CLASSIFICAÇÃO DE CENÁRIO ACÚSTICO UTILIZANDO ESPECTROGRAMAS

Gustavo Zanoni Felipe (PIC/UEM), Yandre Maldonado e Gomes da Costa (Orientador), e-mail: yandre@din.uem.br

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

Ciências Exatas e da Terra / Ciência da Computação

Palavras-chave: classificação automática, espectrograma, reconhecimento de padrões

Resumo:

Este trabalho objetivou investigar o desempenho de um sistema de classificação automático utilizando a base do desafio DCASE 2016, que reconhece ambientes/locais em que uma amostra de áudio foi colhida, considerando 15 possíveis categorias distintas. O sinal de áudio foi convertido para espectrogramas e posteriormente foi feita a extração de características utilizando-se descritores de textura. Foi investigada a complementaridade entre os sinais que existem nos canais esquerdo e direito das amostras de áudio digital gravadas em estéreo. Foram também testadas combinações de características acústicas e visuais, em busca de melhorar a taxa de classificação. Ao final, foi alcançada uma acurácia de 80,17%, superando o baseline originalmente previsto pelo próprio desafio DCASE 2016, que apresentava acurácia de 77,2%. Concluiu-se que a classificação utilizando de características extraídas no domínio visual, foi tão eficiente quanto sistemas de classificação que utilizam de características acústicas.

Introdução

Para nós humanos, ouvir o som de um ambiente e identificar em que tipo de ambiente ele foi gravado, é uma tarefa fácil. Porém, para as máquinas, é diferente. Pensando nisso, surgiu a tarefa de classificação automática de cenário acústico, que possui o objetivo de identificar o lugar/ambiente em que uma determinada amostra de gravação de áudio foi originalmente gravada. Com isso, novas pesquisas e experimentos com a classificação de cenário acústico começam a aparecer.

Recentemente, diversos trabalhos apontaram métodos que caracterizam tarefas de classificação automática de som. Em (COSTA, 2013) foi avaliado a performance de características obtidas da textura de uma imagem tempo-frequência de som, também conhecida por espectrograma, na classificação de gêneros musicais. Neste trabalho, o objetivo é classificar cenários

acústicos tendo como base as características extraídas de espectrogramas gerados a partir de sinais de áudio. O trabalho também investiga a eventual diferença entre os canais esquerdo e direito das amostras de áudio gravadas em estéreo, além de averiguar a complementaridade entre os mesmos e entre as características extraídas de fontes visuais e acústicas.

Materiais e métodos

Base de Dados

Para a realização deste projeto, foi utilizada a base de dados disponibilizada publicamente pelo desafio DCASE 2016 (em *http://www.cs.tut.fi/sgn/arg/dcase2016/task-acoustic-scene-classification*). A base é composta por 1170 arquivos de áudio, que foram originalmente gravados em diferentes lugares da Finlândia. Estes foram gravados em uma frequência de 44,1kHz e foram segmentados para que seguissem o padrão de 30 segundos de duração. As amostras de áudio da base são rotuladas em 15 diferentes classes: ônibus, cafeteria/restaurante, carro, centro da cidade, floresta, loja de conveniência, casa, orla do rio, livraria, estação de metrô, escritório, área residencial, trem, bonde e parque urbano. Além disso, é pré-organizada em 4 folds, criados para a competição do desafio DCASE 2016.

Métodos

A metodologia utilizada neste trabalho, é baseada em utilizar espectrogramas gerados a partir dos sinais de amostras de áudios, seguidos pela extração de características por descritores de textura, classificação e fusão de classificadores para a obtenção da decisão final. Espectrogramas, são uma representação visual de um áudio, onde o eixo “x” representa a frequência em Hertz e o eixo “y” representa a duração em segundos. Para gerar os espectrogramas, foi utilizada a ferramenta *Sound EXchange* (SOX), pois a mesma permite a variação de diversos parâmetros (como por exemplo limite inferior da amplitude, frequência, etc.), além de possibilitar diversas outras operações, como mudar a cor dos espectrogramas para cinza e somar canais de áudio. Pode-se trabalhar com a extração de características dos espectrogramas como um todo, chamado de método “Global”, e também, extração de modo local, para que seja possível preservar as características locais de um espectrograma. Dentre as possibilidades de Extração Local, têm-se: Segmentação, Zoneamento e a conhecida Escala Mel, onde 15 zonas, posicionadas de um modo pré-definido, dividem o espectrogama. Neste caso, cada Zona ou Segmento é tratado como um espectrograma diferente, o que possibilita que para cada um seja criado um classificador específico. Para extrair as características, foram utilizados conhecidos Descritores de Textura, dentre eles, foram utilizados o *Local Binary Pattern* (LBP), *Local Phase Quantisation* (LPQ) e *Robust Local Binary Pattern* (RLBP). Além destes, foram utilizados também

os Descritores Acústicos: *Rhythm Patterns* (RP), *Statistical Spectrum Descriptor* (SSD) e *Rhythm Histogram* (RH). Descritores deste tipo operam sobre o áudio em si e foram utilizados para testar a complementaridade quanto aos Descritores de Textura. A classificação foi realizada através de uma biblioteca que implementa o algoritmo *Support Vector Machine* (SVM), chamada de LIBSVM, onde a mesma nos retorna uma estimativa de probabilidade de acerto para cada classe. As predições podem ser combinadas através de diferentes regras (como Soma, Multiplicação, etc.) e a partir disso, foram combinados de diferentes formas descritores de textura e descritores acústicos, a fim de que a fusão entre informações complementares existentes possibilitassem melhores resultados.

