2022

4(2), 128-142

VarioGram – Zaman Serileri İçin Renkli Bir Zaman-Graf Temsili

İlker Türker[1], Serkan Aksu*[2]

Anahtar Sözcükler

Graf temsili
Ses sınıflandırma
Zaman serilerinin sınıflandırılması
Karmaşık ağlar

Makale Hakkında

Gönderim Tarihi
20 Eylül 2022
Kabul Tarihi
03 Kasım 2022
Yayın Tarihi
28 Aralık 2022
Makale Türü
Araştırma Makalesi

Öz

Bu çalışmada zaman serilerinin ağ tabanlı temsili için bir çerçeve sunulmuştur. Önerilen yöntemde öncelikle, zaman domenindeki sinyaller %50 örtüşmeli sabit genişlikli zaman pencerelerine bölünerek segmentasyon işlemi tamamlanır. Her segment, ana sinyalin mutlak maksimum genlik değerinin ve negatif karşılığının tanımladığı aralık baz alınarak normalize edilir ve normalize sinyaller 2^n seviyesine kuantize edilir. 3 farklı atlama değerinin ifade ettiği 3 kanaldan ilerleyen bu dönüşüm, kanalların katmanlar şeklinde birleştirilmesiyle düşey bir RGB görüntü temsilini oluşturur. Sinyalin her zaman penceresinden elde edilen bu düşey RGB imajlarının yan yana döşenmesinin sonucunda yatay eksenin zamanı ve düşey eksenin sinyal dalgalanmalarını temsil ettiği *VarioGram* olarak adlandırılan bir zaman-graf temsili elde edilmiş olur. Çevresel ses sınıflandırma problemlerinde sıklıkla kullanılan ESC-10 veri setindeki ses sinyallerinin dönüşümü ile elde edilen *VarioGram* temsilleri bir ResNet modeline girdi olarak verildiğinde %82.08'lik bir sınıflandırma başarısı elde edilmiş, mel-spectrogram görüntüleri ile hibritleştirilerek kullanılan *VarioGram* temsilleri ile bu başarı %93.33'e kadar çıkarılmıştır. Dolayısıyla VarioGram temsilleri, tek başına mel-spectrogram ile elde edilebilen en yüksek sınıflandırma başarısını küçük bir farkla iyileştirme yönünde etki yapmıştır.

VarioGram – A Colorful Time-Graph Representation For Time Series

Keywords

Graph representation
Sound classification
Time-series classification
Complex networks

Article Info

Received
September 20, 2022
Accepted
November 03, 2022
Published
December 28, 2022
Article Type
Research Paper

Abstract

In this study, a framework for network-based representation of time series is presented. In the proposed method, initially, a segmentation procedure is completed by dividing the signals in the time domain into fixed-width time windows with 50% overlap. Each segment is normalized based on the range defined by the absolute maximum amplitude value of the main signal and its negative counterpart, and the normalized signals are quantized to 2^n levels. This transformation, proceeding through 3 channels expressed by 3 different jump values, generates a vertical RGB image representation by combining the channels in layers. As a result of tiling these vertical RGB images from each time window horizontally, a time-graph representation called VarioGram is obtained, where the horizontal axis represents time, and the vertical axis represents signal fluctuations. Feeding a ResNet model with VarioGram representations obtained by the transformation of the audio signals in the ESC-10 dataset which is frequently used in environmental sound classification problems, a classification success of 82.08% has been obtained, while this success has been 93.33% with the VarioGram representations hybridized with mel-spectrogram images. The VarioGram representations therefore acted to slightly improve the highest classification success achievable with the mel-spectrogram alone.

