Pervasive Sound Sensing: A Weakly Supervised Training Approach

Kelly, D., & Caulfield, B. (2015). Pervasive Sound Sensing: A Weakly Supervised Training Approach. *IEEE Transactions on Cybernetics*, *46*(1), 123-135. https://doi.org/10.1109/TCYB.2015.2396291

Link to publication record in Ulster University Research Portal

Published in:
IEEE Transactions on Cybernetics

Publication Status:
Published (in print/issue): 14/12/2015

DOI:
10.1109/TCYB.2015.2396291

Document Version
Author Accepted version

General rights
The copyright and moral rights to the output are retained by the output author(s), unless otherwise stated by the document licence.

Unless otherwise stated, users are permitted to download a copy of the output for personal study or non-commercial research and are permitted to freely distribute the URL of the output. They are not permitted to alter, reproduce, distribute or make any commercial use of the output without obtaining the permission of the author(s).

If the document is licenced under Creative Commons, the rights of users of the documents can be found at https://creativecommons.org/share-your-work/cclicenses/.

Take down policy
The Research Portal is Ulster University's institutional repository that provides access to Ulster's research outputs. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact pure-support@ulster.ac.uk

Download date: 13/06/2025

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Pervasive Sound Sensing: A Weakly Supervised Training Approach

Daniel Kelly and Brian Caulfield

***Abstract*—Modern smartphones present an ideal device for pervasive sensing of human behavior. Microphones have the potential to reveal key information about a person's behavior. However, they have been utilized to a significantly lesser extent than other smartphone sensors in the context of human behavior sensing. We postulate that, in order for microphones to be useful in behavior sensing applications, the analysis techniques must be flexible and allow easy modification of the types of sounds to be sensed. A simplification of the training data collection process could allow a more flexible sound classification framework. We hypothesize that detailed training, a prerequisite for the majority of sound sensing techniques, is not necessary and that a significantly less detailed and time consuming data collection process can be carried out, allowing even a nonexpert to conduct the collection, labeling, and training process. To test this hypothesis, we implement a diverse density-based multiple instance learning framework, to identify a target sound, and a bag trimming algorithm, which, using the target sound, automatically segments weakly labeled sound clips to construct an accurate training set. Experiments reveal that our hypothesis is a valid one and results show that classifiers, trained using the automatically segmented training sets, were able to accurately classify unseen sound samples with accuracies comparable to supervised classifiers, achieving an average *F*-measure of 0.969 and 0.87 for two weakly supervised datasets.**

***Index Terms*—Diverse density (DD), pattern recognition, pervasive sensing, sound classification, weak supervision.**

I. Introduction

MOBILE phones are rapidly emerging as the ultimate pervasive sensor of human dynamics [1], [2]. Sensing human behavior has huge potential in the area of health and wellbeing. Sensors could potentially be utilized to generate objective daily life measures such as health related quality of life and these objective health measures could, for example, be utilized by clinicians to better understand a patients reaction to particular treatments [3].

While motion and location sensors have been utilized in many lifestyle monitoring works, the use of the microphone, in a smartphone, as a means of lifestyle monitoring is a relatively new area. A smartphone's microphone has the potential to sense lifestyle related measurements that cannot be sensed by motion and GPS sensors. For example, by detecting "voice" and different "ambient" sounds, measurements of an individuals social activity could be obtained.

Automatic detection of sound types is not a new concept. Voice activity detection (VAD) is a sound classification task aimed at detecting the presence of voice [4]. VAD is a mature technology and techniques discussed in the literature perform very well at detecting the presence of speech. However, most VAD techniques take advantage of specific speech characteristics, therefore applying these techniques to general sound classification problems would likely not produce the same levels of detection accuracy.

General sound classification typically deals with the classification of speech, ambient noise, "music" and subcategories of these sound types. For example, music-based classification can look at identifying music instruments or detecting music genre [5], [6]. Other sound classification applications include voice analysis for the purpose of person identification [7] and musical instrument detection [5]. For example, Lin *et al.* [8] propose a wavelet-based feature set which is used with a bottom-up support vector machine (SVM), to classify 16 different sound categories such as "animals," "bells," "machines," "laughter," and "violin" among others. Training was based on a preannotated dataset (DS) and the authors report a classification accuracy of 97%. Mogi and Kasai [9] proposed a classification system which takes into account the more difficult task of classifying sounds captured in noisy environments using smartphones. Experiments, based on six different sound categories, were carried out to evaluate the proposed system. On average, 122 sounds samples of four seconds duration, were used to train classifiers for each sound category. Testing was carried out on an average of 52 sound samples per category and results reported an average recognition accuracy of 76%.

There exists a large number of related works addressing the problem of general sound classification, however, there is a limitation in the majority of studies carried out to date. This limitation is that a sizable training DS, and detailed labeling of the training set, is required. Labeling data is a time consuming process and the need to label data limits the scalability of the sensing applications. This problem is of particular interest when considering a general purpose sound classification system. A behavior sensing application implemented on a smartphone can encounter a potentially unlimited set of sound categories that could be of interest. A flexible and general

Manuscript received February 12, 2014; revised May 29, 2014 and December 11, 2014; accepted January 9, 2015. This work was supported by Science Foundation Ireland. This paper was recommended by Associate Editor B. W. Schuller.

D. Kelly is with the School of Computing and Intelligent Systems, Ulster University, Derry BT48 7JL, U.K. (e-mail: dnl.kelly1@gmail.com).

B. Caulfield is with the Insight Center for Data Analytics, University College Dublin, Dublin, Ireland.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2015.2396291

2168-2267 © 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

purpose classification framework is, therefore needed where the types of sounds to be recognized can be easily tailored to the particular application. A key problem which restricts the implementation of a general purpose sound classification system is training data is needed for different sound categories which in turn requires a detailed and time consuming data collection and labeling process. A simplification of the training data collection process would make way for a more flexible sound classification framework. It is, therefore important that studies are carried out to investigate the problem of data labeling and potential methods of reducing the need for labeled data.

There is a very limited number of works in the literature that deal with the area of weak supervision and sound classification. One work which carries out experiments in the area is by Zhang and Schuller [10]. In this paper, a classifier A is initially training on a DS A using standard supervised methods. However, an additional unlabeled DS B is automatically labeled by the output of classifier A. The combination of the supervised DS A and the automatically labeled DS B are then utilized to train an overall classifier which results in a classifier with overall improved recognition performance. While this is an interesting solution to producing larger training DSs, the training of the models still require an initial labeled DS.

Another promising study, carried out by Lu *et al.* [11], deals with classifying sounds, recorded from smartphones, with limited labeling. The authors propose "SoundSense," where a two stage classification process is used. Firstly, sounds are categorized as voice, music, or ambient using a decision tree classifier. Secondly, intra category classification is carried out depending on the initial category. Specifically, if a sound is defined as ambient, a further unsupervised classification stage is performed to determine further information about the ambient sound. Unsupervised classification is based on clustering each new ambient sound into a set of B clusters where B is the predetermined number ambient sounds that will be modeled. A simple Bayes classifier is used to assign a sound to a cluster. If it is not appropriate for a sound to be assigned to an existing cluster, a new cluster is created. Less significant sound clusters are removed to make way for new and more significant sounds. Users of the sensing application are asked to input some semantic meaning about each cluster. While this paper is an important step toward less supervised training of sound classifiers, there is still an element of traditional labeling required for this method to work. Specifically, the first stage of the classification requires that a decision tree classifier be trained on labeled examples of voice, music, and ambient sounds.

In this paper, we propose a learning framework that can classify sounds without the need for detailed training data labeling. Specifically, we propose a weakly supervised framework, where data labeling is carried out in a much less detailed and less time consuming way.

In a traditional supervised machine learning scenario, each training instance x requires an associated class label y. Multiple instance learning (MIL), however, is a learning framework which uses significantly weaker labeling information where labels are not assigned to the individual training instances, but instead assigned to sets of instances named bags. A bag is labeled positive if at least one instance in the bag is positive. Conversely, a bag is labeled negative if all instances in the bag are negative. Labels of individual instances are not known and the aim of MIL is to find the optimal labeling of the individual instances in positive bags.

There have been numerous MIL solutions proposed in the literature. Maron and Lozano-Pérez [12] proposed a diverse density (DD)-based method where the goal is to identify a target concept which is similar to positive bags but dissimilar to negative bags. Classification is then based on a weighted similarity between an unseen bag and the target concept. Zhang and Goldman [13] proposed an improvement to the DD method by applying an expectation maximization (EM) layer to the target concept identification process. An alternative approach, and one of the most utilized techniques used for MIL, is the work proposed by Andrews and Tsochantaridis [14]. Andrews and Tsochantaridis [14] proposed a multiple instance SVM (miSVM) where modifications and extensions are applied to the standard SVM framework such that SVM kernel-based classifications can be carried out in a MIL scenario.

