GRADIENT-BASED MUSICAL FEATURE EXTRACTION BASED ON SCALE-INVARIANT FEATURE TRANSFORM

Tomoko Matsui[1], Masataka Goto[1,2], Jean-Philippe Vert[3,4] and Yuji Uchiyama[5]

[1]Institute of Statistical Mathematics, Tokyo, Japan
[2]National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan
[3]Centre for Computational Biology, Mines ParisTech, Fontainebleau, France
[4]Institut Curie, Inserm U900, Paris, France
[5]Picolab Co., Ltd, Tokyo, Japan
phone: + (81) 50-5533-8537, fax: + (81) 42-526-4335, email: tmatsui@ism.ac.jp
web: www.ism.ac.jp/~tmatsui

ABSTRACT

We investigate a novel gradient-based musical feature extracted using a scale-invariant feature transform. This feature enables dynamic information in music data to be effectively captured time-independently and frequency-independently. It will be useful for various music applications such as genre classification, music mood classification, and cover song identification. In this paper, we evaluate the performance of our feature in genre classification experiments using the data set for the ISMIR2004 contest. The performance of a support-vector-machine-based method using our feature was competitive with the contest even though we used only one fifth of the data. Moreover, the experimental results confirm that our feature is relatively robust to pitch shifts and temporal changes.

1. INTRODUCTION

A tremendous amount of music-related data has recently become available either locally or remotely over networks, and technology for searching this content and retrieving music-related information efficiently is demanded. This consists of several elemental tasks such as genre classification[1-13], artist identification, music mood classification, cover song identification, fundamental frequency estimation, and melody extraction. These tasks have been main research topics at several international conferences (e.g., ICASSP[14] and ISMIR[15]) and evaluation competitions (e.g., MIREX[16]) for music-related data.

Generally, for almost all the tasks, audio data is analyzed frame-by-frame using a Fourier or Wavelet transform, and the spectral feature vectors or chroma features extracted for several tens or hundreds of milliseconds are used[17-19]. Longer dynamic features of several seconds in audio data are not utilized in usual. However, we consider that such dynamic features include useful information for discriminating various musical phenomena and can be effectively utilized for most of the tasks.

In this paper, we propose gradient-based musical feature extraction that is based on the scale-invariant feature transform (SIFT) with the objective of extracting spectral features that are dynamic and, moreover, independent of time and frequency. SIFT is an algorithm originally reported in the computer vision field to detect and describe local features in images by using key points[20]. The key points of SIFT are extracted so as to identify "objects" in an image invariantly with respect to scale and orientation. Here, several seconds of audio data is represented as a 2D spectrogram image. We consider that "objects" in each image correspond to partial areas with locally distinctive spectral features. In our feature extraction, the temporal independence is achieved by utilizing chains of adjacent cluster IDs obtained through clustering SIFT key points extracted from the images, and the frequency independence is achieved by utilizing local dynamic features in the logarithmic frequency domain.

In the following section, our musical feature extraction based on SIFT is described. A genre classification method using our musical feature is introduced in section 3. The performance is evaluated in SVM-based genre classification and the details are explained in section 4. In section 5, our musical featured based on constant Q and FFT spectrograms are compared, and the temporal and frequency independence is discussed. Finally we summarize our findings in section 6.

2. MUSICAL FEATURE EXTRACTION BASED ON SIFT

In SIFT, key points were originally defined as maxima and minima of the results of differences between Gaussian functions applied in scale-space to a series of smoothed images[20]. Low-contrast candidate points and edge response points along an edge are discarded. Dominant orientations are assigned to localized key points. These steps ensure that the key points are more stable for matching and recognition. SIFT descriptors robust to local affine distortion are then obtained by considering pixels around the radius of each key point and by blurring and re-sampling the local image orientation planes.