Atıf: Türker, İ. & Aksu, S. (2022). VarioGram – Zaman serileri için renkli bir zaman-graf temsili. *Bilgi ve İletişim Teknolojileri Dergisi*, *4*(2), 128-143. https://doi.org/ 10.53694/bited.1177504

Cite: Türker, İ. & Aksu, S. (2022). VarioGram – A colorful time-graph representation for time series. *Journal of Information and Communication Technologies*, *4*(2), 128-143. https://doi.org/ 10.53694/bited.1177504

* **Sorumlu Yazar/Corresponding Author:** aksu@bartin.edu.tr

[1] Assoc. Prof. Dr., Karabuk University, Faculty Of Engineering, Karabuk/Turkey, iturker@karabuk.edu.tr, https://orcid.org/0000-0001-7577-4658

[2] PhD., Bartın University, Vocational School, Bartin/Turkey, aksu@bartin.edu.tr, https://orcid.org/ 0000-0001-6920-7219

Introduction

Scientific interest on time series classification and representation techniques have increased together with the increasing availability of temporal datasets from diverse domains such as medicine, finance, aviation, telecommunication, industry, weather, multimedia etc. (Bao et al., 2017; Canizo et al., 2019; Dafna et al., 2018; Gharehbaghi & Lindén, 2017; Pourbabaee et al., 2018; Soares et al., 2018). These available datasets facilitate several data-driven tasks including classification, clustering, segmentation, anomaly detection and so on (Ares et al., 2016; Canizo et al., 2019; Kanani & Padole, 2020; Ruiz et al., 2021). The main motivation for developing different representation techniques is reducing the complexity mostly caused by the high dimensionality of the data. Additionally, while converting time series data into diverse representations, scientists focus on improving classification accuracy, removing noise and accelerating the overall task (Deng et al., 2016).

Recently, sparse representation has been an attractive topic for time series, for which a test sample is represented as a sparse combination of training samples. The solution of the sparse representation problem further gives sparse representation coefficients and the test sample is assigned to the class minimizing the residual between itself and the reconstruction from the training samples of its class (Chen et al., 2015; Yin et al., 2012).

A mainstream of representation techniques involve in transformation domain methods including discrete Fourier transform (DFT) and discrete wavelet transform (DWT), Karhunen-Loeve (KL) transform or Singular Value Decomposition (SVD), either mapping the time sequences to a new feature space of a lower dimensionality, or providing a multi-resolution representation with time-frequency localization property (Chan & Fu, 1999). Transformation domain methods further evolved into image-based representations like mel-spectrograms especially used for sound signals, as a robust representation to noise content compared to mel-frequency cepstral coefficients (MFCCs), the coefficients capturing the envelope of the short time power spectrum. The spectrogram image is simply generated by applying discrete Fourier transform (DFT) to the fixed-size segments of the original sound, further finalized with a mel-filter provided from MFCCs (Sharan & Moir, 2015). A similar time-frequency image representation known as Cochleagram handles the frequency band in logarithmic fashion similar to human cochlea. By the way, it includes more frequency components in the lower and less frequency components in the higher frequency range (Peng et al., 2021).

An alternative framework for representing time series in a more structured format is powered by network science, which proposes that each complex interconnected system can be represented as a network (Baydilli et al., 2017; Demir & Türker, 2021; Türker et al., 2016; Türker & Sulak, 2018). Time series are converted into graph representations, which evaluate quantized amplitude levels as nodes and neighborhood between consecutive levels as edges between them (Lacasa et al., 2015). A well-known member of this convention is *visibility graph*, constructed by keeping the convexity information of sequential samples by establishing connections between quantized amplitude levels if they are visible from the top of another level in a pre-defined neighborhood (Lacasa et al., 2008). The resulting graph is a mean-adjacency matrix calculated over a whole time series, that can be assessed a stationary representation. Türker and Aksu achieved 2-3% more successful results than mel-spectrogram based methods in ESC-10 dataset classification with the Connectogram method, which they developed based on the graph representation of time series (Türker & Aksu, 2022).

The main motivation of this study is to develop a graph-inspired representation for time series data which can capture the convexity properties of a given signal, also capturing the time-dependent deviations in this graph representation. Instead of capturing edge formations between consecutive signal levels (nodes), we focus on differences between these nodes and encode them to a vertical array calculated from each frame of the segmented signal. These vertical arrays are then concatenated horizontally to form a time-graph representation, which in turn captures both signal envelope in its shape and power of alternance in the color levels. The color info is provided by applying the same procedure for three different subsampling rates, each of which forms a single layer of a resulting RGB image. The details of the conversion procedure together with its classification performance is presented in the coming sections.

