Enhancing Music Information Retrieval by Incorporating Image-Based Local Features

Leszek Kaliciak[1], Ben Horsburgh[1], Dawei Song[2,3], Nirmalie Wiratunga[1], Jeff Pan[4]
[1]The Robert Gordon University, Aberdeen, UK
[2]Tianjin University, Tianjin, China
[3]The Open University, Milton Keynes, UK
[4]Aberdeen University, Aberdeen, UK
{l.kaliciak,b.horsburgh,n.wiratunga}@rgu.ac.uk; Dawei.Song@open.ac.uk;
jeff.z.pan@abdn.ac.uk

Abstract. This paper presents a novel approach to music genre classification. Having represented music tracks in the form of two dimensional images, we apply the "bag of visual words" method from visual IR in order to classify the songs into 19 genres. By switching to visual domain, we can abstract from musical concepts such as melody, timbre and rhythm. We obtained classification accuracy of 46% (with 5% theoretical baseline for random classification) which is comparable with existing state-of-the-art approaches. Moreover, the novel features characterize different properties of the signal than standard methods. Therefore, the combination of them should further improve the performance of existing techniques.
The motivation behind this work was the hypothesis, that 2D images of music tracs (spectrograms) perceived as similar would correspond to the same music genres. Conversely, it is possible to treat real life images as spectrograms and utilize music-based features to represent these images in a vector form. This points to an interesting interchangeability between visual and music information retrieval.

Keywords: Local features, Co-occurrence matrix, Colour moments, K-means algorithm, Fourier transform

1 Introduction

Almost every representation of music used in the field of Music Information Retrieval (MIR) involves extracting features from music transformed into the frequency domain. These features include chromatic, melodic, harmonic, rhythmic, and timbral measures.

Thus, [1] represents the distribution of chroma within a song as a histogram. The songs with similar chroma histogram distributions are considered similar. The temporal aspects of pitch are taken into account by [2] and [3]. In [4] authors try to capture the rhythm by constructing a self-similarity matrix based upon the similarity of each short time frequency spectra extracted from the audio. Bello [5] presents a method to describe a novelty function (a common method for identifying onsets) of a waveform inspired by Foote's [4] similarity measure. One of the challenges in MIR is how to interpret the classification confusions. Some incorrectly classified instances cast doubt on whether

the ground truth is correct (for example a pop song that could be labeled as funk). The solution to this problem might be the incorporation of fuzzy logic.

Our approach implements methods from both visual IR (VIR) and MIR research areas. The motivation behind this work was the hypothesis, that 2D images of music tracs (spectrograms) perceived as similar would correspond to the same music genres (perhaps even similar music tracs). Conversely, we can treat real life images as spectrograms and utilize music-based features to represent these images in a vector form. This would point to an interesting interchangeability between visual and music information retrieval. Thus, instead of extracting features directly from frequency domain, we generate an image of each song that shows how the spectral density of a signal varies with time. Two geometric dimensions represent frequency and time, and the colour of each point in the image indicate the amplitude of a particular frequency at a particular time. The advantage of such visual representation is that it does not rely on musical