The SIFT key points on a constant Q spectrogram image of 5-s audio data are shown in Figure 1. They were extracted using Lowe's software[21]. They were found to be located near partial areas with distinctive spectral features. The number of the points is roughly 2000. Around the key point location, the 4x4 descriptors are created by calculating the

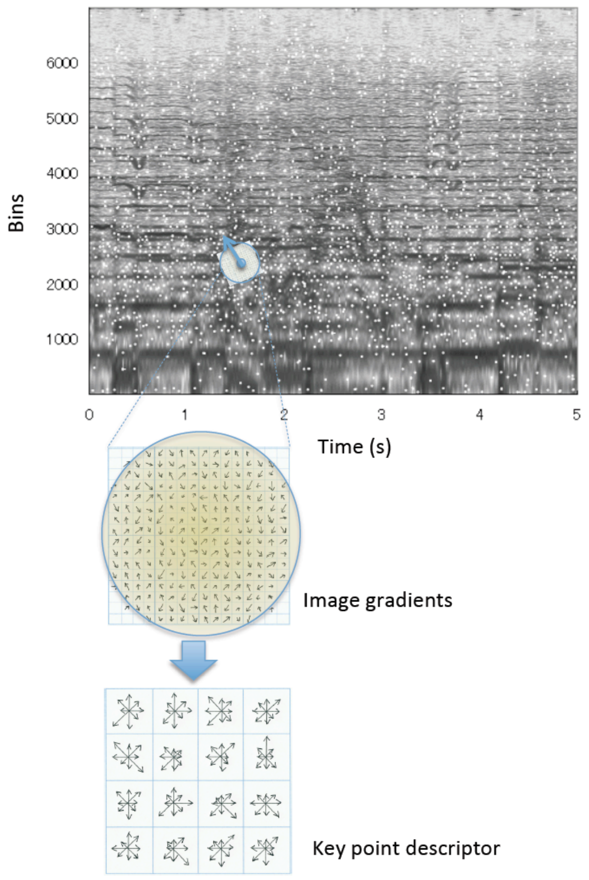

Figure 1 – SIFT key points on a spectrogram image of 5-s audio data of "rock_pop" (top) and illustrations of image gradients for a key point (middle) and key point descriptor (bottom).

gradient magnitude and orientation at image sample points as shown in the middle of Figure 1. Those are weighted by a Gaussian window which is indicated by the overlaid circle. The samples are then accumulated into orientation histograms for 8 directions summarizing the contents over 4x4 regions, as shown in the bottom of Figure 1, with the length of each arrow corresponding to the sum of the gradient magnitudes near that direction within the region. At the result, each key point is represented using a 128(=4x4x8)-dimensional feature vector. Note that absolute coordinate information is not included in the feature vector, so the feature vectors are frequency-independent and represent local image characteristics around the point. Especially since the frequency axis is in a constant (logarithmic) scale here, the local image characteristics reflect the octave structure in musical data.

In order to manage an enormous total number of key points for all training images (e.g., roughly 150,000 when the number of training images is 1000), we first cluster all the key points by, for instance, using the k-means method. Then, each key point is represented using the cluster ID, each image is basically represented using the appearance frequency

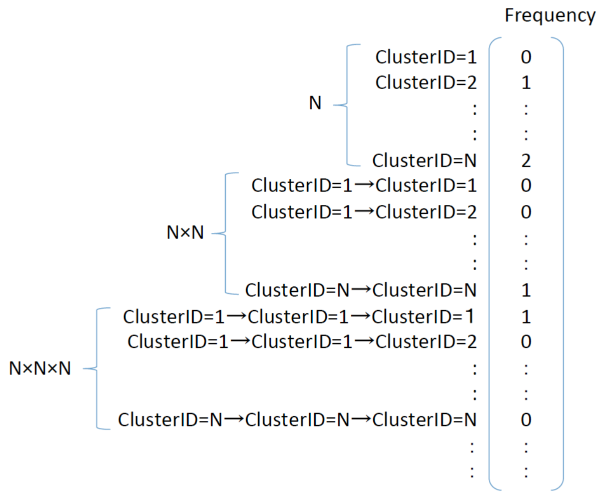
Frequency
N
ClusterID=1 0
ClusterID=2 1
ClusterID=N 2
N×N
ClusterID=1→ClusterID=1 0
ClusterID=1→ClusterID=2 0
ClusterID=N→ClusterID=N 1
N×N×N
ClusterID=1→ClusterID=1→ClusterID=1 1
ClusterID=1→ClusterID=1→ClusterID=2 0
ClusterID=N→ClusterID=N→ClusterID=N 0

Figure 2 – Musical feature composed using frequencies of singlet, doublet, and triplet SIFT key points.

of each key point, and a musical feature is composed as in Figure 2. Here, the elements for "NxN" represent the frequencies of the nearest two key points on the time axis, and the elements for "NxNxN" represent the frequencies of the nearest three key points.