Method

Data and preprocessing

In this study, we focus on an Environmental Sound Classification (ESC) task, dealing with recognizing some classes of sounds from the real environment. As a subfield of audio processing, ESC is a complex task involving classification of several sound events as one of the predefined sounds such as helicopter, sea waves, crying baby, crackling fire etc. (Zhang Zhichao and Xu, 2018). The most commonly used datasets in ESC tasks are known as ESC-10, ESC-50 (Piczak, 2015b) and UrbanSound8k (Salamon et al., 2014) datasets. The bigger ESC-50 dataset is a multiclass source consisting of 2000 separate environmental sound excerpts subdivided into 50 classes and 5 major categories. A filtered subset of ESC-50 is known as ESC-10, that is reduced to 400 samples consisting of 5-second records from 10 classes (Zhang Zhichao and Xu, 2018). Having a mono-channel signal form, environmental sound records can be assessed as a one-dimensional time series data which naturally holds both time and frequency components. We pick the ESC-10 dataset to apply the proposed conversion method. The steps of preprocessing are detailed in the next subsection.

Before the proposed method is applied, a unified preprocessing procedure is applied to all samples. We sampled the sound waves at 22050 Hz, filtered out components below 60dB and min-max normalized the signals into [-1, 1] range. An augmentation procedure is run in signal level first, applying time stretch, pitch shift, dynamic range compression, white noise addition to the raw sounds, augmenting the data size 6 times, to 2400 samples. A second augmentation procedure will also be applied through the TensorFlow library in Python environment, including image distortion methods (rotation, horizontal and vertical shift, brightness, shear, zoom) to the image representation of the sounds (Mushtaq et al., 2021).

After the signal-level preprocessing steps are applied, we apply a segmentation procedure subdividing the whole samples into frames with a window size of 1024 samples and hop-length of 512 samples, corresponding to 50% overlap between consecutive frames. Alternatively, it is possible to give a predefined number of frames to the function, that would result in VarioGrams of desired width and height, employing a different overlap percentage. Sound processing tasks are performed via Librosa library in Python environment, while all classifier models are trained using Keras library with TensorFlow backend on Google Colab environment.

Converting Sound Waves Into *VarioGrams*

We handle each frame including 1024 samples separately, performing a graph-based conversion into a vertical array demonstrating the varying patterns between consecutive signal levels.

- We first fix a *bit-depth* to quantize the signal, for which we employ a default value of 6, resulting a node count of $2^6 = 64$ signal levels (*scale* parameter). This means we further map the normalized amplitude range [-1,1] into [0, 63]. By the way, we achieve signal arrays including integers within this interval, that is ready for generating *variance arrays* (a term used for representing the varying rates, not statistical variance).
- A *variance array* is a resulting array with height $2 \times 2^6 = 2^7 = 128$, width 1, and depth 3.
    - *Height:* Since difference between neighboring amplitude levels are encoded, these differences can take values in range $[0 - 63]\ to\ [-63 - 0] \rightarrow [-63,63]$. Therefore, a size of 2^7 is needed regarding this scale.
    - *Width:* This array calculated from a separate sound segment, will be represented as a 1 pixel-width sequence, which are to be tiled horizontally to form the resulting *VarioGram*.
    - *Depth:* The conversion procedure will be held for 3 different subsampling rates to the original signal, resulting in 3 different $2^7 \times 1$ arrays, those are evaluated as RGB channels from a single variance array.
- Each *variance array* is calculated with 3 different subsampling rates given in a *jump* parameter holding default values as: [3,5,7]. The first value 3 means that a subsampling rate of 3 is applied to the original segment, before the neighborhood relations are encoded into the variance array. These subsampling rates are used to calculate R, G and B layers of the colored *variance array*.
- After each subsampling, the differences between neighboring amplitude levels in proceeding order within the range $\pm 2^n$, are further added with a constant 2^n to achieve nonnegative differences, and the corresponding element of the variance array of size $2^{n+1} \times 1 \times 3$ is increased by 1. For example, if the sequence is like […47, 33…], a difference of $33 - 47 = -14$ is calculated first. Then, it is shifted by the *scale* parameter as $-14 + 64 = 50$. If we are encoding for the first element of the jump parameter, i.e. 3, we increase the element of the *variance array* with index [50, 0, 0] by 1.
- After processing each frame, all variance arrays are normalized to 0-1 interval.
- This procedure is run for each jump layer (3, 5 and 7; corresponding the resulting RGB layers) to achieve a variance array of size $2^{n+1} \times 1 \times 3$.
- Also applied for each sound frame, this procedure produces vertical RGB arrays as much as the count of consecutive frames, further tiled horizontally to form time-variance images named as *VarioGram*. The complete procedure is illustrated in Fig. 1.