In general, the task of classic MIL is to train a classifier that can predict labels of new bags. Classic MIL has been widely applied in various areas, for example, two seminal works in the area of MIL are based in the areas of drug design and image scene classification. Dietterich [15] proposed a molecule classification system, used for drug design, where each molecule is represented by a bag of molecule conformations. Later, Maron and Ratan [16] proposed an image classification system, where an image is represented by a bag of image patches.

MIL can be utilized for sound classification by using bags of sound feature vectors such that a bag would correspond to a sound clip recording. Adopting an MIL approach means that, for training, sound labels do not have to be provided for each sound feature vector. Instead, labels are provided on a very course level such that sound feature vectors are grouped into bags and labels are provided for the bags/sound clips. MIL has successfully been utilized in other domains of human behavior analysis, such as gesture recognition [17] and activity recognition [18]. However, to our knowledge, no other work has investigated the use of MIL or weak supervision in the area of general sound classification.

A factor which limits the use of MIL in the context of sound classification is that classic MIL predicts a single label for an entire bag. A single label for an entire sound clip would provide very limited information in a sound classification scenario. Rather than classifying bags, a sound classification framework requires a classifier that can predict labels of individual sound instances. This problem is known as key instance detection (KID) [19] and has only recently received any attention in the literature. Most MIL training methods are based on the idea of identifying a single "most positive instance in a bag" and basing the classification of bags on a comparison of these "most positive" instances. KID is based on the concept of identifying all positive instances, also known as key instances, responsible for the bag label. To our knowledge, the only work to investigate the idea of KID was by Liu *et al*. [19], who proposed a voting framework to form a citer kNN graph which

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

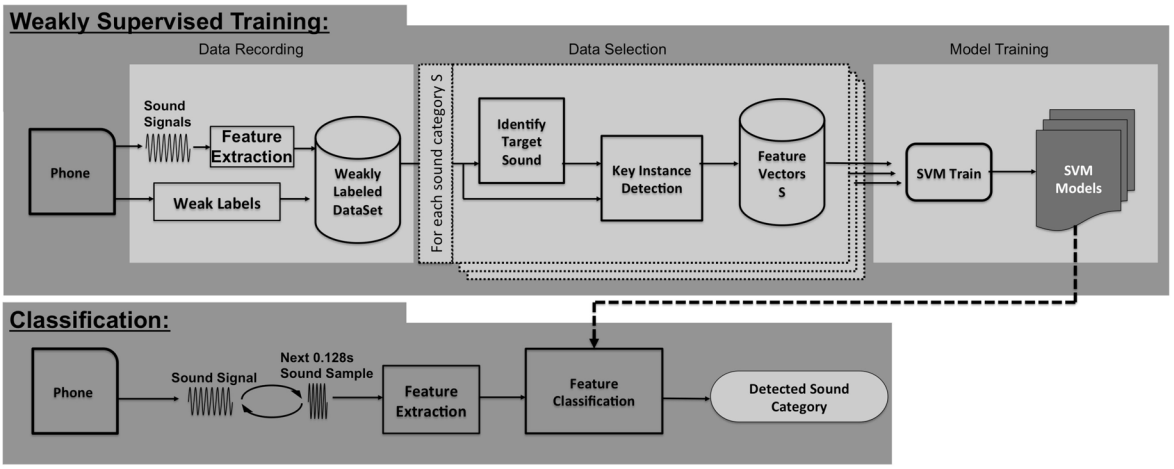

Fig. 1. Overview of proposed system.

could predict the label of individual instances. In this paper, we propose a MIL framework specifically with sound classification in mind where a KID approach is taken to identify instances which are responsible for bag labels. These instances are then utilized to train classifiers such that classification of unseen sound instances can be performed.

Fig. 1 illustrates the overall framework of the work proposed in this paper. It can be seen that sounds are first recorded on a smartphone. Weak labels are then applied to each sound clip by confirming if any of the predetermined sound categories occurred in the sound clip. For each sound category to be learned, a target sound identification and KID process is carried out. Key instances of each sound category, which are identified in the previous steps, are utilized to train a set of SVMs. The trained SVMs are then utilized in the classification stage to predict the sound category of sound feature vectors. In Section II, we will give an overview training stage of the framework, including a description of the feature extraction techniques used as well as details on the target sound identification and KID aspects of the algorithm. In Section III, we discuss evaluations carried out on our framework including tests carried out on the classification stage of the framework.

II. Methods

A. Feature Extraction

In this paper, sound clips are recorded on a standard Android smart-phone. Sound signals are first segmented into 128 ms windows. Windows which contain only silence, determined by thresholding the average magnitude of sound windows, are removed.

For a given frame window w_t, a feature vector, f_t, is extracted in order to describe the audio characteristics of the sound at time t. In this paper, we use mel-frequency Cepstral coefficients (MFCC) as the main set of features to characterize sound frames [20]. In addition to MFCC features, we also utilize a set of time and frequency domain summary features which have been described and validated in previous papers. These features are described in Sections II-A1–II-A8. While we detail the different features used in this paper, the focus of this paper is not on sound features. The overall learning framework which is described does not rely specifically on the features described here and alternative sets of features can be used in place of the described feature set.

1) MFCCs: Mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a signal [21]. In this paper, we extract a set of 12 MFCCs which we denote as $M_t = \{m_1, \ldots, m_{12}\}$.

2) Spectral Flux (SF): SF measures the change in the shape of the frequency spectrum s_t calculated from sound frame w_t using fast fourier transform. Equation (1) details the calculation of SF, where $s_t(i)$ refers to the magnitude of the ith frequency bin [22]

$$\mathrm{SF}_t = \sum_{i=1}^{n} [s_t(i) - s_t(i-1)]. \tag{1}$$

Speech, for example, generally switches quickly between voice and unvoiced sound. This results in quick changes in the frequency spectrum and thus gives speech a higher SF when compared to nonspeech audio.

3) Spectral Rolloff (SR): SR is defined as the frequency bin below which 93% of the distribution is concentrated [11], [23]

$$\mathrm{SR}_t = \max\left[h| \sum_{i=1}^{h} s_t(i) < T\right] \tag{2}$$

$$T = 0.93 \sum_{i=1}^{h} s_t(i). \tag{3}$$

Sounds which tend to have high frequency sounds, such as music, will have a higher SR.

4) Spectral Centroid (SC): Defined as the balancing point of the spectral power distribution [23]

$$\mathrm{SC}_t = \frac{\sum_{i=1}^{n} i \times s_t(i)}{\sum_{i=1}^{n} s_t(i)}. \tag{4}$$

Higher frequency sounds will have higher balancing points.

5) Bandwidth: Spectral spread is the width of the range of frequencies that a sound contains. It calculates the extent

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

at which the frequencies are spread out over the spectrum in relation to the SC SC_t [23]

$$\mathrm{BW}_t = \frac{\sum_{i=1}^{n}(i - \mathrm{SC}_t)^2 \times s_t(i)^2}{\sum_{i=1}^{n} s_t(i)^2}. \tag{5}$$

Most ambient sounds consist of a limited range of frequencies and thus have a small spectral spread.

6) Normalized Mel-Frequency Bands: Frequency bins are used to represent the basic distribution of sound frequencies on the Mel-frequency scale. We use eight normalized Mel-frequency bins $B_t = \{b_{t1}, \ldots, b_{t8}\}$ where b_{ti} is a frequency bin representing the ith frequency range and $\sum_i^8 b_{ti} = 1$. This set of features is used to discriminate between high frequency, mid range frequency and low frequency sounds.

7) Zero Crossing Rate (ZCR): ZCR is defined as the number of time-domain zero-crossings within a frame. For example, human voice shows a higher variation in ZCR when compared to music and ambient sound [24].

8) Low Energy Frame Rate: Low energy frame rate corresponds to the number of sub-windows (windows of 0.5 ms duration within the 128 ms window) that have a root mean square (RMS) value less than 50% of the RMS of the overall 128 ms window. In human speech, there are more quiet frames, thus this measure will be higher for speech related sounds [22].

B. Bags of Sound

A feature vector instance f_t describes the characteristics of a sound at time t using the features described above. A recorded sound clip is used to compute a set of feature vector instances $F = \{f_1, \ldots, f_T\}$. Each sound clip will also have an associated set of weak labels $L = \{l_1, \ldots, l_N\}$ and each label is an element of a discrete set of predefined sound categories $l \in C$. It should be noted that in a weakly supervised learning framework, such as the one described in this paper, the number of labels N is significantly smaller than the number of feature vectors T. For example, over 4500 feature vectors will be extracted for a 10 min sound clip, while the number of labels given to the sound clip would typically be in the order of 1–3. The combination of the set of sound clip feature vectors and the set of weak labels, for the purpose of MIL notation, is called a bag $B_n = \{F_n, L_n\}$ where n denotes the index of the sound clip which the feature vectors and labels are associated with.