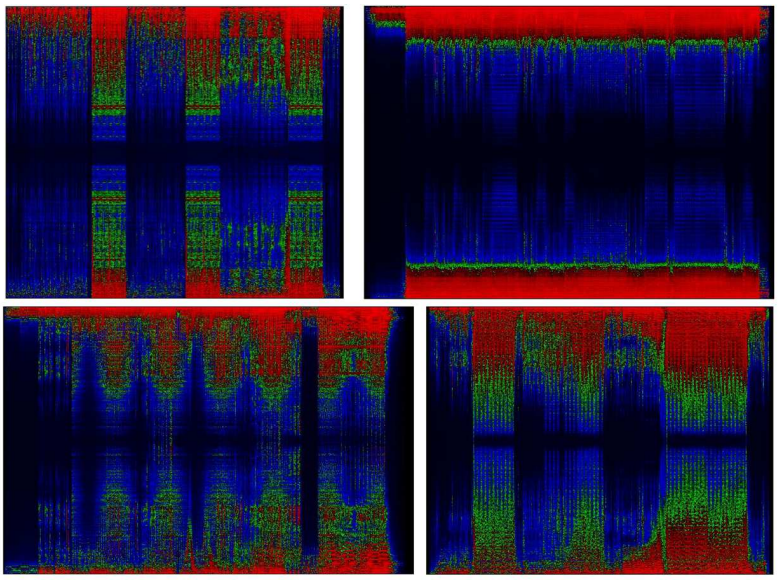

Fig. 1. Music representation in the form of 2D images

concepts (melody, timbre, rhythm, etc.). It is also intuitive that two songs can be compared visually based on their spectrogram representations. The next step involves the application of local features from Content Based Image Retrieval and representation of music in the form of histograms of visual words counts. Thus obtained histograms, characterizing individual songs, are used for genre music classification task.

Because global approaches find it hard to capture all the properties of an image, the implemented local features are based on the “bag of visual words” approach. The first

step in the "bag of features" [1] method is to localize the points of interest (point-like, region-like) by using corner or blob detectors. Other sampling techniques include random and dense sampling. The second step involves the representation of regions around the sample points in a form of multidimensional vectors. There are various existing descriptors, the SIFT (Scale Invariant Feature Transform) being one of the most widely used. The initial extraction is performed on a training set of images and the K-means clustering is applied to it. Each cluster will correspond to one "visual word", a local pattern. Finally, each image in a data collection can be characterized by a histogram of "visual words" counts.

The most common interest points detectors are: Harris-affine, Hessian-affine, SIFT (Scale Invariant Feature Transform), Maximally Stable Extremal Regions (MSER). Among descriptors we have SIFT (detector and descriptor), local jets (image derivatives), steerable filters, generalized moment invariants. One of the variations of "bag of features" (B.O.F) method based on SIFT detector and descriptor was first proposed by Lowe in [6]. Other good sources of information about scale space and local features are [7] and [8]. In [9] authors applied the local features to nude or pornographic images detection. Instead of using the well known SIFT descriptor, they implemented Hue-SIFT method in order to take colour into account. [10] is a comparison of different techniques used in B.O.F. approach for image sampling, visual dictionary generation and normalization of the histograms. Yang et al. [11] incorporated and tested some methods derived from textual information retrieval domain into CBIR (B.O.F.): term weighting, stop word removal, feature selection. They also conducted experiments testing the influence of the vocabulary size (number of clusters) and spatial information on the retrieval performance.

In this paper we utilize fast and easy to implement method based on local features [15]. Because quite often randomly generated sample points are more discriminant than the points detected by corner detector (especially when it comes to a large number of sample points when the set of keypoints detected by corner detectors becomes saturated), we decided to implement a hybrid sampling technique. Comparison between random, dense, pure corner-based, and hybrid sampling showed the superiority of this type of sampling. Our descriptor is based on co-occurrence matrix and colour moments. For the description of the image patches around the sample points we used separately: co-occurence matrix computed at eight different orientations, and three colour moments to capture the local colour properties. The co-occurence matrix has proven to be an effective way of texture representation and by considering multiple orientations we make it invariant to rotation. Colour moments are fairly insensitive to changes in viewpoint, and their computation is trivial. Moreover, patches characterized by colour moments are also able to capture the local textural information. Despite the relative simplicity of the model, this method was able to obtain results comparable with current state-of-the-art (ImageCLEF2010 Wikipedia Retrieval Task).

One of the main advantages of our model is that by switching to visual domain we can abstract from the musical concepts and still obtain results comparable with current state-of-the-art. Moreover, the proposed approach could be used for the characterization of signal in general as an alternative to common techniques.

[1] Terms "bag of visual words" and "bag of features" will be used interchangeably in this paper.

The paper is organized as follows. In Section 2 we present the model and describe the developed algorithm in detail. It consists of the sub-sections introducing the local features based on the "bag of visual words", the Fast Fourier Transform and a sub-section on spectrograms generation. Section 3 is devoted to the experiments and discussion. It describes the data collection used in the experiment, the detailed experimental setup and results with their analysis. We draw conclusions in Section 4 and finally present some ideas for future research (Section 5).

2 The Proposed Model

Here, we give a detailed description of the proposed model. The framework establishes the link between MIR and VIR research areas.

