MIREX 2019: USING IMAGE RECOGNITION TECHNIQUES IN MULTIPLE FUNDAMENTAL FREQUENCY ESTIMATION & TRACKING

Anton Runov
anton.runov.oca@gmail.com

ABSTRACT

This submission presents a multiple fundamental frequency tracking system that utilizes image recognition techniques to identify structures in short-time Fourier transform (STFT) output. An extended version of Haar-like rectangular features relevant to both amplitude and phase parts of the STFT is proposed. AdaBoost algorithm is used to train strong classifiers by selecting the most efficient features from a randomly generated initial set.

1. INTRODUCTION

The fundamental frequency estimation and music recognition in general are usually tied to the speech recognition field. Even though there are obvious similarities between the tasks, the polyphony is what makes them different. To extract all the information from a polyphonic signal it is not sufficient to identify time regions with different sets of features. As the regions intersect, the number of possible combinations makes the separation virtually impossible. On the other hand two-dimensional time-frequency representation does provide enough information to distinguish different voices in a polyphonic signal. While spectrogram visualization has become a standard approach in the music analysis, it seems like there were not very many attempts to use image recognition algorithms and techniques for the task.

This work is an attempt to adopt the well-known Haar-like features originally developed for face recognition problem [1] to the polyphonic music recognition task. The STFT output is considered as an input image in which specific structures should be detected. Despite some similarity there are also important differences between the STFT and typical image recognition objects. First of all the musical 'objects' are not located in a single area but spread across frequencies in a form of sequences of lines or ridges. Another difference is that the significant part of the information in the STFT is in its phase part. The latter is often ignored in music analysis but it does contain an important information. To take into account these factors a new type of features called qH features was introduced.

Another challenge which is very common for MIR tasks was lack of relevant datasets. Training reliable classifiers requires large (at least 10^5–10^6 points) training sets. As it was repeatedly stated before, due to a number of reasons there are insufficient publicly available datasets suitable for MIR tasks [2, 3]. To overcome this issue a technique of synthesizing random loosely-music-like samples was developed. The excellent open dataset introduced recently by B. Li, X. Liu et al. [4] was also used at the final training stage.

2. QH FEATURES

Haar-like features introduced by Viola and Jones in their rapid object detection framework [1] are very simple combinations of the sums of the pixels within rectangular regions. The main advantage of this approach is that the calculation can be performed very fast for any rectangle size. This allows building strong classifiers by selecting most efficient features among hundreds of thousands and more possible locations. The approach seems to be meaningful for STFT as well. Combining rectangular areas with different proportions in the time-frequency space may be interpreted as balancing time-frequency resolutions.

The Haar-like feature set consists of a fixed number of shapes. On the contrary the proposed qH feature set includes virtually all possible rectangle configurations. There are two basic feature types

1. A single rectangle $[t_1, t_2, f_1, f_2]$ with non-zero area (i.e. $t_1 < t_2$ and $f_1 < f_2$). The value of the feature is the average of the sample values within the rectangle.

2. Two nested rectangles $[t_1, t_2, f_1, f_2], [t_3, t_4, f_3, f_4]$ where both rectangles have non-zero area, the second rectangle is located completely inside the first one and is not equal to the first one. The value of this feature is the difference between the average sample values of the two rectangles.

In both cases any proportions of rectangles are permitted (except zero area case) and any relative rectangle's positions are possible as far as the constraints are met.

To take into account harmonic partials of musical signals an additional integer field J_f (a harmonic index) was added to the feature vector. A value of a feature at a point (t, f) is actually calculated at frequency fJ_f. This allows to associate information from different harmonic partials with their fundamental frequencies.

Fast feature calculation is based on the integral image technique described in [1]. To handle the phase information from the STFT the qH features are calculated across 3 different integral images. The first one is computed from the absolute values of the STFT. The logarithm of the values is used and a normalization procedure is applied to limit the influence of frequency corrections in the recordings. The next two images are constructed from the phase part of the STFT. They use an efficient phase difference between consequent samples in the time and frequency dimensions correspondingly.

The resulting qH feature vector has 11 elements (8 rectangle's coordinates, a harmonic index, an image index and a specific frequency scaling parameter aimed to decrease the difference between frequency regions). During the training procedure qH vectors were constructed randomly within the appropriate constraints. Then the AdaBoost algorithm was used the same way as in [1] to find the most efficient vector at each iteration and to update the corresponding weights and coefficients.

The training procedure consisted of three steps using different training sample types at each step. The first two steps used synthesized samples vaguely resembling musical signals. The final one utilized the University of Rochester Multi-modal Music Performance (URMP) Dataset [4]. At each step more than 10^6 training points were used and the corresponding strong classifiers were calculated for 3 different frequency bands (35–130 Hz, 130–512 Hz and 512–2000 Hz).

3. CONCLUSIONS

The results of this work should be considered as very preliminary. The vast majority of many algorithm and training parameters has not been assessed yet. Despite that the algorithm has demonstrated a reasonable accuracy.

One of the main advantages of Viola-Jones approach in the image recognition is very high performance that is achieved due to using classifier cascades [1]. Even though it seems highly possible to apply the similar approach to the qH classifiers, this task is outside of the scope of the presented work. The submitted algorithm does not implement cascading technique and therefore does not demonstrate significant performance. On an efficient 8 core CPU it can run up to 2 times faster than real-time but on most typical systems it runs much slower. Implementing cascading classifier approach should increase the performance dramatically though.

Another important issue is training samples. While the synthesized samples used in the training have demonstrated reasonable training quality, achieving the perfect accuracy will require much better samples. Both in terms of the size of sample sets and in terms of their relevance. In author's opinion, using synthesized samples for high quality training should still be considered as a valuable option. It solves at least two major problems, the lack of available musical data due to copyright restrictions and the inaccuracy of manually annotated samples. But producing high quality synthesized samples requires significant improvements of the technique and further research of the factors affecting the training efficiency.

4. REFERENCES

[1] Paul Viola and Michael J. Jones: "Rapid Object Detection using a Boosted Cascade of Simple Features," *IEEE CVPR*, 2001.

[2] Downie, J. Stephen: "The Music Information Retrieval Evaluation Exchange (2005-2007): A window into music information retrieval research," *Acoustical Science and Technology*, 29 (4), pp. 247–255, 2008.

[3] M. Schedl, E. Gomez and J. Urbano: "Music Information Retrieval: Recent Developments and Applications," it Foundations and Trends in Information Retrieval, Vol. 8, pp. 127–261, 2014.

[4] B. Li, X. Liu, K. Dinesh, Z. Duan and G. Sharma: "Creating a Multitrack Classical Music Performance Dataset for Multimodal Music Analysis: Challenges, Insights, and Applications," *IEEE Transactions on Multimedia*, Vol. 21, No. 2, pp. 522–535, Feb. 2019.