In order to find the individual feature vectors in each bag that optimally describe a particular sound class $c \in C$, the bags are arranged into a set of positive and negative bags. A bag B_n is labeled as positive if L_n contains sound class c. An example scenario is as follows; the overall set of sounds we are training is $C = \{\text{Voice, Music, Ambient}\}$ and the current sound we are training is voice. In this scenario, a bag is labeled positive if, at some point during the recording of the associated sound clip, a voice occurred and this occurrence was flagged in the set of weak labels. Conversely, a bag is labeled as negative if at no point during the recording of the associated sound clip did a voice occur.

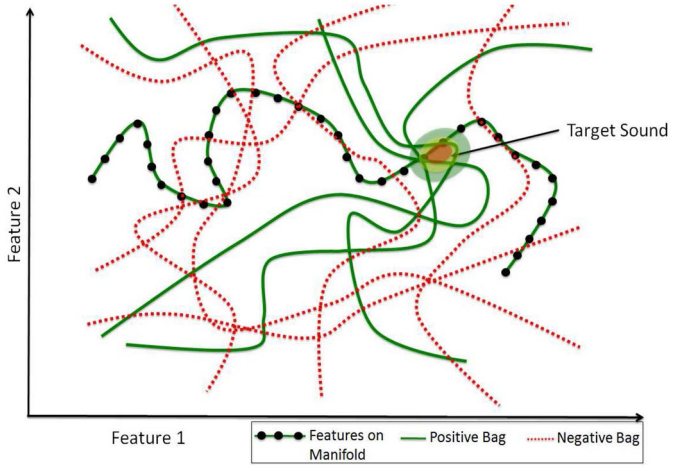

Fig. 2. Multiple positive and negative sound manifolds. Target sound found by finding area where positive manifolds intersect and no negative manifolds intersect.

C. DD

DD is a general framework for solving MIL problems [12]. In this paper, we extend this general framework in order to apply it to the problem of KID in general sound classification. A single sound feature vector f_t represents a point in m-D feature space, where m is the total number of features used to characterize a sound. A set of feature vectors F, for a particular sound clip, will trace out a manifold through this m-D space. Fig. 2 illustrates multiple manifold with toy data. For illustration, one of the manifolds (with black dots) has a set of 37 2-D feature vectors plotted as a manifold.

If a sound clip is labeled as positive, at least one location exists on the manifold where a feature vector represents the target sound. In reality, it is likely than a number of points on the manifold will be very similar to the target sound. Conversely, if the sound clip is labeled as negative, we know that none of the sound characteristics along the manifold represent the target sound. If there are multiple positive and negative bags, and we assume that there exists a point in the feature space that optimally represents the target sound, the goal of DD is to identify an optimal point on the feature space where all positive feature manifolds intersect without intersecting any negative feature manifolds. Fig. 2 also illustrates other positive and negative manifolds, each of which would be constructed from sets of feature vectors. It can be seen that the target sound is identified by finding the area where positive manifolds intersect and no negative manifolds intersect.

To identify a target sound in a real world DS, a measure of DD can be utilized. The main principal of the DD framework is made up from probability density measures $P^+(x = h|B_i^+)$ and $P^-(x = h|B_i^-)$, which compute the density of positive points and the sparsity of negative points, for a given concept sound x, respectively. Assuming there exists an optimal target sound h, the goal is to identify the target sound by simultaneously maximizing the density of positive points and sparsity of negative points over all concept sounds x in the feature space. This is formally described in (6)–(8), where x is maximized in order to identify a point in the feature space which has a

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

high density of positive points and a low density of negative points

$$\underset{x}{\operatorname{argmax}} \prod_i P^+\left(x = h|B_i^+\right) \prod_i P^-\left(x = h|B_i^-\right) \tag{6}$$

$$P^+\left(x = h|B_i^+\right) = 1 - \prod_j \left(1 - P\left(x = h|B_{ij}^+\right)\right) \tag{7}$$

$$P^-\left(x = h|B_i^-\right) = \prod_j \left(1 - P\left(x = h|B_{ij}^+\right)\right) \tag{8}$$

$$P(x = h|B_{ij}) = exp\left[-\sum_k x_s^2(k)\left(B_{ijk} - x_f(k)\right)^2\right]. \tag{9}$$

The individual density probability, $P^+(x = h|B_i^+)$, is modeled on the probability that not all points are different from the concept sound. Thus, $P^+(x = h|B_i^+)$ is high if at least one instance in the bag is close to x. Conversely, the sparsity probability, $P^-(x = h|B_i^-)$, is modeled on the probability that all points are different from the concept sound. If every positive bag has an instance close to x and no negative bags are close to x, x will have a high DD.

The probability that an individual sound instance, B_{ij}, is the same as a concept sound is based on a distance between them. Different features will have different levels of importance in terms of accurately measuring the similarity of two sounds. The similarity between a feature vector, B_{ij}, and a concept sound x is, therefore defined in (9) as a weighted distance between individual features. Where B_{ijk} is the kth feature of the jth feature vector in the ith sound bag. The target sound, h, comprises a target feature vector, h_f, and a scaling vector component, h_s where $h_s(k)$ is a weighting for the kth feature.

Since the target sound is made up of both a feature vector component and a scaling component, the goal of maximization is to find a combined optimal point in the feature space and an optimal weighting for each individual feature dimension. In this paper, we use the limited memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) [25] method of gradient descent and optimization in order to identify the optimal target sound h which produces the maximum DD. L-BFGS will perform a hill climbing optimization process in the feature space in order to steer an approximate solution feature toward a solution which maximizes (6).

D. Dimension Reduction for Target Optimization

In this paper, the feature vector utilized to characterize a sound is comprised of a total of 26 features, which were discussed in Section II-A. An additional 26-D scaling vector is also utilized during the target maximization process, thus, the overall target maximization is based on a 52-D feature space. Preliminary evaluations of this paper showed that the computation time to identify the target sound h was quite large. Based on a basic set of positive and negative bags, where each bag contained, on average, 5 minutes of sound feature vectors and positive and negatives sets contained approximately 10 bags each, computation of an optimal target sound took over two hours on a machine with a 2.4 GHz Intel Core i7 CPU and 4 GB RAM.

In order to reduce the computation time of the target maximization process, we first carry out a dimensionality reduction process on the feature vectors using principal component analysis (PCA). We denote a sound feature vector which has been projected onto a lower dimension using PCA as $\bar{f}_t$. For the purpose of target sound maximization, all bags are comprised of lower dimension sound feature vectors $\bar{f}_t$. Through preliminary classification experiments, we concluded that reducing the feature space to $N = 5$ dimensions produced the best trade-off between computation time and overall classification performance. The complexity of the target sound identification is linear with respect to the feature dimension, therefore reducing the feature vector from $N = 26$ to $N = 5$ improved the target sound identification computation time by an order of approximately 21.

E. Bag Trimming

On completion of the target maximization process, the target sound, h, optimally represents a single point within the feature space, along with the associated weights for individual feature dimensions, relating to a sound category given sets of positive and negative bags. In previous DD-based works [12], [13] discussed in Section I, a bag is classified through a weighted distance between all feature instances in the bag and the target concept. A bag is classified as positive if the weighted distance for at least one instance is below a set threshold. We postulate that a problem with this approach, in the context of sound classification, is that it is unlikely that a single target concept will sufficiently characterize an entire sound category. We, therefore propose a KID approach where the target sound is utilized as the basis for identifying a set of key instances which more accurately characterize sound categories.

The proposed KID approach is a bag trimming algorithm which iteratively removes feature vector instances from positive bags until each bag contains only feature vectors that relate to the target sound. We now describe this process.

Each individual positive bag is considered in conjunction with all negative bags. We consider the set of feature vectors in negative bags as a single set of feature vectors $\Omega = \{F_1^-[1], \ldots, F_1^-[T_1], \ldots, F_M^-[1], \ldots, F_M^-[T_M]\}$. Where $F_i^-[j]$ is the jth feature vector in the ith negative bag, T_i denotes the total number of feature vectors in the ith negative bag and M is the total number of negative bags.

The proposed bag trimming technique is based on an unsupervised clustering technique where feature vectors of each positive bag are assigned to a positive or negative cluster. For each positive bag B_i^+, all feature vectors are initially assigned to the positive cluster and all points in the set of negative feature vectors are initially assigned to the negative cluster. An iterative process is performed whereby positive feature vectors, which are deemed to be a closer fit to the negative cluster than the positive cluster, are removed from the positive cluster and reassigned to the negative cluster. This process is repeated until it is no longer appropriate to assign any remaining feature vectors in the positive cluster to the negative cluster. Feature vectors are reassigned based on a comparative distance metric where the distance between a positive instance and the target sound h is compared with the distance between the instance and the negative set Ω.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

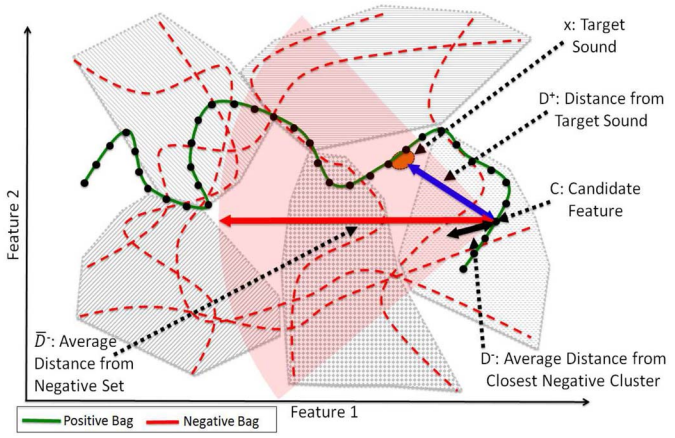

Fig. 3. Distance comparison carried out using negative subsets. Problem using average distance from entire negative set ($\overline{D}^-$) therefore, distance comparison carried out using negative subsets (D^-).

