5R-9

Audio-visual musical instrument recognition

Angelica Lim[†] **Keisuke Nakamura**[‡] **Kazuhiro Nakadai**[‡] **Tetsuya Ogata**[†] **Hiroshi G. Okuno**[†]
Graduate School of Informatics, Kyoto University[†] Honda Research Institute Japan Co., Ltd., Japan[‡]

1 Introduction

In 2008, a humanoid robot developed by Honda conducted a symphonic orchestra in front of a live audience. It perfectly imitated the pre-recorded actions of the orchestra's human conductor. Despite its realistic movements, it was suggested that the robot could not listen to nor interact with the orchestra as a true conductor would. For a robot to truly direct an orchestra, it would need to hear, distinguish, and respond to the sounds of different instruments. As a first step to making a musically trained robot, we implement monophonic instrument recognition on Hearbo, developed by HRI-JP for audio-based human-robot interaction.

Until now, musical instrument recognition has been limited to audio recording analysis. In the field of solo musical instrument recognition, acoustic features such as MFCCs and LPCC's [1] have been widely studied. Classifiers like k-nearest neighbors [1], SVM and GMM [2] to classify these features have also been examined. Using a priori musical knowledge, Martin's classification system in [3] used a hierarchy of musical instrument classes, such as grouping string instruments like guitar and violin together. Both his and Klapuri and Eronen's [4] work showed better results classifying at the instrument family level rather than at the specific instrument. Brown [5] investigated the use of features such as attack and decay to distinguish between four similar woodwind instruments. Indeed, a common problem comes down to distinguishing instruments of the same family. So far, no one has yet exploited the visual differences between instruments to overcome this problem.

2 Multi-modal instrument recognition

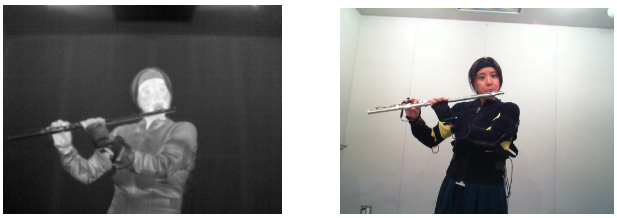
In our approach, we use a robot's thermo camera images to improve audio-only recognition of musicians playing 12 different instruments. As explained in the following sections, we extract features for both audio and video and train Gaussian Mixture Model classifiers over each separately. We then fuse the two scores using a weighted linear combination scheme.

2.1 Audio feature Extraction

We use the HARK [6] implementation of mel-scale log spectrum (MSLS) as our acoustic feature. It is a vector of 27 features: 13 values representing the spectrum in mel-scale, 3 delta values computed using a linear regression on a sequence of five consecutive frames, as well as 1 value representing delta logarithmic power. We use a frame length of 512 on a one-channel audio signal sampled at 16000Hz.

2.2 Video feature extraction

We choose the Histogram of Oriented Gradients (HOG) [7] as the feature for video recognition. As the name indicates, the feature vector contains a histogram of the image's edge gradients, binned over orientations. The idea is that this feature vector would represent the shape of the human body pose while playing their instrument. Although body pose estimation methods exist, they often assume the person is not holding any object in their hands, which cannot be assumed for our instrument playing participants.



(a) (b)

Figure 1: (a) Thermo input image for flute and (b) RGB image for reference.

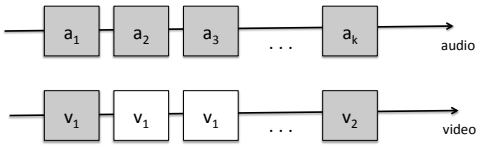


Figure 2: Upsampling of video to match audio sampling rate. The thermo frame is repeated at the same rate as audio until a new thermo image is received.

In our experiments, thermo camera images (see Fig. 1(a)) of 320 × 240 pixels were used, producing a feature vector of 128 values at a rate of 5 Hz. As a pre-processing step to remove the effect of small changes in background or clothing, we perform a Gaussian smoothing operation. Since the audio signal is sampled at a higher rate than thermo video, we perform a simple up-sampling step as shown in Fig. 2. 3D time-of-flight camera images were also tested, but results were poorer, likely due to high frame-to-frame noise. For now, this approach has only been tested where the player faces the robot at a consistent angle, so future work should include taking posture data from multiple viewpoints.

2.3 Training and classification

We used a Gaussian Mixture Model (GMM) classification scheme for both audio and video. For each modality m, we modeled each instrument class C with a K-component gaussian mixture model, where each component k is a triplet weight-mean-covariance:

$$(w_k, \mu_k, \Sigma_k) \in \mathbb{R} \times \mathbb{R}^K \times \mathbb{R}^{K \times K}) \tag{1}$$

with $\sum_{k=1}^{K} w_k = 1$. The parameters of those models are obtained by training with the expectation-maximization algorithm on manually labelled MSLS and HOG features. For each feature vector $x_{m,t}$, the likelihood is:

$$p(x_{m,t}|C,m) = \sum_{k=1}^{K} w_k \mathcal{N}(x_{m,t}; \mu_k, \Sigma_k) \tag{2}$$

We experimentally set the number of components to K=32; multiple mixtures could capture, for example, all