Our algorithm consists of the following stages:

1. **Transformation to frequency domain:** Transform the music data from the time to frequency domain using Fast Fourier Transform (FFT). Since audio signals are periodic over time, it is convenient to represent them as a sum of infinite number of sinusoidal waves. It makes it easier to analyze sinusoidal functions than general shaped functions.

2. **Music representation, visual data generation:** Generate spectrograms in two dimensional space, where the geometric dimensions represent frequency and time, and the colour of each point in the image indicate the amplitude of a particular frequency at a particular time. These spectrograms are generated from the signal transformed by FFT. Our method of spectrograms generation was designed in such a way as to produce images containing easy to capture visual properties.

3. **Image sampling:**
   – **Keypoints detection:** Apply Shi and Tomasi (see [14]) method to find the points of interest.
   – **Random sampling:** Apply random points generator to produce another half of the sample points.
   – **Random sampling:** Alternatively, dense sample images (divide images into a number of uniform, non-overlapping rectangular sub-images).

4. **Description of local patches:** Characterize local patches in the form of co-occurence matrix or colour moments. In case of co-occurence matrix extract the meaningful statistics - energy, entropy, contrast, and homogeneity. Compute the features for individual colour channels.

5. **Feature vector construction:** Represent local patterns as 9 dimensional (moments) or 12 dimensional (co-occurence) vector.

6. **Visual dictionary generation:** Apply K-means clustering to the training set in order to obtain the codebook of visual words.

7. **Histogram computation:** Create a histogram of visual words counts by calculating the distance between image patches and cluster centroids.

8. **Music genre classification:** The classification of music data into music genres is performed by k-nearest neighbour algorithm, based on Minkowski's fractional similarity measure.

Steps 3 to 7 are related to generation of visual representations of the spectrograms. In the course of this research, global methods like colour moments, co-occurrence matrix (texture), colour correlograms were also tested. We utilize local features because of their superior performance over various global methods. The local features may also have another advantage over other models. An interesting future work would be to investigate if image patches identified by corner detectors (roughly speaking - locations of a sudden change of pixel intensities) and "visual words" correspond to some important characteristics of audio signal.

2.1 The Fast Fourier Transform

Let $x_0, ..., x_{N-1}$ be complex numbers. The Discrete Fourier Transform (DFT) is defined by

$$X_k = \sum_{n=0}^{N-1} x_n e^{-i 2\pi k \frac{n}{N}} \qquad k = 0, ..., N-1.$$

Computing DFT requires $O(N^2)$ operations, while FFT reduces the number to $O(N \log N)$. The implemented FFT method incorporates Cooley-Tukey algorithm, which breaks down a DFT into smaller DFTs. The audio sampled size is 65536 bytes (1.486 seconds), with sampling rate of 44100Hz.

2.2 Spectrogram Generation

The resulting spectrum is split into 512 bins (64Hz / bin). The power of each bin is converted into a pixel as follows:

```
Let colour = power / meanPower

IF {colour > 1}
   r = ((1 - (1 / colour)) / 2) + 0.5f
ELSE
   IF {colour > 0.5}
      g = (((colour-0.5)/0.5) / 2) + 0.5f
ELSE
   b = ((colour / 0.5) / 2) + 0.5f

   ENDIF
ENDIF
```

The horizontal dimension of each image represents the time (1 pixel = 1.486sec), vertical dimension represents frequency (1 pixel = 64Hz), and pixel intensities represent power.

This method of spectrogram generation produces images that varies in colour and texture. These properties make the images suitable for application of visual features. An interesting observation is that some genres are easily recognizable directly from our spectrograms. Classical music, for instance, is characterized by a presence of the blue colour joining the top and the bottom part of an image.

2.3 The Sampling Technique

Recently, an approach based on local features extraction has become quite popular in Visual Information Retrieval. Global approaches find it hard to capture all the properties of an image. The most recent state-of-the-art in Image Retrieval is based on so-called "bag of visual words". The first step in the "bag of features" method is to localize the points of interest (point-like, region-like) by using corner/blob detectors. Other sampling techniques include random and dense sampling. The second step involves the representation of regions around the sample points in a form of multidimensional vectors. There are various existing descriptors, the SIFT (Scale Invariant Feature Transform) being one of the best. The initial extraction is performed on a training set of images and the K-means clustering is applied to it. Each cluster will correspond to one "visual word", a local pattern. Finally, each image in a data collection can be characterized by a histogram of "visual words" counts. The most popular interest points detectors are: Harris-affine, Hessian-affine, Scale Invariant Feature Transform (SIFT), Maximally Stable Extremal Regions (MSER). Among descriptors we have SIFT (detector and descriptor), local jets (image derivatives), steerable filters, generalized moment invariants.