Initially, the distance between a positive point and the negative set was measured using an average distance between the positive point and all negative points in the negative set. However, a problem with this method was observed during initial experiments, where it was found that representing the entire set of negative feature vectors as a single cluster was an over-generalization. The set of negative sounds most likely will contain a multitude of different sound types where some of the sound types may be similar to the target sound category and some may be very different. A single metric, evaluating whether a sound is closer to a multitude of different sound categories or a specific sound category, will not accurately reflect the key differences between the sound categories. For example, a candidate sound feature vector could relate to an ambient noise which is very different to the target sound category of voice. We would expect that this candidate sound should be measured as relatively close to the negative cluster and thus reassigned to the negative cluster. However, if the entire negative set contains a majority of other sounds which are more similar to voice than the candidate sound, then the computed distance between the candidate sound and the negative set will be relatively large even though the candidate feature vector should not be assigned to the positive cluster. The distance metric $\overline{D}^-$ in Fig. 3 illustrates this problem where it can be seen that D^+ is less than $\overline{D}^-$, which is undesired because the candidate feature should not be assigned to the target sound.

A potential solution to this might be to simply compare the distance between a candidate and a single negative point that is closest to the candidate. However, we observed that due to the extreme imprecision of the labeling, and noisiness of the data, it was common to have some instances within the negative DS that were extremely close to the target sound. An alternative solution, and the solution we implement to overcome this problem, is to split the negative set into a number of subsets, with the aim that each subset will approximately represent each of the different sound categories within the negative set Ω. We denote the negative subsets as $\hat{\Omega} = \{\Upsilon_1, \ldots, \Upsilon_K\}$, where K is the total number of subsets. The data is split into subsets using a K-means clustering algorithm. It is likely that the number of different sound categories represented in the negative set will vary for different DS recordings. The number of clusters, K, needed to accurately partition the data into different sound categories is, therefore not known *a priori*. In order to identify K, clustering is carried out using different values of K, where $2 \leq K \leq 10$, to evaluate which K best partitions the data. A cluster validation algorithm is used to measure the suitability of different values of K. A number of cluster validation techniques have been proposed in the literature to measure how well the clustering results fit the underlying data [26] and, in this paper, a validity metric known as *SDbw* is implemented [27]. The *SDbw* metric measures cluster compactness and separation while also taking into consideration the density of the clusters. The value of K which produces the lowest *SDbw* measure is used to partition the negative DS. A key component required for the bag trimming technique is a measure of the fit of the candidate feature vector with the negative set. With the introduction of the negative subsets, the fit can now more accurately be measured by calculating the average distance between the candidate feature vector and all negative instances within the negative cluster which is closest to the candidate feature. The distance measure D^+ in Fig. 3 illustrates how a candidate feature vector is evaluated using the negative subsets where the candidate feature is correctly deemed closer to the negative subset than the target sound.

Equations (10)–(13) define the comparative distance metric s_{ij} where $-1 \leq s_{ij} \leq 1$. The metric evaluates the fit of each candidate feature vector, B_{ij}^+, with the target sound h and the negative sets $\hat{\Omega}$. The metric is based on positive and negative distance measures a_{ij} and b_{ij}, respectively, where a_{ij} is the weighted distance between the candidate point and the target sound and b_{ij} is the average weighted distance between the candidate sound and the closest negative subset. The comparison metric yields results close to 1 when the candidate sound is close to the target sound while results are close to -1 when the candidate sound is closer to one of the negative subsets than it is to the target sound

$$s_{ij} = \frac{b_{ij} - a_{ij}}{\max(a_{ij}, b_{ij})} \tag{10}$$

$$a_{ij} = D(B_{ij}^+, h_f, h_s) \tag{11}$$

$$b_{ij} = \min_k \sum_l^{N_{\Upsilon k}} \frac{D\left(B_{ij}^+, \Upsilon_{kl}, h_s\right)}{N_{\Upsilon k}} \tag{12}$$

$$D(p_1, p_2, s) = \sqrt{\sum_i s_i^2 (p_{1i} - p_{2i})^2}. \tag{13}$$

Algorithm 1 details the complete bag trimming procedure where each positive bag, B_i^+, is analyzed in conjunction with the negative subsets, Υ, in order to identify positive instances, B_{ij}^+, which should be reassigned to negative clusters. Upon completion of the bag trimming algorithm, each positive bag contains only key instances. Fig. 4 illustrates the set of positive points after the bag trimming procedure has been applied.

F. Classifier Training

The final stage of the learning process is a classifier training step. A classifier is trained using the set of key

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Algorithm 1: Bag Trimming Algorithm

```
for each B_i^+ ∈ B^+ do
    Construct Negative Set Ω;
    Compute Negative Subsets Ω̂ = {ϒ_1, ..., ϒ_K};
    Complete = false;
    while !Complete do
        S_min = 1;
        j_min = −1;
        for each B_ij^+ ∈ B_i^+ do
            if S_ij < S_min then
                S_min = S_ij;
                k̂ = Closest Negative Cluster Index;
                j_min = j;
            end
        end
        if S_min < 0 then
            Add B_ij_min^+ to ϒ_k̂;
            Remove B_ij_min^+ from B_i^+;
        else
            Complete = true;
        end
    end
end
```

instances, which remain in the positive bags after the bag trimming procedure, and the set of all negative points Ω. The goal of the training procedure is to compute a model that can accurately discriminate between the key instances and negative points. Through extensive preliminary experiments, carried out using DSs which will be discussed in the later section, we found that a SVM classifier using a radial basis function (RBF) kernel performed the best when compared to different decision trees, neural networks, and naive Bayes classifiers. This is consistent with recent works, such as the work carried out by Geiger *et al.* [28] and Guo and Li [29], where SVMs have been successfully applied in general sound classification problems. There are two parameters to consider while training RBF kernels: C and γ. We carry out a V-fold cross-validation to compute optimal values for C and γ. The parameter combination which produced the maximum average F-measure was chosen to train the SVM.

It should be noted that the target sound identification and bag trimming techniques are carried out using bags which are comprised of lower dimension sound feature vectors $\bar{f}_t$. However, we found that during the classification stage, classifiers perform with a higher classification accuracy when trained on the corresponding raw higher dimension feature vectors f_t. This was evaluated by carrying out preliminary evaluations, similar to the classification accuracy experiments which will be discussed in Section III-C, by evaluating the classification accuracy of the system trained on lower dimension sound feature vectors and the original feature vectors. It was found the classifiers trained on the original feature vectors achieved an 8% improvement in classification accuracies in comparison to the classifiers trained on the PCA reduced features.

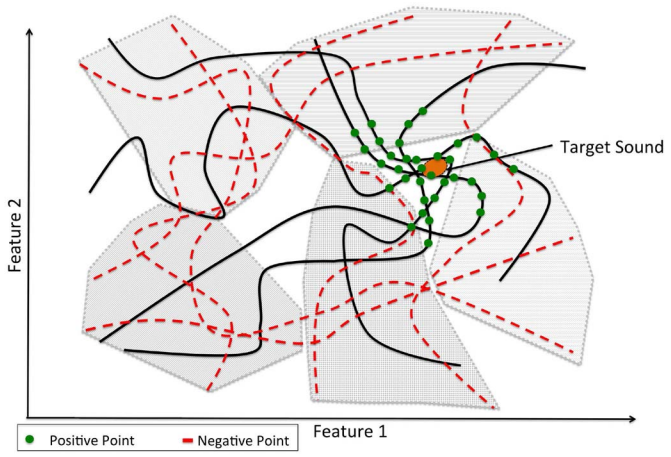

Fig. 4. Final set of positive points after bag trimming process using comparative distance measures s_{ij}.

III. Evaluation

A. DSs

Two sound DSs, each DS comprising of a training set and test set, were recorded using smartphones (Samsung Galaxy S3) with a custom built sound recording Android app. Audio was recorded with an 16 bit PCM encoded single channel with a sampling rate of 44 100 Hz. The training sets comprised of sound clips and associated weak labels. As previously discussed, detailed labeling is a difficult and time consuming process which limits the application of general purpose sound classification. We have, therefore proposed a weakly supervised solution in this paper where a weak label simply defines the presence of a sound category during an associated training sound clip. Weak labels do not hold any other information such as duration or timestamps of the actual sound category.

