

# O Uso do Descritor HOG no Problema da Classificação de Cenas Acústicas sobre o Domínio Wavelet

Lucas Angelo Mattesco<sup>1</sup>, Caio Cesar Enside de Abreu<sup>2</sup>

<sup>1,2</sup>Universidade do Estado do Mato Grosso - UNEMAT  
Rua Santa Rita, 128 – Centro – Alto Araguaia – MT – Brazil

lucas.mattesco@unemat.br, caioenside@unemat.br

***Abstract.** The problem of Acoustic Scene Classification (ASC) consists in analyzing and assigning semantic labels to a determined audio recording, through algorithms, in order to classify the acoustic scene where the audio signal was recorded according to a finite number of classes. The application of Histogram of Oriented Gradients descriptor (HOG), paired with Wavelet analysis is the focus of our investigation.*

***Resumo.** O problema de Classificação de Cenas Acústicas (CCA) consiste em analisar e atribuir um rótulo semântico a uma determinada gravação de áudio, por meio de algoritmos, de forma a classificar o cenário acústico onde o sinal de áudio foi capturado de acordo com um número finito de classes. A aplicação do descritor Histograma de Gradientes Orientados (HOG), em conjunto com a análise Wavelet é o foco da nossa investigação.*

## 1. Introdução

O problema de Classificação de Cenas Acústicas é tratado em diversas pesquisas e em eventos como o IEEE AASP Challenge Detection and Classification of Acoustic Scenes and Events (DCASE). A CCA consiste em analisar e atribuir um rótulo semântico a uma determinada gravação de áudio, por meio de algoritmos, para classificar o cenário acústico onde foi gravado, de acordo com um número finito de possíveis classes.

As transformadas integrais são ferramentas utilizadas para a análise desses sinais, em especial a Transformada de Fourier (TF) e a Transformada Wavelet (TW). As transformadas são usadas para analisar uma função, por analogia um sinal, em outro domínio. A TF transforma um sinal no domínio do tempo para o domínio da frequência, sem mudar a informação contida. Já a TW fornece uma representação do sinal no domínio tempo-frequência [Daubechies 1992; Oppenheim et al. 1998].

Os descritores amplamente utilizados pela literatura para a tarefa de CCA são os coeficientes mel-cepstrais, que são calculados tendo como base a TF [Nogueira et al. 2013]. No entanto, pesquisas recentes tem mostrado que a utilização do descritor Histograma de Gradientes Orientados (HOG) na CCA pode ser promissor [Rakotomamonjy et al. 2013]. Para isso, os pesquisadores utilizam o espectrograma como uma imagem digital, para então extrair os atributos HOG. Em tais metodologias, porém, o espectrograma é gerado utilizando a TF, ou algumas de suas variações. Dessa forma, o objetivo deste trabalho é investigar a utilização do descritor HOG em conjunto com a análise Wavelet, configurando uma abordagem inovadora dentro da área.

## 2. Material e Métodos

Para a realização deste trabalho, propõe-se utilizar o escalograma wavelet, gerado pela TW em sua versão contínua, em conjunto com o descritor HOG. A Transformada Wavelet Contínua (TWC) é uma ferramenta que representa um sinal no domínio do tempo e da frequência simultaneamente, sendo que estes parâmetros variam de forma contínua. A TWC de uma função  $x(t)$  em escala  $a > 0$  ( $a \in \mathbb{R}^{+*}$ ) e tempo  $b \in \mathbb{R}$ , é expressada por:

$$X_w(a, b) = \frac{1}{|a|^{\frac{1}{2}}} \int_{-\infty}^{+\infty} x(t) \bar{\psi}\left(\frac{t-b}{a}\right) dt. \quad (1)$$

A fim de trabalhar com a equação (1) em um ambiente computacional, é necessário realizar a discretização dos parâmetros  $a$  e  $b$ . O escalograma é definido como  $|X_w(a, b)|^2$  [Daubechies et al. 1992].

