Overlapping Sound Event Recognition using Local Spectrogram Features with the Generalised Hough Transform

Jonathan Dennis\textsuperscript{1,2}, Huy Dat Tran\textsuperscript{1}, Eng Siong Chng\textsuperscript{2}

\textsuperscript{1}Institute for Infocomm Research, A*STAR, 1 Fusionopolis Way, Singapore 138632
\textsuperscript{2}School of Computer Engineering, Nanyang Technological University, Singapore

\{stujwd,hdtran\}@i2r.a-star.edu.sg, ASESChng@ntu.edu.sg

Abstract

We present a novel approach for recognition of overlapping sound events based on the Generalised Hough Transform (GHT) – a technique commonly used for object recognition in the domain of image processing. Unlike our previous work on image-based sound event classification, where we focussed on global image features, here we extract local features from detected interest-points in the spectrogram. These form a robust representation of the local region, and when the information from all interest-points in the spectrogram are combined using the GHT, we can form hypotheses for the location of one or more overlapping sound events in the image. Our experiments show promising results, and demonstrate the ability of our approach to recognise overlapping sounds.

Index Terms: overlapping, sound events, recognition

1. Introduction

The spectrogram of an audio signal provides valuable visual information that, for example, has been historically used to analyse the phonetic structure of speech. A trained expert in “spectrogram reading” \cite{1} is able to pick out the important structures and use these to recognise the underlying speech. Despite this, visual-based techniques for automatic classification of speech have not been heavily researched, in part due to the complicated lexical structure. Sound events, on the other hand are typically more sparse and distinct, making the visual information more tractable for automatic classification.

In our previous works \cite{2,3}, we developed an image feature framework for sound event classification. These approaches perform well, particularly in noisy conditions, which vindicates the use of such image processing techniques for sound classification. Related works on the topic of image-based audio classification have focussed on both environmental sounds and music genre classification \cite{4,5,6}, as music spectrograms have similarities with texture images. Often these approaches just extend frame-based techniques, commonly used in speech processing, or combine block-based techniques with Support Vector Machines (SVM) to classify the whole spectrogram.

In this paper, we address the problem of recognising overlapping sound events, inspired by the visual information. According to the LogMax principle \cite{7}, when two sounds are added together each cell of the spectrogram belongs to one of the sounds, as the log-magnitude of the stronger sound will always dominate. Hence, we can still visually see overlapping sounds in the spectrogram. However, overlapping conditions is one of the most challenging in audio classification, where performance degrades greatly. While present research focuses more on sound source separation, few works have considered the simultaneous recognition of overlapped sound signals. A related area is Computational Auditory Scene Analysis (CASA) \cite{8}, where the aim is to segment the spectrogram to form an Ideal Binary Mask (IBM). However, this segmentation is prone to errors that affect the subsequent classification. Among the available literature on speech separation, the Factorial Hidden Markov Model (FHMM) technique \cite{7} is the most popular. Recently, a method based on cepstral processing and subband distributions has been applied for overlapping audio classification \cite{9}, however this method is applicable only for certain sound classes, not the general case of sound events.

Here, we propose a novel approach that combines the detection and classification of overlapping sound events into a single framework using local spectrogram features with the Generalised Hough Transform (GHT) \cite{10}. The fundamental idea of the GHT is to map consistent interpretations of a shape into local maxima in the Hough space, which are then separable. In object recognition, the GHT is realised through a sophisticated voting procedure \cite{11}. In this work, we develop a method that first extracts local features surrounding interest-points in the spectrogram. These features, and their temporal locations, are used to create a model which is used for detecting each sound class. The GHT is used as a scoring mechanism, enabling us to detect multiple overlapping sound events in a single clip. Our experiments show promising results, and we discuss how further could incorporate robustness to noise and distortion.

The paper is organised as follows. Section 2 first introduces the idea behind our approach. Section 3 then details the recognition algorithm for the proposed method. Section 4 describes our experiments and the results obtained. Finally, Section 5 concludes this work.
2. Overview of our Proposed Approach

2.1. The Hough Transform

The principle of the Hough Transform is to detect shapes in an image, though a voting procedure that is robust to imperfections in the image. It was first used to detect parametrised lines and curves, and later expanded to cover arbitrary shapes through the Generalised Hough Transform (GHT) [10].