For testing, feature vectors were extracted from the test sound clips but they were not stored in bags. Instead, each test instance was stored along with a ground truth label. It is, therefore important to note that, while training is carried out on weakly labeled bags, testing is carried out on the more difficult task of frame level classification as opposed to bag classification. Each feature vector, corresponding to a 128 ms window, has an associated label as defined during a manual labeling process. This is a different approach to many MIL-based works where testing is usually carried out on bags as opposed to feature vector instances. This manual labeling process was carried out by listening back to each of the test clips and carefully noting the start and end frame when a predefined sound category occurred. Feature vectors which occurred between the noted start and end frame were labeled as the corresponding sound category. It should be noted that manual labeling discussed in this paper is only carried out for testing and evaluation purposes and only weakly supervised labels are required in order to train our proposed system.

1) DS 1: When recording sound clips for DS 1, a multitude of different scenarios that occur in a kitchen were recorded.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

During these scenarios sounds which occurred included cooking sounds, chopping, "water" boiling, kettle boiling, drawers and cupboards opening and closing, brushing, vacuuming, scrubbing pots, and many other common kitchen sounds. Additionally, the two chosen classification sound categories "cutlery" noise and water pouring occurred. All sounds where recorded by leaving a smartphone approx. Five meters away from the source of the sounds. A set of common kitchen tasks were then carried out. When either of the two sound categories occurred during a sequence of tasks, weak labeling was performed by noting the occurrence of the sound, by simply ticking a checkbox, when the sound clip recording was completed.

For the training set, 20 sound clips were recorded. In total, 204 min of sound data was recorded with an average sound clip duration of 10 min 12 s. At least one of the chosen sound categories occurred in each of the 20 sound clips. The cutlery sound had an average duration of approximately 5 s in each sound clip and the water sound had an average duration of approximately 9 s. The cutlery and water sounds occurred in 12 and 11 of the sound clips, respectively. Both sounds occurred in 3 of the 20 sound clips.

For the test set, an additional 20 sound clips were recorded. In total, 27 min of test sounds were recorded with an average sound clip duration of 1 min 21 s. Manual labeling was carried out on the test sound clips such that each feature vector instance was assigned a ground truth label. We make the assumption that labeling training data more frequently than at 10 min intervals would be unlikely. From all test data, a total of 5 min 37 s of the cutlery sound was labeled while a total of 8 min 26 s of the water sound was labeled. The remaining 12 minutes of test sound represented different sounds which typically occur in a kitchen.

2) DS 2: Three general sound categories were chosen for DS 2 in order to robustly test our framework under different conditions compared to DS 1. The chosen sound categories are: voice, music, and ambient sound. During the recording of DS 2 sound clips, various types of voice, music, and ambient sound was encountered. Voice sounds included various numbers of people talking in different scenarios such as meetings, social and phone conversations. Different genres of music were recorded from music playing through different devices such as high quality speakers, basic car speakers and laptop speakers. It should be noted that vocals during music were not labeled as voice. Finally, ambient sounds were recorded from a variety of scenarios including general office noise, cafe/restaurant noise, traffic noise, building/roadworks, and general sounds in the home. If either of the three sound categories occurred during the sound recording, weak labeling was performed by noting the occurrence of the sound when the sound clip recording was completed. For the test set, we recorded a varied set of sound clips where, for example, different people were recorded for the voice recording, different songs were recorded for the music recording and different locations were recorded for the ambient sounds. It should be noted that the training set and test set for this DS was recorded in uncontrolled and noisy environments. The DS, as such, represents real world sounds that would be encountered by a smartphone in everyday life.

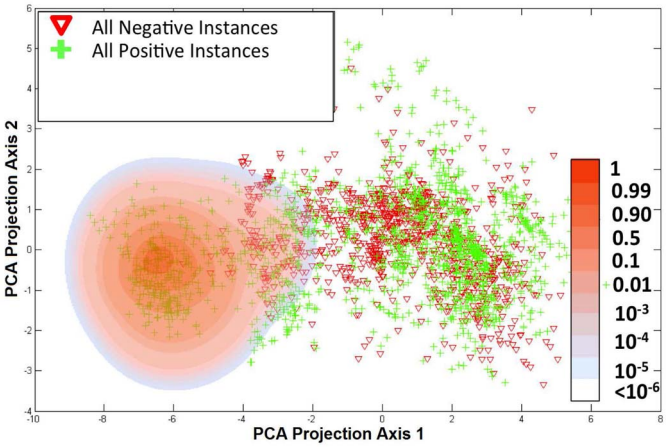

Fig. 5. DS 1: finding cutlery from general sounds—feature vectors in positive and negative bags and DD probability heatmap.

For the training set, 26 sound clips were recorded. In total, 255 min of sound data was recorded with an average sound clip duration of 9 min 49 s. At least one of the sound categories occurred in each of the 26 sound clips. The voice, music, and ambient sounds all had a similar average duration of approximately 4 min in each of the sound clips which they occurred. Of the 26 sound clips, 23 contained at least 2 of the sound categories. For example, training sound clips were recorded in an office kitchen area where both voice and ambient noise occurred. For the test set, an additional 20 sound clips were recorded. In total, 211 min of test sounds were recorded with an average sound clip duration of 10 min 55 s. Manual labeling was carried out on the test sound clips such that each feature vector instance was assigned a ground truth label. From all test data, a total of 52 mins of voice was labeled, 85 mins of music was labeled and 74 min of ambient was labeled.

B. Qualitative Analysis

In this paper, we implement a target sound identification process, as described in Section II, to identify positive and negative samples. In this section, we visually examine feature vectors computed from our DSs with the aim of further understudying the identification of a target sound. While sounds are represented by 26-D feature vectors, for the purpose of illustration in this section, PCA was used to reduce the dimension of the feature vectors to 2-D. In order to not overcrowd the visualizations, the data is illustrated using subsets of the original DSs. Each subset is comprised of feature vectors extracted, from its corresponding overall DS, at uniform intervals.

Fig. 5 illustrates the set of positive feature vectors and the set of negative feature vectors. Additionally, a heat map illustrates the DD probability for different candidate target sounds for the cutlery noise sound. The target sound which relates to the cutlery noise sound corresponds to the point in the heat map with the maximum probability. It can be seen that the heat map area of interest covers an area where there is many positive points but very few negative points. Fig. 6 illustrates the set of key instances which have been deemed similar to the target sound using the bag trimming procedure.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

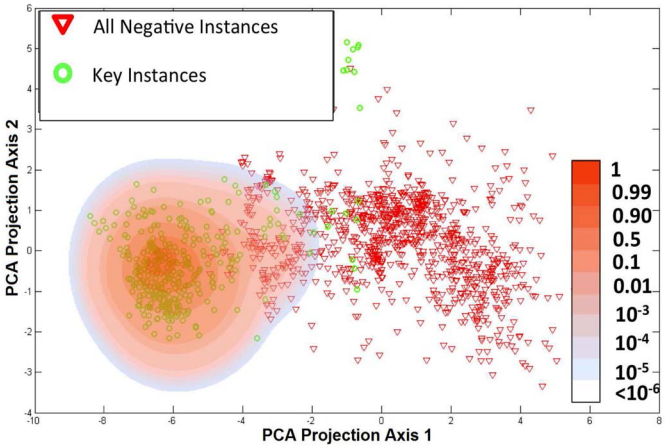

Fig. 6. DS 1: finding cutlery from general sounds—automatically identified key instances.

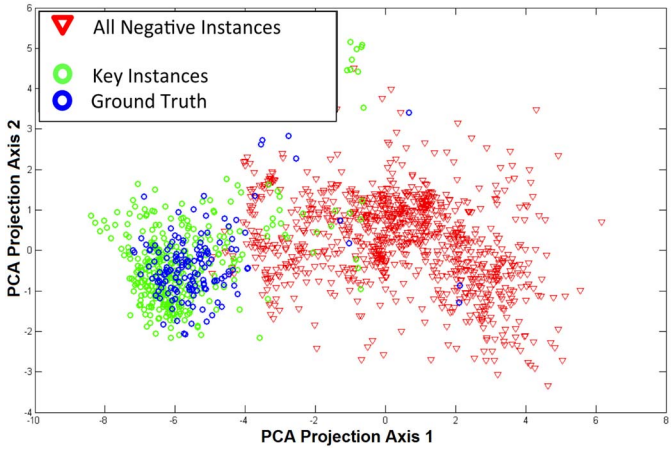

Fig. 7. DS 1: finding cutlery from general sounds—comparison with ground truth positives.

