CLASSIFICAÇÃO AUTOMÁTICA DE EMOÇÕES EM MÚSICA UTILIZANDO ESPECTROGRAMAS

Juliano Cézar Chagas Tavares (PIBIC/FA/Uem), Yandre Maldonado e Gomes da Costa (Orientador), e-mail: jcezarchagas@gmail.com

Universidade Estadual de Maringá / Centro de Tecnologia / Maringá, PR.

Área: Ciências Exatas e da Terra / Subárea: Ciência da Computação

Palavras-chave: classificação automática de músicas, reconhecimento de padrões, espectrogramas

Resumo:

Este trabalho tem como objetivo apresentar um sistema para classificação automática de emoções em música baseado em características visuais extraídas da mesma. Para tal, foram extraídas também características acústicas de modo a avaliar a eficácia das características separadamente e também quando utilizadas em conjunto com as visuais. As características visuais são obtidas a partir de espectrogramas e as acústicas diretamente do sinal de áudio. Os descritores de textura utilizados são *Local Phase Quantization* (LPQ), *Local Binary Pattern* (LBP) e *Robust Local Binary Pattern* (RLBP). As características acústicas são descritas utilizando *Rhythm Patterns* (RP), *Rhythm Histogram* (RH) e *Statistical Spectrum Descriptor* (SSD). Os experimentos realizados foram feitos sobre um subconjunto da *Latin Music Mood Database* (LMMD), considerando três classes de emoções: positiva, negativa e neutra. Na etapa de classificação, o classificador SVM foi usado e os resultados finais foram obtidos usando validação cruzada com 5 *folds*. A melhor taxa de acerto de classificação individual obtida corresponde à frequência de 1.700 Hz a 3.400 Hz, usando o descritor RLBP. Nesse caso a taxa de acerto *F-measure* obtida foi de 59,55%.

Introdução

A música não é utilizada apenas para o divertimento, ela pode ser utilizada também na formação da personalidade, para a cura de algumas doenças e alteração do humor. Graças aos avanços dos componentes eletrônicos é possível ter acesso às músicas em qualquer lugar de maneira rápida. Ouvir música já é um hábito na vida de muitas pessoas e buscar formas de melhorar a interação com a mesma é uma das motivações para o desenvolvimento deste trabalho. Nesse contexto, é possível justificar o desenvolvimento de pesquisas relacionadas ao reconhecimento automático de humor em músicas baseado no sinal de áudio.

Espectrogramas são representações do espectro das frequências do sinal de áudio e, a partir deles, é possível obter características visuais. Tem sua representação gráfica mais comum em duas dimensões onde o eixo horizontal corresponde ao tempo e o eixo vertical à frequência do espectro. Desde 2011, espectrogramas vêm sendo utilizados como uma fonte para a obtenção de características bastante úteis para trabalhos de classificação de áudio, inicialmente voltados para a classificação de gêneros musicais [1] e posteriormente voltado a outros domínios de aplicação [2].

Materiais e métodos

Base de Dados

A base de dados utilizada nesse trabalho é um recorte da *Latin Music Mood Database* (LMMD) [3]. A LMMD foi criada a partir da *Latin Music Database* (LMD) [4], que originalmente é uma base de músicas latinas classificadas em gêneros. A LMMD conta com 3136 amostras distribuídas em seis classes diferentes: amor, paixão, alegria, entusiasmo, tristeza e decepção. Essas classes ainda podem ser agrupadas duas a duas criando uma distribuição em três classes sendo elas: positivo (alegria e entusiasmo), neutro (amor e paixão) e negativo (tristeza e decepção). O recorte da base utilizado neste trabalho conta com 1005 amostras distribuídas no formato de três classes distintas explicado anteriormente. Esse formato foi utilizado por ser o mesmo utilizado no trabalho de Przybysz [5], o que permite uma comparação de resultados. A redução da quantidade de amostras neste caso, deu-se pelo fato de que não foi possível obter cifras e/ou letras (fontes para obtenção de características utilizadas naquele trabalho) para todas as músicas da LMMD original.

Descritores de Textura

Para a extração de características visuais dos espectrogramas foram utilizados os métodos *Local Binary Pattern* (LBP), *Robust Local Binary Pattern* (RLBP) e *Local Phase Quantization* (LPQ). O LBP extrai um padrão binário da imagem fazendo um processamento a partir de um pixel central C e comparando com seus P pixels vizinhos equidistantes a uma distância R. Além disso o LBP diferencia padrões uniformes e não-uniformes.

O RLBP é semelhante ao LBP, a diferença é que o RLBP trata os ruídos de maneira diferenciada. Tanto o LBP quanto o RLBP extraem um vetor de características composto por 59 elementos. Para tal a configuração dada a ambos foi uma vizinhança de 8 pixels a uma distância igual de 2 pixels cada a partir do ponto central.

O LPQ foi idealizado para operar sobre imagens com borramento, porém apresenta resultados satisfatórios em imagens que não possuem borramento também. Para a criação dos vetores de características a partir do LPQ se fez o uso de uma janela de dimensão 5 x 5 e ao se computar o método foi obtido um vetor de 256 elementos.

