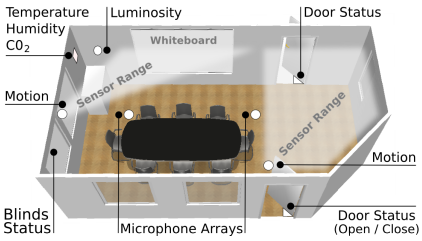
The network topologies are directly derived from the data. The number of processing units in the input layer corresponds to the dimensionality of the data, while the output nodes are assigned to the different context scenarios to be classified. The neurons in the hidden layer can not be estimated automatically, hence some trial runs have been conducted in that sense. Considering the audio data, a 10-10-4, while for the environmental sensor data a 15-10-4 topology is applied. In order to integrate the two classifiers, first the activations of the output neurons of each network are normalized, i.e. estimates of class posterior probabilities are obtained. Assuming class independence between the different sensor information, combined estimates can be calculated by simple multiplication (right part of Fig. 1).

4 Evaluation

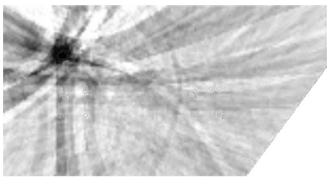
In order to demonstrate the effectiveness of the presented approach we conducted a number of practical experiments in a smart conference room. The objective of the experiments was to evaluate the given approach in a real world setting during regular usage of this environment.

4.1 Setting

The work described in this paper is part of a greater research project, the FINCA [13]. This intelligent environment, amongst others, consists of a conference room equipped with a multitude of ambient sensors, microphones and cameras. Figure 2a gives an overview of the sensor setup in general. Sixteen microphones grouped in two distributed ceiling-mounted arrays are employed for ASL. A visualization of the SLF (2) for an arbitrary time instance as used for the ASL approach is shown in Fig. 2b. The location of the audio source is approximately the center of the high energy area (dark spot in the two-dimensional map).



(a) Configuration of acoustic and ambient sensors.



(b) SLF $P(\boldsymbol{q})$ evaluated over the two-dimensional FINCA grid.

Fig. 2: The smart conference room of the FINCA.

4.2 Data Set and Methodology

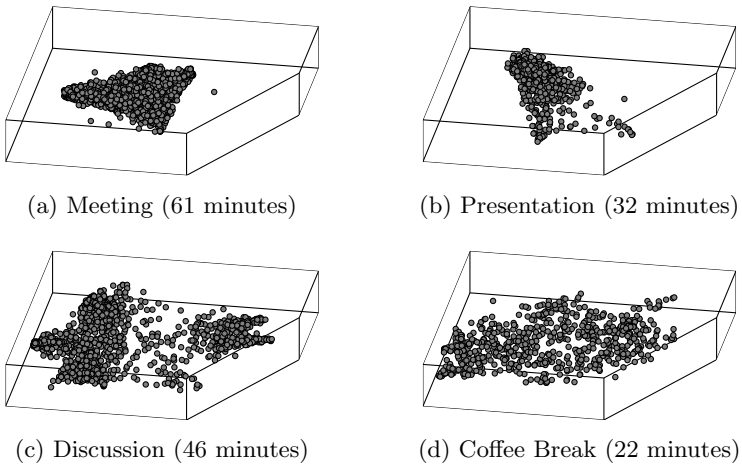
Sensor data was recorded over a time period of several months with multiple recording sessions for each situation. The final data set consists of four situations that are typically related to meeting rooms.

Meeting: Team meeting, using possibly the white-board for collaborative tasks (122 minutes of data)
Presentation: A talk given by one person while slides are projected on the white-board. (107 minutes of data)
Discussion: Questions from the audience after a given presentation. (67 minutes of data)
Coffee break: Coffee break between subsequent presentations and meetings. (45 minutes of data)

Regarding the acoustic sensor information, separating these classes is a challenging task as audio events occur rather widespread throughout the different situations. Figure 3 visualizes the localization results for exemplarily selected, complete recording sessions. Note that for the classification only short time periods of such sensor data, i.e. several seconds, are used. Moreover, not the raw coordinates as plotted in Fig. 3, but the more general statistics are utilized. Hence, not the positions of speakers are learned by the classification system, but the patterns of interaction between multiple persons.