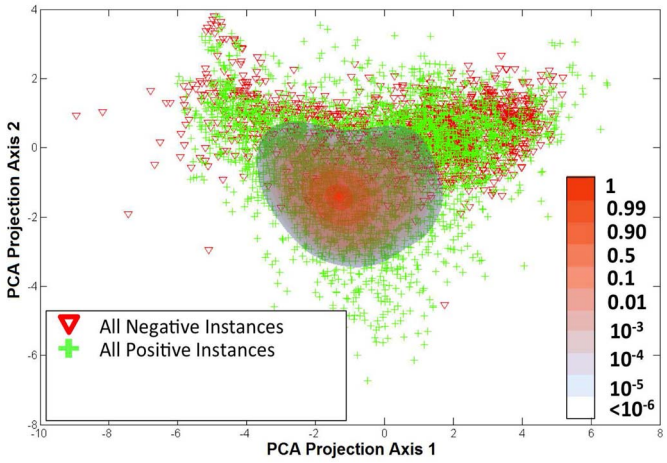

Fig. 8. DS 2: finding voice from general sounds—feature vectors in positive and negative bags and DD probability heatmap.

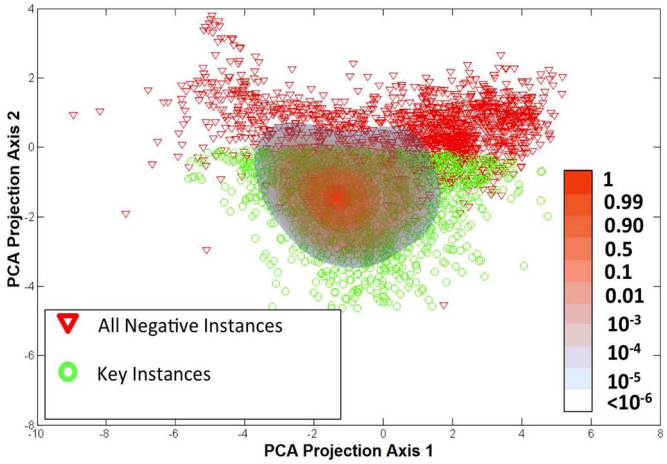

Fig. 9. DS 2: finding voice from general sounds—automatically identified key instances.

Finally, Fig. 7 illustrates the automatically chosen set of key instances, all negative feature vectors, and a set of ground truth feature vectors, manually labeled as cutlery sounds, extracted from an additional sound clip. It can be seen that the manually labeled ground truth feature vectors occupy the same area of the feature space as the automatically labeled key instances. This is a strong indication that our approach is a valid one and shows that the MIL, DD, and bag trimming techniques are an appropriate set of techniques to identify key instances from weakly labeled data.

As with DS 1, we carry out a qualitative examination of DS 2. Fig. 8 illustrates the set of positive bag feature vectors and the set of negative bag feature vectors. Additionally, a heat map illustrates the DD probability for different candidate target sounds for the voice sound. The key difference between DSs 1 and 2 is the chosen sound categories in DS 2 represents more general sound categories compared to the specific sound categories being targeted in DS 1. Evidence of this can be seen by observing the voice heat map, in Fig. 8, compared to the cutlery sounds heat map in Fig. 5. The cutlery noise heat map occupies a more distinct area of the feature space when compared to the voice heat map. Fig. 9 illustrates the set of key instances which have been deemed similar to the target sound, and therefore labeled as voice features, using the bag trimming procedure. Finally, Fig. 10 illustrates the automatically chosen set of key instances, all negative feature vectors, and a set of ground truth feature vectors, manually labeled as voice, extracted from an additional sound clip. As with the previous example, it can be seen that the manually labeled ground truth feature vectors occupy the majority of the same feature space as the automatically labeled key instances. There is, however, a small portion of the ground truth feature vectors which occupy an area outside the automatically identified key instances (upper left side). Further inspection of the feature vectors revealed that this area of the feature space related to ambient sounds and the ground truth feature vectors which occupy this area relate to ambient sounds which occurred between voice activity. As an aside, further analysis also revealed that top right area of the feature space, shown in Fig. 10, related to music sounds. The ambient sounds, which were recorded during the voice ground truth sound clips, occurred for very short periods of time. Due to their short duration, these sounds were missed, or overlooked,

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

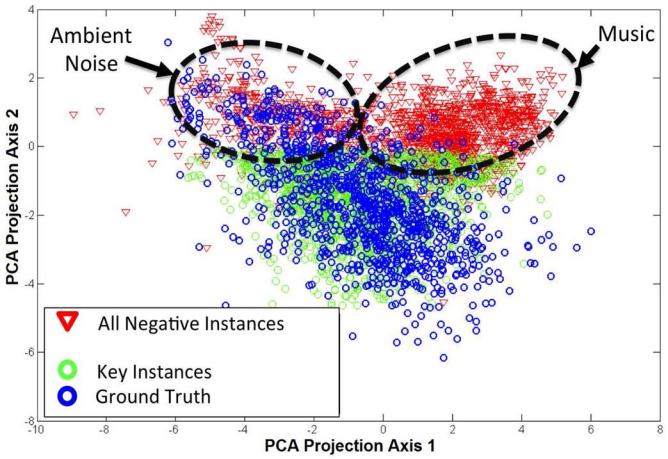

Fig. 10. DS 2: finding voice from general sounds—comparison with ground truth positives.

by the human labeler. As a result, the main segment of the sound clips were labeled as voice as opposed to a sequence of shorter voice segments with intermittent ambient sound segments. These errors in labeling illustrate the potential problems that can occur during manual sound labeling and demonstrate the level of accuracy that is required for manual labeling.

C. Quantitative Analysis

The protocol for the quantitative experiments is as follows.

1) For each sound, a set of key instances are identified using the DD and bag trimming techniques.
2) For each sound, l, in the set of sounds, $C = \{l_1, \ldots, l_c\}$, a RBF SVM, S_l, is trained using a set of positive and negative feature vectors. The positive training features correspond to the identified set of key instances while the negative training features corresponds to the set of all feature vectors in the set of negative bags for the current sound.
3) Each SVM is trained to obtain class probability estimates by computing pairwise class probabilities using Lin *et al.*'s [30] improved implementation of Platt's method [31].
4) The set of SVMs, $S = \{S_{l_1}, \ldots, S_{l_c}\}$, is then utilized to predict the class of individual test feature vector instances f by computing the set of probabilities $P = \{P(f = 1|S_{l_1}), \ldots, P(f = 1|S_{l_c})\}$, where $P(f = 1|S_{l_i})$ is the probability that f belongs to the positive class given S_{l_i}. If $P(f = 1|S_{l_i}) > 0.5$, then f is classified as belonging to class i. Since each probability is independent from other probably measures, it is possible that a feature can be assigned to more than one class. If this occurs, the feature is assigned to the class which produces the maximum probability.
5) Evaluation metrics are computed using the SVM predicted labels, computed from test set feature vector instances, and compared with the corresponding test set ground truth labels.

Tables I and II detail the confusion matrix and precision, recall and F-measure scores achieved by the classifiers.

TABLE I
DS 1 CLASSIFICATION RESULTS

	Classified As:					
	'Cutlery'	'Water'	Other	Precision	Recall	F-Measure
'Cutlery'	0.987	0.002	0.012	0.987	0.987	0.987
'Water'	0.008	0.951	0.041	0.973	0.951	0.962
Other	0.005	0.024	0.971	0.948	0.971	0.959
			Average	**0.969**	**0.969**	**0.969**

TABLE II
DS 2 CLASSIFICATION RESULTS

	Classified As:					
	'Voice'	'Ambient'	'Music'	Precision	Recall	F-Measure
'Voice'	0.926	0.06	0.012	0.841	0.926	0.881
'Ambient'	0.165	0.768	0.066	0.847	0.768	0.805
'Music'	0.009	0.078	0.918	0.921	0.918	0.919
			Average	**0.870**	**0.870**	**0.870**

TABLE III
LEARNING TECHNIQUES COMPARISON

Method	'Cutlery'	'Water'	**Avg-DS1**	'Voice'	'Ambient'	'Music'	**Avg-DS2**
Proposed	0.987	0.951	**0.969**	0.881	0.805	0.919	**0.870**
Supervised	0.989	0.973	**0.978**	0.906	0.824	0.916	**0.882**
miSVM[14] (Poly)	0.99	0.947	**0.968**	0.72	0.685	0.852	**0.752**
miSVM[14] (RBF)	0.99	0.957	**0.973**	0.878	0.734	0.793	**0.801**
DD[12]	0.979	0.932	**0.955**	0.546	0.642	0.559	**0.582**
EM-DD[13]	0.983	0.939	**0.961**	0.711	0.513	0.718	**0.647**

The results show that the SVMs achieved an overall F-measure of 0.969 and 0.870 when classifying feature vectors from DSs 1 and 2, respectively.