Descritores de Acústicos

Para a extração de características acústicas foram usados os métodos *Rhythm Pattern* (RP), *Rhythm Histogram* (RH) e *Statistical Spectrum Descriptor* (SSD). Todos os extratores anteriores trabalham diretamente no sinal de áudio, ou seja, sem o intermédio de uma imagem de descrição. Os vetores de características gerados a partir do SSD possuíam 168 elementos, do RH 60 e do RP 1440. Todos foram obtidos através da biblioteca *Rhythm and Timbre Feature Extraction from Music*.

Zoneamento da Imagem

Foram feitos zoneamentos lineares de três, cinco e dez zonas, com a finalidade de verificar qual zoneamento oferece o melhor desempenho nas taxas de acerto. A escala Mel foi aplicada criando-se classificadores exclusivos para cada uma das suas quinze zonas não lineares.

Esquema de Classificação

O esquema proposto consiste em: divisão das amostras de áudio em *folds*, geração de espectrogramas utilizando a ferramenta SoX, extração das características acústicas ou visuais, classificação utilizando a SVM e manipulação dos dados de saída da classificação utilizando técnicas de fusão. A classificação por meio do SVM foi realizada utilizando-se a biblioteca LIBSVM. Como a divisão da base foi realizada em cinco *folds* para a validação cruzada, um dos *folds* é tomado para teste e os *folds* restantes são utilizados como treino. O processo é repetido até que todos os *folds* tenham sido utilizados como teste.

Resultados e Discussão

Para as classificações zoneadas, foi realizado *late fusion* e, dentre os resultados gerados com as diferentes regras de fusão, foi escolhido o que apresentou a *F-measure* mais alta. A regra de fusão que apresentou os melhores resultados foi, na maioria dos casos, a regra da Soma. As características acústicas individualmente não tiveram taxas de acerto muito relevantes em relação às taxas das características visuais. RP, possivelmente pela sua alta quantidade de características, se saiu um pouco melhor do que os outros classificadores. Nas classificações iniciais o RLBP atingiu a maior taxa de acerto, em *F-measure*, com 55,06% seguido do LBP e do LPQ respectivamente com 54,71% e 51,32%.

Nenhuma das fusões unimodais superou o resultado obtido anteriormente pelo RLBP. Na fusão de predições visuais o maior acerto, em *F-measure*, ficou com a fusão de LBP com LPQ, alcançando uma taxa de acerto de 46,96%. Já na fusão de predições acústicas o maior acerto ficou para a fusão de RH e SSD, atingindo uma taxa de acerto de 44,12%.

Ainda foram realizados experimentos com uma abordagem multimodal. Foram utilizados na fusão os melhores classificadores visuais combinados com os três classificadores acústicos. Nenhum dos resultados superou o resultado inicialmente alcançado por RLBP.

Por fim, ao avaliar que as fusões não ajudaram no aumento das taxas de acerto, foi realizada uma análise dos resultados das zonas lineares dos classificadores individualmente. Foi encontrada a zona 4, que corresponde à faixa de frequências no intervalo entre 1700 Hz e 3400 Hz, do zoneamento linear em 5 zonas do RLBP. Essa zona obteve uma taxa de acerto de 59,55%, superando a taxa de acerto tida pelo RLBP desde o começo dos testes.

Conclusões

Tendo por base as explicações apresentadas acima, é possível mostrar que as características visuais podem ser utilizadas no auxílio à classificação de emoções em música, porém quando combinadas com descritores acústicos, não houve ganho nas taxas de acerto, possivelmente porque as taxas piores dos classificadores acústicos prejudicaram o desempenho da fusão. Uma boa hipótese de pesquisa para trabalho futuro é a de que a combinação de características visuais e/ou acústicas com outras obtidas a partir de meta-dados da música (como letra ou cifra) permitam que se alcance melhor desempenho nessa tarefa de classificação.

Agradecimentos

Ao programa CNPq/Fundação Araucária/PIBIC pelo financiamento do projeto de pesquisa e à UEM pela concessão da bolsa de IC.

Referências

[1] COSTA, Y. M. G.; OLIVEIRA, L. E. S.; KOERICH, A. L.; GOUYON, F., **Music genre recognition using spectrograms**, in 18th International Conference on Systems, Signals and Image Processing, Sarajevo, 2011.

[2] MONTALVO, A.; COSTA, Y. M. G.; CALVO, J. R**., Language Identification using Spectrogram Texture**. in Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, 2015. p. 543-550.

[3] SANTOS, C. L.; SILLA JR., C. N., **The Latin Music Mood Database**, in EURASIP Journal on Audio Speech and Music Processing, 2015.

[4] SILLA JR., C. N.; KOERICH, A. L.; KAESTNER, C. A. A., **The latin music database**, in ISMIR 2008, 9th International Conference on Music Information Retrieval, 2008, pp. 451–456.

[5] PRZYBYSZ, A. L., **Classificação automática de emoções em músicas latinas utilizando diferentes fontes de informação**. in Dissertação - Programa de Pós-Graduação em Informática, Universidade Federal do Paraná, Cornélio Procópio, 2016.