Let us now return to the local features utilized in this paper. As aforementioned, the implemented hybrid sampling method combines Shi and Tomasi [14] corner detection with a random number generator. The Shi and Tomasi method is based on the Harris corner detector. The change of pixel intensities is characterized as

$$S(x,y) = \sum_u \sum_v w(u,v)(I(u,v) - I(u+x, v+y))^2. \tag{1}$$

From Taylor approximation of the first order we get

$$I(u+x, v+y) \approx I(u,v) + I_x(u,v)x + I_y(u,v)y. \tag{2}$$

Substituting (2) in (1) we obtain

$$S(x,y) \approx \sum_u \sum_v w(u,v)(I_x(u,v)x + I_y(u,v)y)^2. \tag{3}$$

We can rewrite equation (3) in the following form

$$S(x,y) \approx \begin{pmatrix} x & y \end{pmatrix} A \begin{pmatrix} x \\ y \end{pmatrix} \tag{4}$$

where $A =$

$$\sum_u \sum_v w(u,v) \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix} = \begin{bmatrix} \langle I_x^2 \rangle & \langle I_x I_y \rangle \\ \langle I_x I_y \rangle & \langle I_y^2 \rangle \end{bmatrix}. \tag{5}$$

We define "cornerness", a measure of corner response as

$$M_c = \lambda_1\lambda_2 - \kappa(\lambda_1 + \lambda_2)^2 = det(A) - \kappa trace^2(A). \quad (6)$$

We assume that the corner was detected if M_c is sufficiently large. Shi and Tomasi found that the good corners can be obtained by setting a minimum threshold and checking if the smaller of the eigenvalues is greater than the threshold.

Another sampling technique implemented for comparison purposes was dense sampling. In this case, each image was divided into the same number of 900 identical rectangular sub-images.

2.4 Region Descriptors

Each local patch in an image was represented as

- The 8 orientational co-occurence matrix.
- Colour moments.

A simple co-occurence matrix is defined as follows

$$C_{\Delta x,\Delta y}(i,j) = \sum_{p=1}^{n}\sum_{q=1}^{m}\begin{cases} 1, if\ I(p,q) = i\ and\ I(p+\Delta x, q+\Delta y) = j \\ 0, \quad otherwise \end{cases}$$

The matrix describes the way certain grayscale pixel intensities occur in relation to other grayscale pixel intensities. It counts the number of such patterns. The most discriminating statistics extracted from co-occurence matrix are: contrast, inverse difference moment, entropy, energy, homogeneity, and variance.

The method based on three colour moments assumes that the distribution of colour can be treated as probability distribution. Three statistics extracted from individual colour channels are

- Mean $E_i = \sum_{j=1}^{n} \frac{1}{N} p_{ij}$
- Standard Deviation $\sigma_i = \sqrt{\left(\frac{1}{N}\sum_{j=1}^{N}\left(p_{ij} - E_i\right)^2\right)}$
- Skewness $s_i = \sqrt[3]{\left(\frac{1}{N}\sum_{j=1}^{N}\left(p_{ij} - E_i\right)^3\right)}$

The first moment can be interpreted as an average colour value, second as a square root of the variance of the distribution, and third as the measure of asymmetry in the distribution. One can construct the weighted similarity measure as an analogy to manhattan metric, for example:

$$d(H,I) = \sum_{i=1}^{r} w_{i1}\left|E_i^1 - E_i^2\right| + w_{i2}\left|\sigma_i^1 - \sigma_i^2\right| + w_{i3}\left|s_i^1 - s_i^2\right|.$$

Colour moments can also capture the textural properties of an image and are fairly insensitive to viewpoint changes. By computing them in HSV colour space we can make the statistics insensitive to illumination changes.