As a simple example, consider the image shown on the left of Fig. 1, which contains two lines against a noisy background. Here, each point, with coordinates \((x, y)\), is considered an interest-point, and therefore casts votes into the Hough accumulator, \(H(r, \theta)\) (shown on the right), where \(r\) is the perpendicular distance from the origin and \(\theta\) the angle from the horizontal axis. The points simply vote for all possible straight lines that they could belong to (all possible rotations: \(-90 < \theta < 90\)), according to the polar equation of a straight line, as follows:

\[
H(r, \theta) \equiv 1 \rightarrow r = x \cos \theta + y \sin \theta \quad \forall \theta
\]  

Local maxima in the Hough accumulator correspond to the combined evidence for a line existing with a given \((r, \theta)\) from many individual points lying on the line. Hence the two lines can be detected and separated easily, as shown in Fig. 1.

The extension of the approach to the GHT allows for arbitrary shape detection, but maintaining the principle of independent voting and local maxima separation in the Hough accumulator. The process now requires the detection of interest-points, local feature extraction and matching to a codebook, where the codebook contains a weight function, \(W\), in place of the parametrised equation, to perform voting. In addition, as the codebook entries are learnt separately for each class of shape, the codebook entries cast votes into a separate Hough accumulators for each class. This has the powerful property of enabling the detection of overlapping shapes, simply by finding the local maxima in the Hough accumulator belonging to each class, and performing verification.

2.2. Approach for Overlapping Sound Event Recognition

In this paper, we bring this principle to the task of overlapping sound detection using the visual information in the spectrogram. Our approach, based on the GHT, combines the detection, classification and segmentation of the sound event image into a single process that is robust to occlusion of overlapping events.

When used for object recognition, the Hough accumulator typically has four dimensions – the \(x, y\) position in the image, and the object scale and rotation. However, the spectrogram has a number of notable differences from typical images, but most importantly the frequency dimension of the spectrogram is fixed. Hence we assume that the frequency position, scale, and rotation cannot change, and focus on the time dimension of the spectrogram. Therefore we use a one dimension Hough accumulator of the time information of the sound events, and combine the information from each subband of the spectrogram together. This reduces the problem of finding sound events to the detection of local maxima in the class-specific Hough accumulator, which correspond to strong evidence for the sound.

Note that we considered using the magnitude dimension of the spectrogram as a second parameter for the Hough accumulator. However, to ensure the method is independent of the relative magnitude of the sound class between training and testing, we chose not to, and instead normalise the spectral region around each interest-point to make the local features independent of the absolute magnitude.

3. Proposed Overlapping Sound Recognition System

In this section we describe our approach that brings the GHT recognition framework of [11] to the domain of sound event recognition, using the following steps:

- **Interest-point detection and local feature extraction**
- **Training:** Codebook clustering; extraction of codebook temporal distribution functions.
- **Testing:** Matching codebook entries; Generalised Hough Transform voting; Verification.

3.1. Interest-Point Detection

We start from a short-time Fourier transform (STFT) spectrogram representation of the sound event, \(S(f, t)\), where \(f\) is the frequency bin (using 129 frequency bins), \(t\) is the time frame, and the values are the cube-root of the spectral power. We prefer this over log-power compression, as it does not amplify small changes in the noise as the power tends towards zero.

The first step is to detect a set of “interest-points”, \(K\), that represent interesting structures in the image (sometimes referred to as “keypoints” in the literature). To do this, we consider a “plus” shaped region, \(R(f, t)\), of a 12-pixel diameter surrounding each point in the image, as shown in Fig. 2:

\[
R(f, t) = \left[ S(f - 5 : f + 6, t) ; S(f, t - 5 : t + 6) \right]
\]

where “:” specifies a range of pixels. We choose the “plus” shape, as it gives a glimpse of the time-frequency information at that point, without including the full 2D region, which may vary significantly in the case of overlapped sounds.

We then generate a local estimate of the noise, \(z(f, t)\), surrounding the point as the minimum of the mean of \(R(f, t)\) separately across time and frequency:

\[
z(f, t) = \min(\mu(R(\forall f[t])), \mu(R(\forall t[f])))
\]

In this way, we are assuming the noise to be stationary over the local region surrounding the point, which is reasonable for a small enough local region. It also as the advantage that it bypasses the need for a global noise estimate for the spectrogram. We can now select all interest-points, \(k_n = [f, t]\), where:

\[
S(k_n) - z(k_n) > \beta,
\]

a threshold that allows us to extract only the sparse peaks of the signal and reject less significant maxima. Here we used threshold \(\beta = 5 dB\). For each image we therefore have:

\[
K = [k_1, k_2, k_3, \ldots, k_N]
\]

where \(N\) is the number of interest-points found.