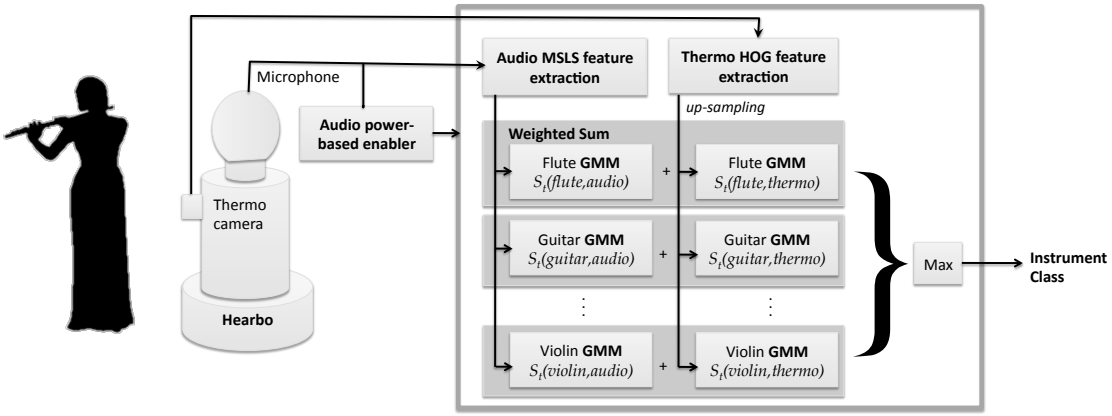


Figure 3: Overview of the musical instrument recognition system.

the various poses of a violin player playing. We smoothed the result by finding the joint likelihood over 50 frames: $S_t(C,m) = \sum_{i=0}^{49} log(p(x_{m,t-i}|C,m))$. Here, we did not try varying the length of the smoothing window number, nor the number of mixtures depending on instrument, though this could be explored in future work.

2.3.1 Fusion

The final score per instrument is calculated with a weighted linear combination of the two scores:

$$S_t(C) = w_1 S_t(C,audio) + w_2 S_t(C,thermo) \qquad (3)$$

We experimentally set $w_1 = 0.2$ and $w_2 = 0.4$. To classify, we found $argmax_C S_t(C)$, the class with the best fused score for the frame at time t.

2.4 Real-time implementation

The system, shown in Fig. 3, was implemented on the Hearbo robot, equipped with an 8ch circular microphone array mounted on its head and a thermo camera (5 [frames/sec]) on its chest. Only one channel audio was used. The open source audio-processing software HARK [6] was used to extract the audio features, and ROS middleware [8] was used to communicate through the network. An off-board notebook PC performed the signal processing. To prevent extraneous detections, we set a audio power enabler such that the system only output a result when sound level exceeded a certain manually-tuned threshold.

3 Evaluation and results

3.1 Data set

Our data set includes 12 instruments: flute, wooden flute, alto ocarina, soprano ocarina, alto recorder, clarinet*, conga*, snare drum*, violin, viola, guitar and bass guitar (where * denotes a synthetic instrument). The players consisted of 1 female and 3 males ranging from beginner to expert, recorded in an anechoic chamber. HEARBO was equipped with a front-facing thermo camera, and the player stood facing the robot, approximately 1.5 metres away. Each player played a scale and/or several songs. Fan noise from the robot's power convertor was audible.

3.2 Procedure and Results

We trained each instrument/modality model with 3 songs and tested with 1 song which was not used for training. Each song was an average of 32 seconds in length, and was labeled to only take into account when sound was audible. This labeling also ensured that the musician was holding the instrument in its playing position, and not, for example, preparing to play. Our frame-by-frame test results show an average recognition result for audio at 80%, for video at 91%, and using our audio-visual weighted scheme, 96%.

3.3 Discussion

The use of a weighted sum rule is a common method for multi-modal fusion. We tried another method during our tests: 1) concatenate the audio and thermo feature vectors, producing a $27 + 128 = 155$ dimensional feature vector 2) train our GMM's on the 155-D concatenated feature 3) test using concatenated features. The results were not as good as the weighted sum method; the concatenated result landed in between the recognition rates for two modalities, at 86%, compared to 96% for the weighted sum result which outperformed both audio and video alone.

4 Future Work

In this work we have shown promising results for a new multi-modal method for instrument recognition, tested on a dataset of 12 instruments. Although this work dealt with monophonic mixtures, it may easily be extended to complex mixtures of polyphonic music using sound source separation and localization. This embodied knowledge of music sounds may be useful for music-playing robots, particularly in ensembles where it must distinguish and track the sounds of multiple players. In the future, we may also wish to extend the video processing frame-by-frame images to motion analysis over time.

This work was partially supported by the Grant-in-Aid for Scientific Research on Priority Areas (No.22118502), KAKENHI (S) and GCOE.

References

[1] A. Krishna and T. Sreenivas, "Music instrument recognition: from isolated notes to solo phrases," ICASSP, 2004, pp. 265-268.
[2] Marques, J. "An automatic annotation system for audio data containing music. Unpublished master's thesis, Massachussetts Institute of Technology, Cambridge, MA, 1999
[3] K. D. Martin. "Sound-Source Recognition: A Theory and Computational Model." PhD thesis, MIT, 1999.
[4] A. Eronen and A. Klapuri. "Musical instrument recognition using cepstral coefficients and temporal features." ICASSP, 2000, pp. 753-756.
[5] J. C. Brown. "Feature dependence in the automatic identification of musical wood- wind instruments." JASA, 109(3): pp.1064-1072 (2001).
[6] K. Nakadai, H.G. Okuno, H. Nakajima, Y. Hasegawa, H. Tsujino. "Design and Implementation of Robot Audition System "HARK"", Advanced Robotics, vol.24, pp.739-761 (2010).
[7] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," CVPR, 2005, pp. 886-893.
[8] ROS: Robot Operating System, Willow Garage, www.ros.org