2.5 Feature Vector Construction, Visual Dictionary Generation, and Histogram Computation

The local patches are represented as multidimensional vectors constructed from different statistics, extracted from individual colour channels. By taking a sample training set consisting of collection's representative images, we can generate so-called visual vocabulary. The K-means clustering algorithm has been used for that purpose. Each cluster characterizes a local pattern, representing specific "visual word". The histogram of visual words counts is created by computing the manhattan distance between individual patches and cluster centroids, and calculating how many patches belong to specific clusters.

2.6 Music Genre Classification

The classification of music data into music genres is performed by k-nearest neighbour algorithm, using Minkowski's fractional similarity measure

$$d(x,y) = \left(\sum_{i=1}^{n} \sqrt{|x_i - y_i|} \right)^2$$

where $x = (x_i)$ and $y = (y_i)$ are the n dimensional feature vectors. It was experimentally proven (see [12]) that the fractional measures from Minkowski's family of distances yield good results in VIR.

3 Experiment and Discussion

Figure 2 shows the query by visual example retrieval based on the local features and our music representation.

For the experimental purposes we used a data collection consisting of 4759 music tracks. The genre distribution is presented in table 1.

Genre labels were extracted from iTunes. The local feature algorithm uses 900 sample points per image, for each sample point we open a square window 10 by 10 pixels wide. The dimensionality of the histograms of visual words counts is 40. The applied k-nearest neighbour algorithm uses 9-fold cross-validation, 12 nearest neighbours, distance weighting and manhattan metric. The classification accuracy we obtained with dense sampling was approximately **46 per cent** (2176 tracks) of correctly classified instances. The hybrid sampling scored lower, resulting in **43 per cent** (2051 tracks) of retrieval accuracy. The reason for this lies in the worse performance of corner detector in this domain. The local features with hybrid sampling performed better than the one with dense sampling on ImageCLEF2007 and MIRFlickr25000 collections, consisting of real-life images.

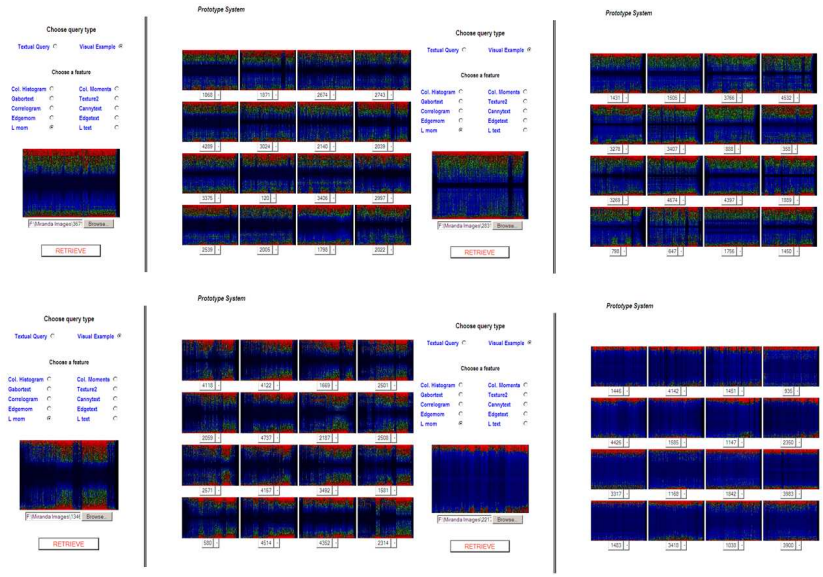

Fig. 2. Local features at work

From the confusion analysis we observed that most incorrectly classified instances were confused with similar genres, and the song-genre correspondence was arguable and subjective. That is why it is so hard to improve the retrieval performance. Good, natural solution to this problem could be the incorporation of fuzzy logic, and associate each song with certain probability of it being in one of the classes. The problem with comparisons with other methods arises because of the lack of the standardized data collections in MIR. Many data collections have unequally distributed data sets, different number of genres, more specialized or generalized classes.