3.2. Feature Extraction

From each \(k_n\), we extract a feature, \(y_n\), from the surrounding spectral region, to give a set of feature vectors, \(Y\):

\[
Y = [y_1, y_2, y_3, \ldots, y_N]
\]

Each \(y_n = \phi(R(k_n))\) – a function of the surrounding plus-shaped spectral region, \(R(f, t)\). Here, we simply use the raw spectral values as the feature, normalised by the magnitude at the interest-point:

\[
y_n = \frac{R(f, t)}{S(f, t)}
\]

The normalisation ensures that during training we can build compact clusters, and that during testing the matching process is independent of any changes in magnitude of the sound event.
3.3. Training the Recognition System

For a sound class \( X \), the aim of the training process is to learn:

1. A set of class-specific codebooks of local feature clusters, \( C_{X,f} \), across each frequency subband, \( f \).
2. The temporal distribution function (TDF) of each codebook, \( W_C(l) \), where \( l = t - t_{ON} \), which specifies the coordinates of codebook feature occurrences, \( t \), relative to the sound’s onset time, \( t_{ON} \).

An overview of this approach is shown diagrammatically in Fig. 2, and described in detail below.

3.3.1. Codebook Clustering

To create the codebook, \( C_{X,f} \), we perform clustering of the extracted features, \( X \), belonging to each subband, \( f \) in turn. This makes the assumption that features can only be matched within a subband, and if they occur in a different subband during testing, then they must belong to a different sound.

Starting with each feature from subband \( Y_{X,f} \) as a separate cluster, agglomerative clustering is performed. Here, the two most similar clusters \( c_{j,1} \) and \( c_{j,2} \) are merged as long as the average Euclidean distance between features is below a certain threshold \( \lambda \). Here we use \( \lambda = 1 \) as our cutoff to give compact clusters. This clustering scheme guarantees that only those features are grouped which represent similar spectrogram regions.

From each resulting cluster, we compute the cluster mean and variance, and store it in the codebook, \( C_{X,f} \):

\[
C_{X,f}(m) = \mu(y_n \in c_{j,m}), \quad \sigma^2(y_n \in c_{j,m})
\]

(8)

where \( m = 1,2,\ldots,M \) is the number of codebook clusters in each subband, \( \mu(.) \) is the mean, and \( \sigma^2(.) \) is the variance.

3.3.2. Codebook Temporal Distribution Function

In this step, we find the TDF of each subband cluster, \( W_C(l) \), to build a function that represents the spatial information of the sound over time. We use the position \( l = t - t_{ON} \), relative to the sound onset, \( t_{ON} \), as the spatial function, such that training is not based on the absolute positions of the interest-points, \( k[f,t] \).

As recommended in [11], we first find the probability density function (PDF), \( P_{C}(l) \), of the temporal occurrences, and then normalise it by the peak of the distribution, as follows:

\[
P_{C}(l) = PDF(k[f,t - t_{ON}] \subset c_{j,m})
\]

(9)

\[
W_C(l) = \frac{P_{C}(l)}{\max(P_C(l))}
\]

(10)

Therefore the function, \( W_C(l) \) represents the probability of finding codebook cluster \( C \) at position \( l \) relative to the sound onset. We model this using Gaussian Mixture Models (GMMs), which we find better generalises to testing data better than directly using the histogram. We use the toolkit from [12], which automatically determines the number of Gaussian mixtures.

3.4. Recognition of Overlapping Sounds

Given the class-specific codebook clusters, \( C_{X,f} \), and their associated temporal distribution function, \( W_C(l) \), we can perform recognition of overlapping sound events. We start by extracting interest-points, \( k_{n} \), and features, \( y_{n} \), and then match the features onto the codebook, and use \( W_C(l) \) as a voting function to perform a GHT. Note that we cannot evaluate \( W_C(l) \) directly, as the sound class and onset are not known in advance.

The method is shown diagrammatically in Fig. 3, and described in detail in the sections below.

3.4.1. Matching Codebook Entries

Each interest-point in the testing spectrogram, \( k_{n}[f,t] \), activates all training clusters, \( C_{X,f} \), that match to within a threshold. Here, we use a missing feature approach for this step using the local noise estimate from (3), \( z_{b,w} \). All \( y_{n} < z_{b,w} \) are marked as unreliable, and we then use bounded marginalisation [13], using the mean and variance calculated in (8), to estimate the cluster log-likelihoods, \( L_{C_{X,f}} \). All clusters with \( L_{C_{X,f}} > \alpha \), where we used \( \alpha = -50 \) as our threshold, are considered to be matched, and are used in the Hough voting below.