A five-fold cross-validation was performed for training, parameter optimization and final testing. The dataset was subdivided into five disjoint sets. Training was performed using three fifths of the sets, validation of the MLP on the fourth, and final testing on the remaining set. Final results were obtained by averaging over the results achieved for every such configuration. To account for practical applicability, no effort was taken to manually adjust the data set. For training, classes were automatically balanced by data duplication, whereas the validation and test procedure was run on the unmodified and unbalanced set.

The experiments were designed to analyze the classification performance for ASL based context classification, the baseline system and the combined approach



(a) Meeting (61 minutes) (b) Presentation (32 minutes)

(c) Discussion (46 minutes) (d) Coffee Break (22 minutes)

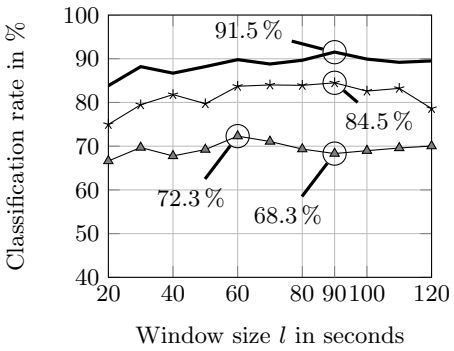
Fig. 3: ASL results for different exemplarily selected audio recordings.

for different window sizes. Moreover, a reduced feature set was tested to investigate the ability of the system to learn interaction patterns apart from any spatial position. As stated above, only derived statistics are used by the classifier, not the raw speaker locations. However, the mean value of x,y and z-coordinates over all audio events within a specific frame – i.e. the center of audio activity in the environment at the given time – is regarded. In contrast, the reduced feature set discards this location specific information.

4.3 Results

Figure 4 (left part) shows the classification results for the different classification systems and window sizes. The ASL based approach outperforms the baseline system and reaches a classification rate of 84.5 % at window size $l = 90$ seconds, while the others best performance is 72.3 % at $l = 60$ seconds. Combining the two further improves the performance and the overall best of 91.5 % is achieved for $l = 90$ seconds. The corresponding confusion matrix and class accuracies are presented in Fig. 4 (right part). While the classes *meeting* (M), *presentation* (P) and *coffee break* (C) are classified correctly with rates lager 90 % the class accuracy of *discussion* (D) only achieves 80.8 % caused by the confusion with the class *presentation*. This can be explained by the similarity between the two situations, i.e. the former presenter might still be the main speaker when answering questions of the audience in detail.

The effect of the window size can be explained with the increasing amount of context information encoded into a frame which positively affects the classification. However, the smoothing characteristic of the framing decreases the performance for larger l. The perfect parameter value depends on the feature set and differs between the ASL approach and the baseline system.



	M	P	D	C	Class Acc.
M	441	7	2	0	98.0 %
P	0	1157	42	0	96.5 %
D	0	138	582	0	80.8 %
C	0	64	62	1243	90.8 %
			Overall Acc.		91.5 %

Fig. 4: Classification results: ASL —⋆—, baseline system —▲—, combination ——. The confusion matrix for the optimal combined result is shown on the right.

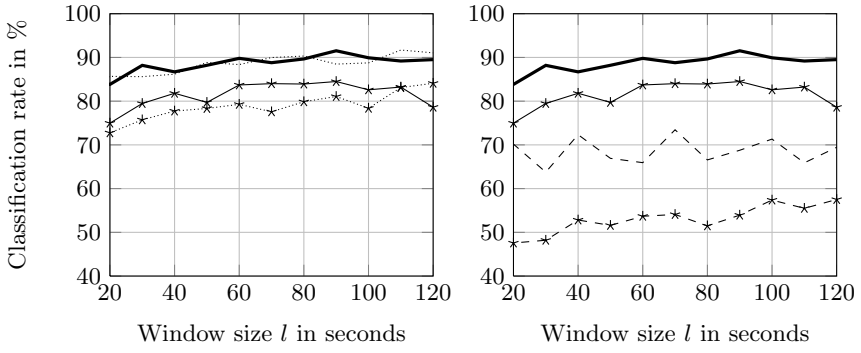


Fig. 5: Classification results for reduced ASL feature sets. Left part: Discarding mean of x, y ···⋆··· and combined ·····; Right part: Additionally discarding mean of z - ⋆ - and combined - - -. In comparison to the results presented in Fig. 4.