All of that affects the behavior of classifier. Meng and others [13] used a multivariate autoregressive feature model, considered as current state-of-the-art in MIR, to capture the temporal information in the window. The data set used consisted of 1210 music tracks with 11 genres. The best mean classification accuracy they obtained were 44 and 40 per cent for the LM and GLM classifiers. It should be noted though that the accuracies obtained by the automatic classification need to be relative to the theoretical baseline for random classification which is 9% for [13], and 5% for our collection. It means that the performance of our method is actually much better. There are also other aspects, mentioned previously, that make the evaluation difficult.

In his PhD thesis on music genre classification, Serra presents a “non exhaustive list for the most relevant papers presented in journals and conferences for the last years” [16]. He concludes that “although accuracies are not completely comparable due to the different datasets the authors use, similar approaches have similar results. This suggest

Table 1. Genre classes

Genre	Tr.	Genre	Tr.
Pop	1024	Country	82
Alternative and Punk	919	Hip-Hop	81
Rock	862	Reggae	80
R&B	516	Easy Listening	80
Classical	293	Musicals	75
Dance	265	Latin	62
Alternative	139	Christmas	42
Folk	115	Rap	15
Metal	89	Soundtrack	11
Blues	9		

that music genre classification, as it is known today, seems to reach a "glass ceiling"". The reported accuracies were then plotted with respect to the number of genres (Figure 3).

The human performance in classifying music genres (10 genres) is around 53% correctly cassified for 250ms samples and around 70% for samples longer than 3s [17]. Thus, the performance of current state of the art models for genre classification is comparable with the human performance.

4 Conclusions

In this paper we propose a novel approach to MIR. Having represented the music tracks in the form of two dimensional images, we apply the "bag of visual words" method from visual IR in order to classify the songs into 19 genres. The motivation behind this work was the hypothesis, that 2D images of music tracs (spectrograms) perceived as similar would correspond to the same music genres (perhaps even similar music tracs). Conversely, we can treat real life images as spectrograms and utilize music-based features to represent these images in a vector form. This would point to an interesting interchangeability between visual and music information retrieval.

First, songs are represented as images generated from Fourier frequency domain. The next step involves indexing thus obtained data collection by applying the following method, derived from content based image retrieval. Because images often consist of different patches of uniform patterns, global features find it hard to capture all the properties. Initially, when it comes to the implemented method, about half of sample points is detected by a corner detector and another half is picked at random. For relatively small number of sample points this technique proved to give better results than sampling based purely on corner detectors, random generators, or dense sampling. Detector-based keypoints tend to concentrate on objects, which is good for object instance recognition but not necessarily good for generic image categorization. For characterization of local patches the regions around selected points are represented as four different statistics extracted from co-occurrence matrix, computed for individual colour channels. We also experiment with a different descriptor based on three colour moments, which is able

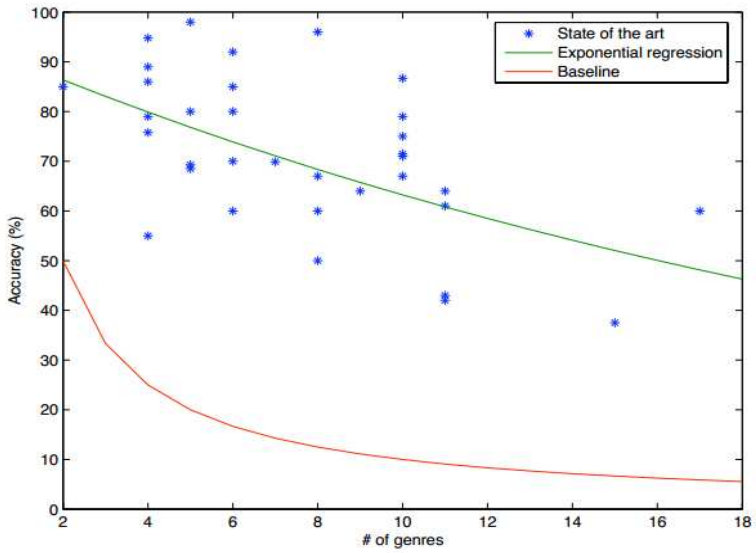

Fig. 3. State of the art in genre classification. Adapted from [16]

to capture local textural properties as well. Finally the k-means algorithm is applied to the training set to generate the visual dictionary and the images from the database are characterized with a histogram of visual words counts. Thus obtained histograms, characterizing individual songs, are used for genre music classification task.