D. Comparative Analysis

In this section, an additional set of classification experiments are carried out on different learning techniques in order to compare our proposed technique with existing learning techniques. An important evaluation to carry out is to compare the performance of our weakly supervised technique with that of standard supervised learning. A supervised learning evaluation was carried out by manually labeling each sound clip in the training sets for DSs 1 and 2. Manual labeling was performed by carefully listening to each sound clip and noting the approximate start frame and end frame of each occurrence of the three sound categories. The same feature vectors, discussed in Section II-A, were extracted from the manually labeled sound clips. Feature vectors were grouped into sets corresponding to their manually assigned label. An SVM was trained on the feature vector sets using the same training procedure used in the weakly supervised method. The supervised SVMs were then evaluated using the test sets and the same test protocol discussed in Section II-A. Table III details the classification results achieved by the supervised SVMs trained on the manually labeled data. Results show the overall F-measure achieved by the supervised classifiers for DSs 1 and 2 was 0.978 and 0.882, respectively. In order to test for statistical significance of performance comparison differences, we perform two-tailed P value tests, utilizing variances calculated over eight folds (this protocol is also carried out for other T-tests in this section). T-tests carried out on DSs 1 and 2 show that the increase

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

in performance of the supervised training over our proposed method was not statistically significant ($P > 0.05$ for weak versus supervised comparison on DS 1 and $P > 0.05$ for weak versus supervised comparison on DS 2).

As discussed in Section I, there have been a number of different solutions proposed to the MIL problem such as DD and miSVM-based methods. We therefore evaluate these techniques, and variations of, in order to directly compare results achieved by our proposed method with commonly used methods. As previously discussed, classification using these MIL techniques have previously been based on classifying an entire bag as opposed to individual instances. In order to test these techniques with the more difficult task of classifying individual instances, test bags were constructed such that each bag contained only a single instance.

Table III details the F-measure results achieved by our proposed method, the supervised classifier and four alternative MIL techniques. We utilize the Waikato Environment for Knowledge Analysis data mining software [32] to evaluate the miSVM technique [14] using a polynomial kernel and a RBF kernel, the DD method [12], and the EM-DD method [13]. Results, for DS 1, show very good performance for all methods. In particular, the miSVM with RBF kernel, performed better than any of the other methods. However, when applied to a much more complex DS, DS 2, the performance of all methods decreased when compared to DS 1. It can be seen that supervised learning and our proposed method performs best for DS 2. The decrease in performance, when compared to DS 1, is 9.6% ($P < 0.05$ for T-test of DSs 1 and 2) for supervised learning and 9.9% ($P < 0.05$ for T-test of DSs 1 and 2) for our proposed technique. In comparison, there is a much larger decrease in performance of 17.2% ($P < 0.05$ for T-test on DSs 1 and 2) for miSVM (RBF), which is the next best performer.

While the overall classification comparison between the weakly supervised technique and the supervised classifier perform with similar accuracies, we perform an additional comparison on the weakly supervised classifier and the supervised classifier in order to understand how the classifiers perform under varying training set sizes. For this experiment we utilize the more challenging DS, DS 2. We train both the weakly supervised system and the supervised classifier using 100%, 80%, 60%, 40%, 20%, and 10% of the training set and evaluate the performance of the classifiers using the full test set (100% training = 26 sound clips). For each training set size, training was performed using multiple folds, where each fold is a random subset of sound clips from the training set. The number of folds used was inversely correlated to the training set percentage being tested. For example, four folds were used when testing on 80% training set size, while 16 folds were used when testing on 10% training set size. Evaluation of each fold is performed using the full test set. Overall results are calculated for each training set size by averaging the F-measure scores for each fold. Fig. 11 illustrates the results, where it can be seen that our proposed weakly supervised system performs with similar results as the supervised system for the discussed DS with training set sizes of 40% (approximately 10 sound clips) and above [T-test shows

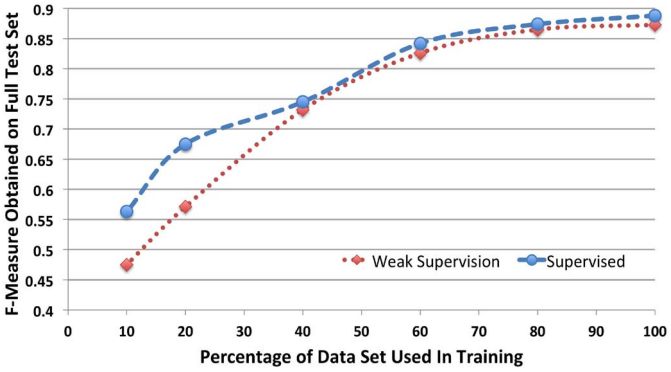

Fig. 11. DS 2 classification performance: weak supervision versus supervised training.

no statistically significant difference in results for training set sizes of 40% and above ($P > 0.05$), while there was a statistically significant difference ($P < 0.05$) for training set sizes of 20% and 10%].

E. Discussion

The experiments described above show that the approach of training a smartphone-based general sound classification framework using weakly supervised labels is a valid one. In particular, experiments validate our particular approach where a DD-based MIL framework can be utilized to identify a target sound and our bag trimming algorithm can be used to identify a set of key instances based on the target sound. Moreover, classification evaluations show that a classifier, trained on automatically extracted key instances from weakly supervised DSs, can accurately predict the correct sound class of individual sound feature vector instances. Experiments based on DS 1, which comprised of sounds heard in a kitchen and evaluated on identifying cutlery and water sounds, showed that a classifier was able to accurately predict whether a sound was a cutlery sound, water sound or neither. The classifiers were able to predict 27 min worth of kitchen sound features with an F-measure score of 0.969. Similarly, for DS 2, the classifiers were able to predict 211 min worth of general sound features with an F-measure score of 0.87. These scores represent promising results considering the very limited and noisy training information that was available in the training set.

An important aim of the experiments was to evaluate if the proposed weakly supervised learning techniques could be applied to very specific sounds as well as very general sounds. Both qualitative and quantitative experiments showed that, due to the hugely varied types of sounds that can occur in the general sound categories, the task of identifying target sounds and classifying general sound categories was more difficult than that of the specific sound categories. A visualization of this can be seen in Figs. 5 and 8 where the Heatmap for the specific sound categories, in DS 1 (Fig. 5), occupies a distinct area of the feature space. This is in contrast with the general sound categories, in DS 2 (Fig. 8), where the Heatmap occupies less a distinctive area of the feature space.

Comparative experiments also reveal that our proposed system performs with similar performances measures when compared to a supervised classification approach when tested

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

on the two DSs discussed in this paper. Results show that a classifiers trained on manually labeled sound clips, achieved F-measure scores of 0.978 and 0.882 for DSs 1 and 2, respectively. This was a nonstatistically significant increase in performance when compared to the F-measure scores achieved by our proposed weakly supervised system. This is an important result as it shows that weakly supervised classifiers, trained using data which has been labeled with very little effort and expertise, have the potential to perform as well as a supervised classifier which has been trained using detailed manually labeled data attained from an expert labeler using a very time consuming process. A further analysis of the weakly supervised and supervised classifiers also showed that our weakly supervised system performs as well as the supervised system using a training set comprising 10 sound clips or more. The performance of the weakly supervised system drops significantly, in comparison to the supervised classifier, for training sets with less than 10 sound clips. This can be attributed to the fact that the weakly supervised target sound identification process requires multiple positive and negative sound clips for each sound category to accurately identify a target sound.

Additional experiments also revealed that handling the less distinctive margin between the sound categories of DS 2 was a challenging task for other well known MIL solutions. Experiments showed that the F-measure performance for DS 1 for five different MIL classification techniques, including our proposed method, was good with results of over 0.95 for all techniques. However, a decrease in performance was reported for all methods when evaluated on DS 2. Interestingly, the decrease in performance of our proposed method was much less than that of the other MIL methods and was similar to the decrease in performance of the supervised learning method. We postulate that the difference in performance between the weakly supervised techniques is due to the KID approach employed in our proposed technique. DS 1 suits the classic MIL approach of identifying the most positive instances in a bag where each positive bag contains at least one instance which represents the target sound, and each of most positive instances are very similar to one and other. However, the classic MIL approach performs with less accurate results for DS 2 because it does not fit this classic MIL paradigm. Each positive bag contains at least one instance which represents the target sound, however, each of these instances could be very different from one and other. For example, each positive bag for the ambient category could have a unique sound representing ambient sound, such as car sounds and kitchen sounds. It can therefore be difficult for the MIL techniques to find a common sound among the positive bags. Through the use of KID and the bag trimming technique, our proposed method deals with this problem much better than the other discussed techniques. This is evident in the fact that our proposed technique achieves classification results comparable with supervised learning for both DSs whereas the other weakly supervised techniques achieve comparable results for DS 1 only. A target sound, which can be thought of as an initial best guess, is first identified using a DD-based approach. However, this target sound is used only as the basis for selecting the set of key instances. Key instances are selected based on similarity between the target sound and the negative set. Thus, if a candidate feature vector represents a unique sound which is not represented in other positive bags, it can still be considered as an element of the key instances if it is more dissimilar to the negative set than it is to the target sound. For example, car sounds and kitchen sounds represent very different sounds within the ambient sound category. However, if there is no car or kitchen sounds within the negative set, then it is probable that these sounds will be assigned to the positive class. Our proposed method, therefore offers more flexibility and is based around constraints of the DS rather than the technique itself.