Since the nearest points do not depend on the absolute time difference between them, our musical feature represents time-independent dynamic features with two or three successive key points. Moreover, since the absolute coordinate information is discarded as mentioned before, our musical feature is frequency independent.

3. APPLICATION TO GENRE CLASSIFICATION

In this section, we propose a genre classification method using our musical feature. The basic procedure for input data having a length of *l* seconds is shown in Figure 3. Here, we used the one-vs.-all method with support vector machines (SVMs) estimated for each genre. Since our musical feature vector has a huge number of dimensions, O(N^3), a robust classifier like an SVM is necessary to handle the high dimensionality problem. For longer input data, in practice, the data is segmented into *l*-s lengths and the SVM scores for each segment are summed for each genre in order to make a final decision.

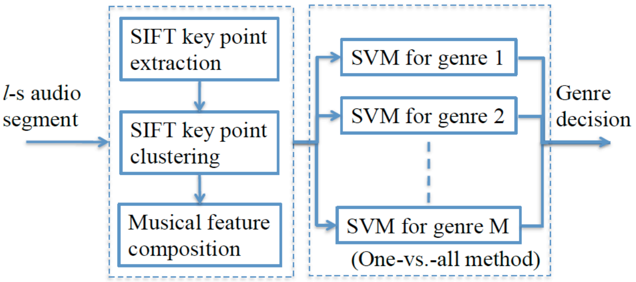

Figure 3 – Genre classification procedure using musical feature.

4. EXPERIMENTS

Here, we examine the performance of SVM-based genre classification using our musical feature.

4.1 Conditions

We used the training and development sets for the ISMIR2004 genre classification contest as our training and testing data, respectively[22]. Both the training and testing data consisted of samples from six genres, and the distributions of the samples over genres were the same. The second column of Table 1 lists the numbers of samples for training (or testing) data for each genre.

Genre	No. of samples	No. of images	No. of SIFT key points
classical [201.0]	320	3200	1769.3
electronic [327.9]	115	1150	2411.4
jazz_blues [230.0]	26	260	2111.4
metal_punk [251.4]	45	450	2415.9
rock_pop [226.1]	101	1010	2285.5
world [351.6]	122	1220	1977.2
Total/Average Average [253.8]	Total 729	Total 7290	Average 2161.8

Table 1 – Numbers of training samples and images and the average number of SIFT key points per image for each genre ([ ]: average length of original samples (s)).

We randomly selected ten 5-s intervals for each sample and represented them as 2D spectrogram images. The total number of training or testing images was 7290. Note that the average lengths of training and testing samples were 253.8 and 242.3 s, respectively. Since we selected the total 50-s length from each sample, the amount of data used in the experiments was roughly one fifth of the ISMIR2004 contest. For 2D spectrogram images, we used the constant Q transform in which 4096 FFT bins were expanded to 6983 bins corresponding to a frequency band from 80 Hz to 14 kHz. We extracted SIFT key points for each image. The fourth column of Table 1 lists the average number of SIFT key points for training data. It is interesting that for the classical genre, which has relatively slow and soft motifs, the number of key points was small.

We used SVMs with a linear kernel and an error term penalty parameter of 1. We examined different numbers of clusters, C = 100, 200, 300, 500 and 1000, to cluster the SIFT key points. The following three cases were studied.

(L1) Using only singlet SIFT key points
(dimensions of musical feature: N)

(L2) Using singlet and doublet SIFT key points
(dimensions of musical feature: N + N^2)

(L3) Using singlet, doublet, and triplet SIFT key points
(dimensions of musical feature: N + N^2 + N^3)

In case L1, no dynamic information was utilized as shown in Figure 2.

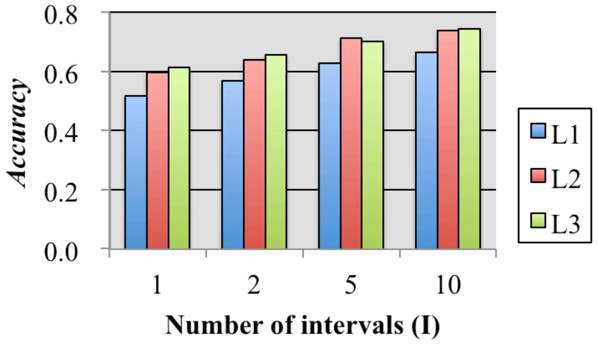

Figure 4 – *Accuracies* for various 5-s interval numbers (C = 1000).