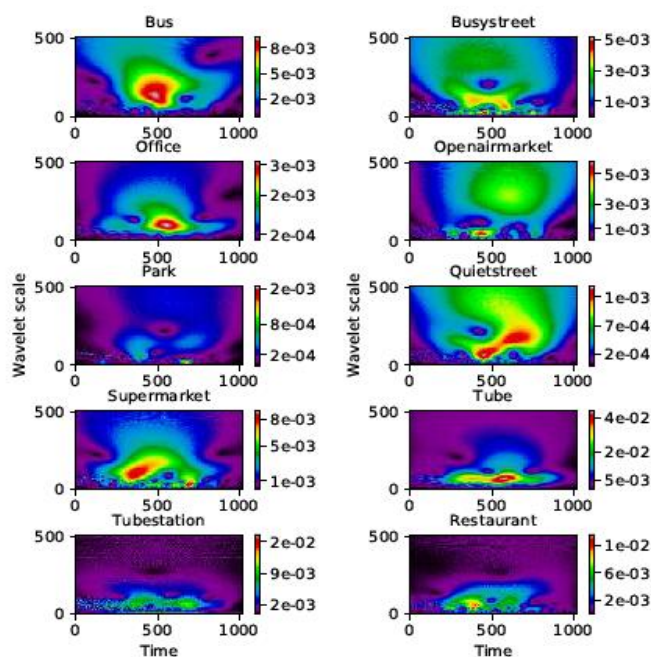
O descritor HOG é comumente utilizado em problemas de visão computacional, mais especificamente em aplicações cujo objetivo é a detecção de objetos. Este descritor se baseia na intensidade e orientação do gradiente calculado em pequenas porções da imagem. Dessa forma, propõe-se utilizar o escalograma wavelet como uma imagem digital, para extrair os atributos de acordo com a distribuição da intensidade dos gradientes. Em seguida, esses atributos irão alimentar um algoritmo de aprendizagem de máquinas que atribuirá um rótulo a uma determinada gravação de áudio.

Para as simulações, é utilizado o mesmo banco de dados utilizado no evento DCASE-2013, que consiste em diversas cenas acústicas gravadas em Londres, em locais e horários diferentes para evitar a classificação sistemática. As gravações podem ser classificadas em 10 classes: ônibus, rua movimentada, escritório, feira, parque, rua silenciosa, restaurante, supermercado, metrô e estação de metrô. Cada cena contém 10 sinais de áudio, totalizando 100 gravações. As implementações estão sendo desenvolvidas em linguagem de programação Python, e a classificação é feita por meio de uma Máquina de Vetores de Suporte (classificador SVM), que utiliza a aprendizagem supervisionada.

Para cada sinal de áudio em processamento, o seguinte procedimento é realizado: carrega-se o arquivo de áudio e realiza-se um janelamento de Hanning com 1024 pontos e 50% de sobreposição; para cada terça parte do sinal, a média aritmética de todas as janelas são calculadas; aplica-se a TWC sobre cada uma das três janelas resultantes e três escalogramas wavelet são gerados; em seguida, as características são extraídas pelo descritor HOG e três classificações são emitidas utilizando o classificador SVM; o rótulo atribuído ao sinal de áudio corresponde a classe com maior número de votos. A fim de evitar overfitting do classificador, utiliza-se a validação cruzada com cinco pastas. Para cada pasta, oito sinais de cada classe são utilizados para treino e dois para teste. Ao final da execução da validação cruzada, todos os sinais de cada classe serão classificados.

## 3. Resultados e Discussões

Para uma maior compreensão sobre o escalograma wavelet, e como ele pode ser utilizado para a CCA, a Figura 1 apresenta um escalograma para cada uma das 10 cenas acústicas presentes na base de dados DCASE-2013.



**Figura 1. Escalograma wavelet obtido para 10 cenas acústicas diferentes.**

Para gerar a matriz dos escalogramas na Figura 1, a função wavelet complexa de Morlet foi utilizada e definiu-se  $a = 1, 2, \dots, 500$  (500 escalas). Após a execução do procedimento descrito na Seção 2, uma acurácia de 43% foi alcançada. Quando comparado com o sistema baseline fornecido pelos organizadores do evento DCASE-2013, que alcançou 52% de acurácia, nota-se que o método proposto alcançou resultados competitivos. Vale destacar que os resultados alcançados são resultados iniciais.

#### 4. Conclusão

Neste trabalho foi investigada a aplicação do descritor HOG sobre o escalograma wavelet. Para isso, foi proposta uma nova interpretação para o escalograma, onde este passa a ser visto como uma imagem digital. Dessa forma, estruturas de tempo-frequência inerentes a diferentes cenas acústicas puderam ser capturadas.

Nos eventos DCASE-2013 e 2016, verificou-se que apenas dois dos 46 sistemas submetidos utilizam a teoria wavelet. Este fato reforça a necessidade de investigação na área, visto que as wavelets têm potencial para utilização em diferentes linhas de pesquisa.

#### Referências

- Daubechies, I. Ten Lectures on Wavelets, SIAM, Philadelphia, 1992.
- Oppenheim, A. V.; Schafer, R. W.; Buck, J. R. Discrete-time signal processing. New Jersey: Prentice Hall, 1998.
- Rakotomamonjy, A.; Gasso, G. "Histogram of gradients of time–frequency representations for audio scene classification", IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 23, n. 1, p.142-153, 2013.
- Nogueira, W.; Roma, G.; Herrera, P.; "Sound scene identification based on MFCC, binaural features and a support vector machine (SVM) classifier", IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events, 2013.