Results on the reduced feature set (discarding mean values of x and y-coordinates) in comparison to the former are illustrated in Fig. 5 (left part). While the ASL based classification decreases slightly (83.1 %), the combined performance remains on the same level. In contrast, additionally discarding the mean of z, dramatically reduces the classification rates (57.5 % ASL, 73.4 % combined)(Fig. 5, right part). These findings show the capability of our approach to learn interaction patterns independent of the location of speakers. However, the spatial height of a localized audio source, e.g. whether different interaction partners are sitting or standing, has a great impact on the classification performance of our system and must be regarded for the given task.

5 Conclusion

In this paper we presented a system for context classification in smart environments. The proposed system is the first which utilizes acoustic information about human interaction patterns for this task. Without any prior knowledge a neural network classifier learns acoustic source localization patterns which are typical for specific contexts. Results of experiments in a smart conference room equipped with multiple microphones showed improved performance compared to a baseline system relying on environmental sensors. The integration of both information sources even increased the classification performance to 91.5 % in a demanding, real world scenario.

References

1. Atallah, L., Yang, G.Z.: The use of pervasive sensing for behaviour profiling - a survey. Pervasive Mob. Comput. 5(5), 447–464 (2009)
2. Benesty, J.: Adaptive eigenvalue decomposition algorithm for passive acoustic source localization. JASA 107(1), 384–391 (Jan 2000)
3. Chen, J., Benesty, J., Huang, Y.: Time delay estimation in room acoustic environments: An overview. EURASIP Journal on Applied Signal Processing (2006)
4. DiBiase, J.H., Silverman, H.F., Brandstein, M.S.: Robust localization in reverberant rooms. In: Microphone Arrays, chap. 8, pp. 157–180. Springer (2001)
5. Dong, B., Andrews, B.: Sensor-based occupancy behavioral pattern recognition for energy and comfort management in intelligent buildings. In: Proc. Int. IBPSA Conf. (2009)
6. Henricksen, K., Indulska, J., Rakotonirainy, A.: Modeling context information in pervasive computing systems. In: Proc. Pervasive. pp. 167–180 (2002)
7. Huang, Y.A., Benesty, J. (eds.): Audio Signal Processing for Next-Generation Multimedia Communication Systems. Springer US (2004)
8. van Kasteren, T., et al.: Accurate activity recognition in a home setting. In: Proc. UbiComp. pp. 1–9 (2008)
9. Kleine-Cosack, C., Plötz, T., Vajda, S., Fink, G.A.: Context classification in smart spaces using environmental sensors (2010), submitted to: Ambient Intelligence Forum 2010
10. Knapp, C.H., Carter, G.C.: The generalized correlation method for estimation of time delay. IEEE Trans. Acoust., Speech, Signal Process. 24(4), 320–327 (1976)
11. Kuncheva, L.I.: Combining Pattern Classifiers. Wiley-Interscience (2004)
12. Niu, W., Kay, J.: Location conflict resolution with an ontology. In: Proc. 6th Int. Conf. on Pervasive Computing. pp. 162–179 (2008)
13. Plötz, T.: The FINCA: A Flexible, Intelligent eNvironment with Computational Augmentation (2007), www.finca.irf.de
14. Ranganathan, A., Campbell, R.H.: A middleware for context-aware agents in ubiquitous computing environments. In: Proc. Int. Conf. on Middleware (2003)
15. Turaga, P., et al.: Machine recognition of human activities: A survey. IEEE Trans. on Circuits and Systems for Video Technology 18(11), 1473–1488 (2008)
16. Waibel, A., Schultz, T., Bett, M., Denecke, M., Malkin, R., Rogina, I., Stiefelhagen, R., Yang, J.: Smart: The smart meeting room task at ISL. In: Int. Conf. Acoustics, Speech and Signal Processing. pp. 752–755 (2003)

17. Wilson, D.H., Atkeson, C.: Simultaneous tracking and activity recognition (STAR) using many anonymous, binary sensors. Pervasive Computing pp. 62–79 (2005)
18. Wren, C.R., Tapia, E.M.: Toward scalable activity recognition for sensor networks. In: Proc. LoCA. pp. 168–185 (2006)