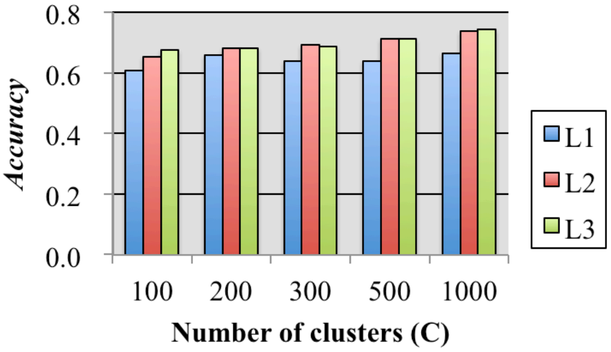

Figure 5 – *Accuracies* for various cluster numbers (I = 10).

In testing, we used the same two evaluation metrics as used for the ISMIR2004 contest. The first metric is the accuracy of correctly guessed genres (*Accuracy*) defined by

$$Accuracy = \sum_{c \in genres} p_c \cdot guessed_c .$$

The samples per genre were not equidistributed. The evaluation metric normalizes the number of correctly predicted genres (*guessed*) by the probability of appearance of each genre (p_c). The second metric is the average percentage of correct answers (*Correct*).

For testing, we evaluated with *Accuracies* and *Corrects* calculated by using the score for each 5-s interval and by using the sum of the scores for two, five, and ten 5-s intervals randomly selected for a sample.

4.2 Results

Accuracies for different numbers of 5-s intervals for each trial are shown in Figure 4. The number of clusters for SIFT key points was 1000 in all cases. As the total data length became longer (i.e., the number of 5-s intervals used to extract SIFT key points became larger), the performance increased and the difference in performance between cases L2 and L3 became smaller. It can be considered that in this method, the dynamic features over a key point doublet is robustly utilized, while the combinatorial number of triplet key points is larger, so the stability for capturing the dynamic features over triplet key points decreased. In cases L1, L2, and L3 with I = 10 (10 intervals), *Corrects* were 0.775, 0.823 and 0.827, respectively. The performance of our method was competitive with the results in the ISMIR2004 genre classification contest even though we used only one fifth the amount of data for training and testing.

Accuracies for different cluster numbers for the SIFT key point clustering are shown in Figure 5. As the number of clusters increased, the performance increased especially for cases L2 and L3. These results indicate that doublet or triplet key points have discriminative information for genre classification. *Corrects* for different cluster numbers are listed in Table 2, which also shows this tendency.

No. of clusters	100	200	300	500	1000
L1	0.711	0..781	0.779	0.778	0.775
L2	0.763	0.796	0.801	0.812	0.823
L3	0.785	0.796	0.804	0.811	0.827

Table 2 – *Corrects* for different cluster numbers (I = 10) when using our musical features based on constant Q spectrograms.

5. DISCUSSION

5.1 Constant Q vs. FFT spectrograms

In the above experiments, we used constant Q spectrograms so as to better capture the octave structure. Here we examine the performance with our musical feature for which the SIFT key points are extracted from FFT spectrograms in order to confirm the effectiveness of use of constant Q spectrograms. Table 3 lists *Corrects* for our musical features based on FFT spectrograms. The training and testing data used here consist of almost the same 5-s intervals (the difference in time is within 1 ms) as ones used in the experiments in section 4. The FFT was calculated using a Hamming window of 4096 length and a frequency band from 0 to 14 kHz on 2D spectrogram images was used in SIFT. When comparing *Corrects* in Tables 2 and 3, we can see that our musical features based on constant Q spectrograms outperform those based on FFT spectrograms especially for larger cluster numbers. The best *Correct* for constant Q based features was 0.827 (L3 and 1000 clusters in Table 2) and it was 5% relative improvement from the best *Correct* of 0.789 for FFT based features (L2 and 1000 clusters in Table 3).

No. of clusters	100	200	300	500	1000
L1	0.726	0.746	0.757	0.760	0.693
L2	0.739	0.761	0.787	0.786	0.789
L3	0.765	0.761	0.786	0.781	0.783

Table 3 – *Corrects* for different cluster numbers (I = 10) when using our musical features based on FFT spectrograms.