Resultados e Discussão

Em um primeiro momento, foram realizados testes utilizando-se da Extração Global, com a finalidade de encontrar parâmetros que mais se adequassem a este problema. Com isso, concluiu-se que para a obtenção de melhores resultados, seriam utilizados a frequência 44,1 kHz e limite da amplitude inferior de -150 dB como parâmetros padrões para os demais testes. Vale ressaltar também, que dentre os Descritores de Textura utilizados, o que obteve melhor êxito na tarefa da classificação foi o LBP (OJALA et al., 2002), assim como mostrado na Tabela 1. Após os testes, pode ser observado que os canais comportavam-se de maneiras semelhantes.

Tabela 1 – Resultados iniciais, onde os testes variavam parâmetros ao se gerar o espectrograma e utilizavam de diferentes Descritores de Textura.

Frequência	Amplitude	Descritores	Acurácia CE*	Acurácia CD*
44,1 kHz	-150 dB	LBP	69,8%	70,5%
44,1 kHz	-150 dB	RLBP	68%	70,4%
44,1 kHz	-150 dB	LPQ	65,8%	65,6%

Notas – As siglas CE e CD referem-se à “Canal Esquerdo” e “Canal Direito” respectivamente.

Procurando por melhores resultados, foram testadas extrações locais e também, diferentes combinações entre descritores e canais. Com isso, pode-se observar que ao realizar a fusão entre os canais esquerdo e direito, houve uma melhoria nos resultados, havendo alguma complementaridade entre os mesmos. Junto a isso, foram combinados os Descritores de Textura aos Descritores Acústicos. Nesta situação, ao combinar o descritor LBP com o descritor RP, obteve-se os melhores resultados, mostrados na Tabela 2.

Tabela 2 – Melhores resultados, alcançados a partir de fusão de classificadores obtidos da combinação de Descritores de Textura e Acústicos.

Descritores	Regra	*Precision*	*F-measure*	Acurácia
LBP + RP	Produto	70,17%	70,45%	80,17%
LBP + SSD	Produto	70,26%	70,31%	79,74%
LBP + RP	Soma	68,64%	68,97%	78,55%

LBP + RLBP	Produto	64,63%	64,93%	73,93%

Os resultados alcançados neste trabalho, superaram o resultado apresentado como baseline do próprio desafio DCASE 2016, o qual mostrava acurácia de 77,2%.
O valor do limite inferior da amplitude utilizado neste trabalho é muito diferente dos valores comumente utilizados em trabalhos realizados em outros domínios de aplicação. Provavelmente isso se explique pelo fato de que boa parte do conteúdo relevante do sinal, e importante para o processo de classificação, se encontre em amplitudes muito baixas. Em testes envolvendo valores mais comuns como -60 dB, -80 dB e outros, foram obtidos resultados ruins. Mesmo utilizando um valor de frequência mais baixo, como por exemplo 8 kHz (região do espectrograma onde o sinal seria mais intenso), os resultados não foram completamente satisfatórios.

Conclusões

O método de classificação automático utilizando espectrogramas é relativamente novo. Através deste trabalho, foi mostrado que a classificação utilizando de características visuais para a tarefa da Classificação de Cenário Acústico, obteve resultados melhores do que os inicialmente propostos pelo desafio DCASE 2016. O melhor resultado encontrado neste trabalho alcançou acurácia de 80,17%, enquanto o baseline do desafio proposto era de 77,2%. Onde o mesmo utilizava de um sistema de classificação automático que extrai características diretamente do sinal do áudio. Concluindo, a fusão de diferentes classificadores foi essencial para a obtenção de melhores acurácias e a combinação de descritores de textura com descritores acústicos, evidenciam a complementaridade de características do domínio acústico e visual. Além da complementaridade entre os sinais retirados dos canais esquerdo e direito das amostras de áudio.

Agradecimentos

Ao professor Yandre Maldonado e Gomes da Costa pela orientação e apoio proporcionados e ao programa PIC-UEM pela oportunidade proporcionada.

Referências

COSTA, Y. M. G. **Reconhecimento de Gêneros Musicais utilizando Espectrogramas com Combinação de Classificadores.** 2013, 106f. Tese (Doutorado) – Programa de Pós-Graduação em Informática, Universidade Federal do Paraná , Curitiba, 2013.

OJALA, T.; et al. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. **IEEE Transactions on Pattern Analysis and Machine Intelligence,** v. 24, n. 7, p. 971-987, 2002.