We obtained classification accuracy of 46% (with 5% theoretical baseline for random classification) which is comparable with existing state-of-the-art approaches. Moreover, the novel features characterize different properties of the signal than standard methods. Therefore, the combination of them should further improve the performance of existing techniques.

The main advantages of our method are: more intuitive, easy way to automatic music classification, classification accuracy comparable with state-of-the-art and new promising research direction.

5 Future Work

The future work may include incorporation of the spatial information for local image patches, experimentation with different sampling techniques and incorporation of temporal information (short time Fourier transform, wavelets), which should further improve the classification accuracy. Additionally, an interesting future work would be to investigate if image patches identified by corner detectors and "visual words" corre-

spond to some important characteristics of audio signal. In other words, new specialized visual features can be developed for this particular task.

References

1. I.S.H. Suyoto, A.L. Uitdenbogerd, and F. Scholer. Searching musical audio using symbolic queries. *IEEE Transaction on Audio Speech and Language Processing*, 16(2):372–381, 2008.
2. N. Hu, R. Dannenberg, G. Tzanetakis. Polyphonic audio matching and alignment for music retrieval. *Proc. IEEE WASPAA, New Paltz, NY*, 2003.
3. N. Collins. Using a pitch detector for onset detection. *Proc. of ISMIR2005*, 100–106, 2005.
4. J. Foote, M. Cooper. Visualizing musical structure and rhythm via self-similarity. *Proceedings of the 2001 International Computer Music Conference*, Citeseer, 419–422, 2001.
5. J. Bello, L. Daudet, L. Abdallah, S. Duxbury, M. Davies, M. Sandler. A tutorial on onset detection in music signals. *IEEE Transaction on Speech and Audio Processing*, 13(5):1035, 2005.
6. D.G. Lowe. Object recognition from local scale-invariant features. *In Proceedings of the International Conference on Computer Vision*, 2:1150–1157, 1999.
7. K. Mikolajczyk, C. Schmidt. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1615–1630, 2005.
8. T. Lindeberg. Scale-space. *Encyclopedia of Computer Science and Engineering*, 4:2495–2504, 2009.
9. A.P.B. Lopes, S.E.F. De Avila, A.N.A. Peixoto, R.S. Oliveira, A.A. Araujo. A bag-of-features approach based on hue-SIFT descriptor for nude detection. *In Proceedings of the 17th European Signal Processing Conference, Glasgow, Scotland*, 2009.
10. E. Nowak, F. Jurie, B. Triggs. Sampling strategies for bag-of-features image classification. *Lecture Notes in Computer Science*, 3954(490), 2006.
11. J. Yang, Y.G. Jiang, A.G. Hauptmann, C.W. Ngo. Evaluating bag-of-visual-words representations in scene classification. *In Proceedings of the international workshop on Workshop on multimedia information retrieval*, 206, 2007.
12. H. Liu, D. Song, S. Ruger, R. Hu, V. Uren. Comparing dissimilarity measures for content-based image retrieval. *AIRS*, 44–51, 2008.
13. A. Meng, P. Ahrendt, J. Larsen, L.K. Hansen. Temporal feature integration for music genre classification. *IEEE Transactions on Audio, Speech and Language Processing*, 15(5):1654–1664, 2007.
14. J. Shi, C. Tomasi. Good features to track. *IEEE Conference on Computer Vision and Pattern Recognition*, 593–600, 1994.
15. L. Kaliciak, D. Song, N. Wiratunga, J. Pan. Novel local features with hybrid sampling technique for image retrieval. *Proceedings of Conference on Information and Knowledge Management (CIKM)*,1557–1560 , 2010.
16. X. Serra. Audio Content Processing for Automatic Music Genre Classification: Descriptors, Databases, and Classifiers. *Doctoral Dissertation*, 2009. *www.tesisenred.net/bitstream/handle/10803/7559/tegt.pdf?sequence=1*
17. D. Perrot, R. Gjerdigen. Scanning the Dial: An Exploration of Factors in Identification of Musical Style. *Proceedings of Social Music Perception Cognition*, 88, 1999.