3.4.2. Generalised Hough Transform voting

All matching codebooks of every interest-point in the spectrogram cast votes into the class-specific Hough accumulator, \( H_X(t) \), for hypothesised sound onsets using their temporal distribution function, \( W_C(l) \) as calculated in (10). Independently for each class, these are summed together across all subbands to give a score for each class as a function of time:

\[
H_X(t) = \sum_{b \subset K} \sum_{W_{matched}} W_C(l + b_w)
\]

(11)

Peaks in this GHT accumulator, \( t_{ON,X} \), correspond to strong evidence for the existence of a sound event.
3.4.3. Verification

Upon finding all peaks across all classes, \( t_{ON} \), we perform the following verification steps. Starting from the hypothesis that explains the most interest-points, we check that \( H_X(t_{ON}) > \gamma \). Here, \( \gamma = \frac{1}{2}(H_X(t_{ON})) \) is a threshold based on training images, \( T \), proportional to the mean onset score. Next, we check that the magnitude of matched interest-points is consistent with the training. After removing any extreme points, if \( H_X(t_{ON}) \) is still greater than the threshold, we accept the hypothesis, and remove all hypothesised points from future matches.

This is repeated until no valid hypotheses are found. Therefore it is possible to detect any combination of the trained sound classes in the spectrogram, and their onset position, \( t_{ON} \).

4. Experiments

4.1. Experimental Methods

For our experiments, we generate a database of overlapping sound events using 5 classes: Horn, Bells, Bottle, Phone and Whistle [14]. We then combine these to give a further 10 overlapped events, each consisting of 2 single events, overlapped by at least 50%. This gives a total of 15 different clip types. For training, we use 20, randomly selected isolated sound event clips. For testing, we generate 50 samples for each of the 15 types, using isolated samples excluded from the training set.

As there are no state-of-the-art methods for overlapping audio event recognition, we compare our proposed method against a method based on the LogMax principle. As in [9], we use a LogMax GMM in place of the full FHMM approach from [7]. In this case, the PDF of overlapping classes can be decomposed from the PDFs and cumulative density functions (CDFs) of the isolated class models. For overlapped class \( Z = X + Y \), where \( X \) and \( Y \) are two clean classes, the PDF, \( p_Z \), of class \( Z \) is:

\[
p_Z(\alpha) = p_X(\alpha)c_X(\alpha) + p_Y(\alpha)c_Y(\alpha)
\]

where \( c_X \) is the CDF and \( \alpha \) is the feature to evaluate. Here we model the PDFs using a 6-component GMM.

As evaluation measure, we calculate the recognition accuracy (TP) and false alarm (FA) of each class over 5 runs of the experiment. TP is calculated as the ratio of correct detections to the number of clips containing occurrences of that class. Analogously, FA is the ratio of incorrect detections to the number of clips not containing that class.

4.2. Results and Discussion

The results are given in Table 1. It can be seen that the average TP of our proposed method is 94.9%, a 10.9% absolute improvement over the baseline. In addition, the low false alarm rate of our method is significant, particularly when considering that we in no way limit our algorithm to the recognition of just two overlapped sound classes, as with the LogMax method.

Our proposed method also has the advantage that it should prove to be noise robust, as we only require a local estimate of the noise, rather than a global mask. In addition, distortion effects should not significantly change the normalised feature in a local area of the spectrogram. However, verification will need to be improved to take account of the non-uniform change in the observed magnitude of the sound across frequency. Finally, despite the seeming increase in complexity of our proposed method over the baseline, the recognition algorithm is actually very simple, and the voting of each subband can be implemented in parallel giving real-time recognition.

<table>
<thead>
<tr>
<th>Proposed Method</th>
<th>LogMax GMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP</td>
<td>FA</td>
</tr>
<tr>
<td>Horn</td>
<td>99.4 ± 0.5</td>
</tr>
<tr>
<td>Bells</td>
<td>93.8 ± 2.7</td>
</tr>
<tr>
<td>Bottle</td>
<td>91.3 ± 1.2</td>
</tr>
<tr>
<td>Phone</td>
<td>100 ± 0</td>
</tr>
<tr>
<td>Whistle</td>
<td>89.8 ± 2.7</td>
</tr>
<tr>
<td>Average</td>
<td>94.9 ± 1.8</td>
</tr>
</tbody>
</table>

Table 1: TP/FA (%) averaged over the 15 overlap combinations.

5. Conclusion

In this paper, we propose a novel method for the simultaneous recognition of an event, and use interest-point detection, local feature extraction, and Generalised Hough Transform voting to transform the information to a domain where the sound events are easily separable. Our preliminary experiments show the potential of the approach, and in future we will expand this to noisy and reverberant conditions.

6. References