5.2 Temporal and frequency independence

We examined the temporal and frequency independence of our method through comparison with a simple method based on the Gaussian mixture model (GMM) with mel-frequency cepstral coefficient (MFCC) feature vectors. We transformed the testing samples using an audio editor, Audacity[23], and simulated five conversions: (1) lower pitch offset by half an octave [pitch_L], (2) higher pitch offset by half an octave [pitch_H], (3) slower tempo: one fourth of the original one [tempo_S1], (4) slower tempo: one third of the original one [tempo_S2], and (5) faster tempo: half of the original one [tempo_F].

For the GMM-based method, a feature vector of 60 components, consisting of 20 MFCCs and their first and second derivatives (dynamic information), was derived once every 16 ms over a 32-ms Hamming-windowed audio segment. In training, a diagonal-covariance GMM with 30 mixture components was created for each genre. The number of

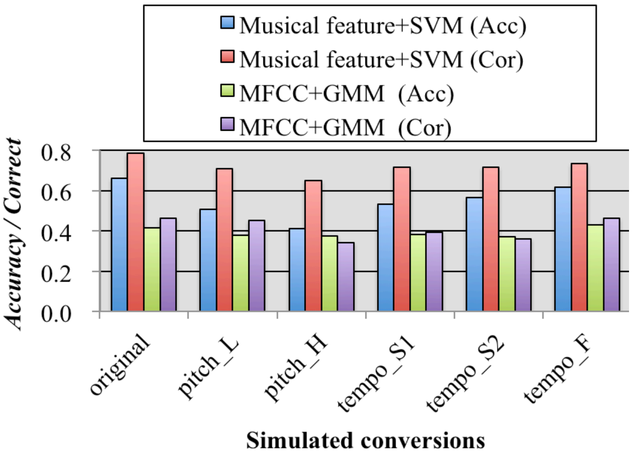

Figure 6 – Comparison of *Accuracy* and *Correct* for musical feature+SVM and MFCC+GMM methods when using simulated samples with higher & lower pitch offsets and with slower & faster tempos.

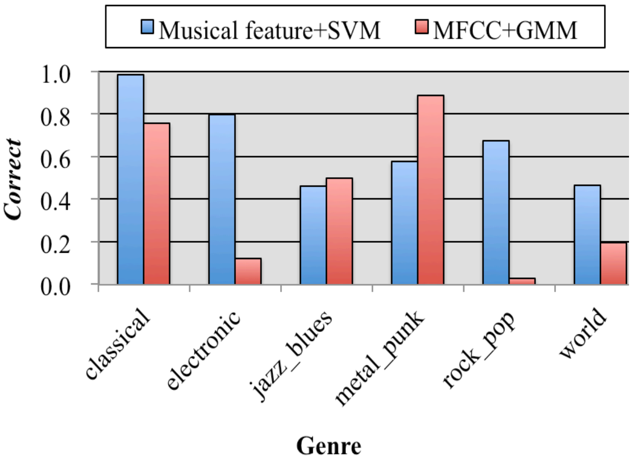

Figure 7 – Comparison of *Corrects* for musical feature+SVM and MFCC+GMM for each genre.

mixture components was decided through preliminary experiments and both training and testing data were the same as in section 4.1 (ten 5-s lengths of data randomly selected from each sample). In testing, the genre of GMM with the highest likelihood was identified as the one for input data, which was a concatenation of ten 5-s intervals.

Accuracies and *Corrects* of our musical feature based on FFT spectrograms+SVM method (L3, C = 1000, and I = 10) and the MFCC+GMM method for simulated conversions are shown in Figure 6. Our method outperformed MFCC+GMM for all the conversions. The *Correct* reduction rates compared with ones for the original samples were 10.2% for our method and 12.8% for MFCC+GMM. Our method is relatively robust to variations in pitch and tempo.

Corrects for musical feature based on FFT spectrograms+SVM (L3, C = 1000, and I = 10) and MFCC+GMM for each genre are compared in Figure 7. The standard deviations for our method and MFCC+GMM were 0.205 and 0.355, respectively. Our method enabled relatively stable performance over genres to be obtained. This confirms that our method captures the dynamic features for each genre effectively.

6. CONCLUSION

We investigated a novel method of musical feature extraction based on SIFT. Our feature can effectively capture the local dynamic information in the logarithmic frequency domain. The experimental results confirm that the SVM method together with our feature is robust to variations in pitch and tempo and has time- and frequency-independent characteristics.