IV. Conclusion

Sound sensing has the potential to sense human behavior in ways that motion sensors cannot. We highlight that a problem with current sound classification techniques is that a very specific and time consuming labeling process is required to segment and assign labels to sound features. This process makes the practical application of sound sensing in behavior sensing very restricted due to the varied types of sounds that could potentially be of interest in behavior analysis. Exploring techniques to carry out sound classification without the need for this detailed and time consuming labeling process is, therefore an important research goal.

In this paper, as an alternative to detailed sound labeling and supervised training, we explore the use of weak supervision in sound labeling where sound clips are labeled with very limited information. More specifically, we take a KID approach to MIL where the goal is to not only train a system using weak supervision, but also to classify individual instances rather than entire bags. We propose a bag trimming technique, an extension to DD, in order to carry out KID by utilizing the target sound to find key instances within the sound clips. The automatically created training sets are then used to train an SVM-based classifier. Experiments, based on two DSs, showed that our approach is a valid one with classifiers, trained using automatically identified training sets, shown to accurately classify sounds from DSs 1 and 2 with an average F-measure of 0.959 and 0.87, respectively. Furthermore, results also show that classifiers trained using our weakly supervised techniques perform with results comparable to results achieved by classifiers which were trained using fully supervised DSs.

To our knowledge, no other work has explored the use of a smartphone to carry out general sound classification without using explicit supervised training at some stage of the training process. This paper represents a positive step away from time consuming expert-based recording, labeling and training of sound classifiers and a step toward a flexible system which could allow nonexpert users to record, label, and train a sound classifier with their own particular application in mind.

References

[1] N. D. Lane *et al.*, "A survey of mobile phone sensing," *IEEE Commun. Mag.*, vol. 48, no. 9, pp. 140–150, Sep. 2010.

[2] D. Kelly, B. Smyth, and B. Caulfield, "Uncovering measurements of social and demographic behavior from smartphone location data," *IEEE Trans. Human–Mach. Syst.*, vol. 43, no. 2, pp. 188–198, Mar. 2013.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

[3] J. W. H. Kocks *et al.*, "Health status measurement in COPD: The minimal clinically important difference of the clinical COPD questionnaire," *Respir. Res.*, vol. 7, p. 62, Apr. 2006.
[4] P. K. Ghosh, A. Tsiartas, and S. Narayanan, "Robust voice activity detection using long-term signal variability," *IEEE Audio, Speech, Language Process.*, vol. 19, no. 3, pp. 600–613, Mar. 2011.
[5] Z. Fu, G. Lu, K. M. Ting, and D. Zhang, "A survey of audio-based music classification and annotation," *IEEE Trans. Multimedia*, vol. 13, no. 2, pp. 303–319, Apr. 2011.
[6] M. Morvidone, B. L. Sturm, and L. Daudet, "Incorporating scale information with cepstral features: Experiments on musical instrument recognition," *Pattern Recogn. Lett.*, vol. 31, no. 12, pp. 1489–1497, Sep. 2010.
[7] M. Chetouani, M. Faundez-Zanuy, B. Gas, and J. L. Zarader, "Investigation on LP-residual representations for speaker identification," *Pattern Recogn.*, vol. 42, no. 3, pp. 487–494, Mar. 2009.
[8] C.-C. Lin, S.-H. Chen, T.-K. Truong, and Y. Chang, "Audio classification and categorization based on wavelets and support vector machine," *IEEE Speech Audio Process.*, vol. 13, no. 5, pp. 644–651, Sep. 2005.
[9] R. Mogi and H. Kasai, "Noise-robust environmental sound classification method based on combination of ICA and MP features," *Artif. Intell. Res.*, vol. 2, no. 1, pp. 107–121, 2012.
[10] Z. Zhang and B. Schuller, "Semi-supervised learning helps in sound event classification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Kyoto, Japan, 2012, pp. 333–336.
[11] H. Lu, W. Pan, N. D. Lane, and T. Choudhury, "SoundSense: Scalable sound sensing for people-centric applications on mobile phones," in *Proc. 7th Int. Conf. Mobile Syst. Appl. Serv.*, Wroclaw, Poland, 2009, pp. 165–178.
[12] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *Proc. Adv. Neural Inf. Process. Syst.*, Denver, CO, USA, 1998, pp. 570–576.
[13] Q. Zhang and S. A. Goldman, "EM-DD: An improved multiple-instance learning technique," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2001, pp. 1073–1080.
[14] S. Andrews and I. Tsochantaridis, "Support vector machines for multiple-instance learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 15. Whistler, BC V0N, Canada, 2003, pp. 577–584.
[15] T. Dietterich, "Solving the multiple instance problem with axis-parallel rectangles," *Artif. Intell.*, vol. 89, nos. 1–2, pp. 31–71, Jan. 1997.
[16] O. Maron and A. L. Ratan, "Multiple instance learning for natural scene classification," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Madison, WI, USA, 1998, pp. 341–349.
[17] D. Kelly, J. McDonald, and C. Markham, "Weakly supervised training of a sign language recognition system using multiple instance learning density matrices," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 41, no. 2, pp. 526–541, Apr. 2011.
[18] M. Stikic, D. Larlus, S. Ebert, and B. Schiele, "Weakly supervised recognition of daily life activities with wearable sensors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2521–2537, Dec. 2011.
[19] G. Liu, J. Wu, and Z. H. Zhou, "Key instance detection in multi-instance learning," in *Proc. JMLR Workshop Asian Conf. Mach. Learn.*, vol. 25. Singapore, 2012, pp. 253–268.
[20] J. D. Deng, C. Simmermacher, and S. Cranefield, "A study on feature analysis for musical instrument classification," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 38, no. 2, pp. 429–438, Apr. 2008.
[21] K. Umapathy, S. Krishnan, and R. K. Rao, "Audio signal feature extraction and classification using local discriminant bases," *IEEE Audio, Speech, Language Process.*, vol. 15, no. 4, pp. 1236–1246, May 2007.
[22] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Munich, Germany, 1997, pp. 1331–1334.
[23] D. Li, I. K. Sethi, N. Dimitrova, and T. McGee, "Classification of general audio data for content-based retrieval," *Pattern Recogn. Lett.*, vol. 22, no. 5, pp. 533–544, Apr. 2001.
[24] J. Saunders, "Real-time discrimination of broadcast speech/music," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Atlanta, GA, USA, May 1996, pp. 993–996.
[25] C. G. Broyden, J. E. Dennis, Jr., and J. J. Moré, "On the local and superlinear convergence of quasi-Newton methods," *IMA J. Appl. Math.*, vol. 12, no. 3, pp. 223–245, 1973.
[26] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *J. Intell. Inf. Syst.*, vol. 17, nos. 2–3, pp. 107–145, 2001.
[27] F. Kovács, C. Legány, and A. Babos, "Cluster validity measurement techniques," in *Proc. 5th WSEAS Int. Conf. Artif. Intell. Knowl. Eng. Data Bases*, Madrid, Spain, 2006, pp. 388–393.
[28] J. T. Geiger, B. Schuller, and G. Rigoll, "Large-scale audio feature extraction and SVM for acoustic scene classification," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, USA, 2013, pp. 1–4.
[29] G. Guo and S. Z. Li, "Content-based audio classification and retrieval by support vector machines," *IEEE Trans. Neural Netw.*, vol. 14, no. 1, pp. 209–215, Jan. 2003.
[30] H.-T. Lin, C.-J. Lin, and R. C. Weng, "A note on Platt's probabilistic outputs for support vector machines," *Mach. Learn.*, vol. 68, no. 3, pp. 267–276, Aug. 2007.
[31] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*. Cambridge, MA, USA: MIT Press, 1999, pp. 61–74.
[32] M. Hall *et al.*, "The WEKA data mining software," *ACM SIGKDD Explor. Newslett.*, vol. 11, no. 1, p. 10, Nov. 2009.

Daniel Kelly received the Ph.D. degree from the National University of Ireland, Maynooth, Maynooth, Ireland, in 2010.

He is currently a Lecturer with the School of Computing and Intelligent System, Ulster University, Coleraine, U.K. His current research interests include human behavior analysis technology in areas of health and well-being, motion analysis, machine learning, and activity recognition.

Dr. Kelly is a member of Computer Science Research Institute, Ulster University.

Brian Caulfield received the Ph.D. degree from University College Dublin, Dublin, Ireland, in 2002.

He is a Director with the Insight Center for Data Analytics, University College Dublin, Dublin, Ireland. His current research interests include clinical physiotherapy, wide-area based around using technology in the assessment and enhancement of human function in health and sport.