Our future work will include evaluating our method using a larger amount of audio data and investigating the use of longer dynamic features with nonlinear SVMs.

REFERENCES

[1] G. Tzanetakis and P. Cook, “Musical genre classification of audio signals, IEEE Transactions on Speech and Audio Processing, Vol.10, No.5, pp.293-302, 2002.

[2] T. Li, M. Ogihara, and Q. Li, “A comparative study on content-based music genre classification,” in Proc. the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pp.282-289, 2003.

[3] C. Xu, N. C. Maddage, X. Shao, F. Cao, and Q. Tian, “Musical genre classification using support vector machines,” in Proc. ICASSP 2003, pp.429-432, 2003.

[4] C. McKay and I. Fujinaga, “Automatic genre classification using large high-level musical feature sets,” in Proc. ISMIR, 2004.

[5] E. Pampalk, A. Flexer, and G. Widmer, “Improvements of audio-based music similarity and genre classificaton,” in Proc. ISMIR, 2005.

[6] T. Lidy and A. Rauber, “Evaluation of feature extractors and psycho-acoustic transformations for music genre classification,” in Proc. ISMIR, 2005.

[7] A. Meng, P. Ahrendt, J. Larsen, and L. K. Hansen, “Temporal feature integration for music genre classification,” IEEE Transactions on Audio, Speech, and Language Processing, Vol.15, No.5, pp.1654-1664, 2007.

[8] T. Lidy, A. Rauber, A. Pertusa, and J. M. Inesta, “Improving genre classification by combination of audio and symbolic descriptors using a transcription system,” in Proc. ISMIR, 2007.

[9] A. Holzapfel and Y. Stylianou, “Musical genre classification using nonnegative matrix factorization-based features,” IEEE Transactions on Audio, Speech, and Language Processing, Vol.16, No.2, pp.424-423, 2008.

[10] C.-H. Lee, J.-L. Shih, K.-M. Yu, and H.-S. Lin, “Automatic Music Genre Classification Based on Modulation Spectral Analysis of Spectral and Cepstral Features,” IEEE Transactions on Multimedia, Vol.11, No.4, 670-682, 2009.

[11] C. McKay, J. A. Burgoyne, J. Hockman, J. B.L. Smith, G. Vigliensoni, and I. Fujinaga, “Evaluating the Genre Classification Performance of Lyrical Features Relative to Audio, Symbolic and Cultural Features,” in Proc. ISMIR, 2010.

[12] N. A. Draman, C. Wilson, and S. Ling, “Modified Ais-based Classifier for Music Genre Classification,” in Proc. ISMIR, 2010.

[13] K. K. Chang, J.-S. R. Jang and C. S. Iliopoulos, “Music Genre Classification via Compressive Sampling,” in Proc. ISMIR, 2010.

[14] The 36th International Conference on Acoustics, Speech and Signal Processing (http://www.icassp2011.com).

[15] The International Society for Music Information Retrieval Conferences (http://www.ismir.net).

[16] The Music Information Retrieval Evaluation eXchange (http://www.music-ir.org/mirex/wiki/MIREX_HOME).

[17] R. Typke, F. Wiering, and R. C. Veltkamp, “A survey of music information retrieval systems,” in Proc. the International Conference on Music Information Retrieval, pp. 153–160, 2005.

[18] M. Goto and K. Hirata, “Recent studies on music information processing,” Acoustical Science and Technology (edited by the Acoustical Society of Japan), Vol. 25, No. 6, pp. 419–425, 2004.

[19] M. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney, “Content-Based Music Information Retrieval: Current Directions and Future Challenges,” in Proc. the IEEE, Vol.96, No.4, pp.668-696, 2008.

[20] D. G. Lowe. “Distinctive image features from scale-invariant keypoints,” International Journal of Computer Vision, Vol. 60 (2) pp. 91–110, 2004.

[21] D. G. Lowe, “Demo Software: SIFT Keypoint Detector” (http://people.cs.ubc.ca/~lowe/keypoints/).

[22] ISMIR2004 Audio Description Contest—Genre/Artist ID Classification and Artist Similarity (http://ismir2004.ismir.net/genre_contest/index.htm).

[23] Audacity (http://audacity.sourceforge.net/).