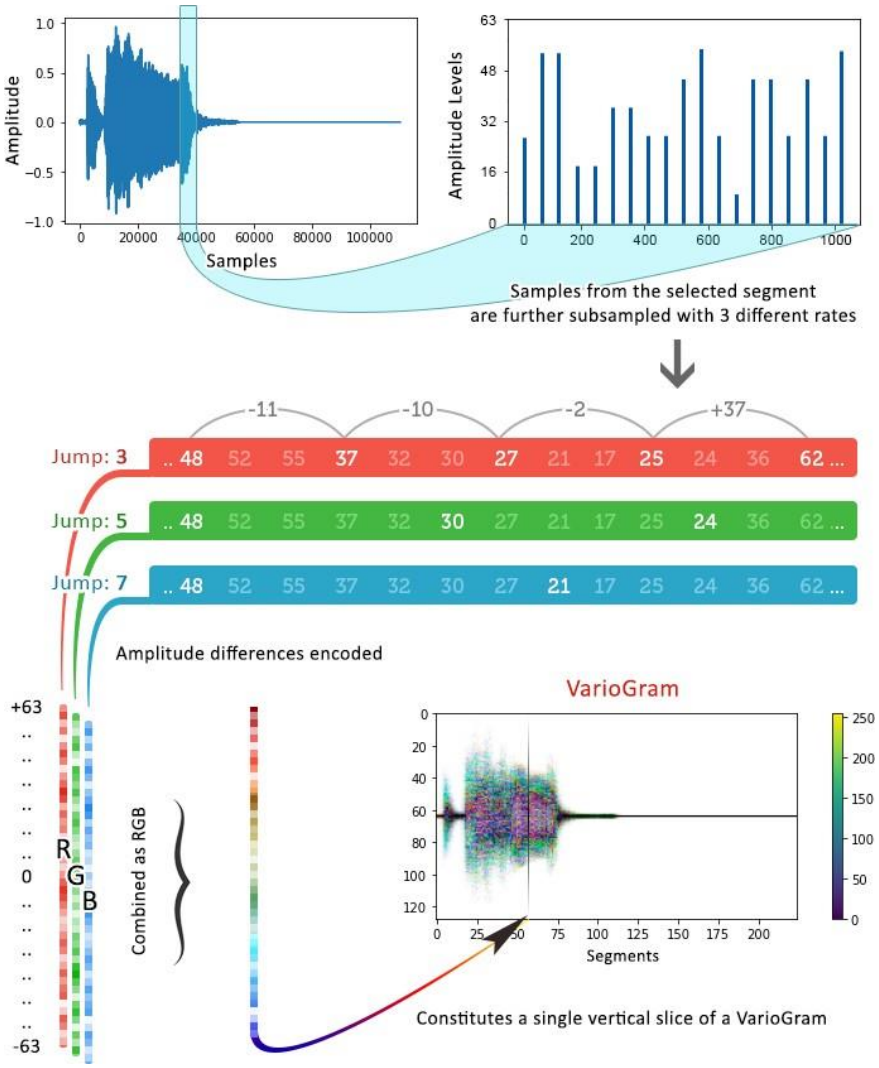

Figure 1. Procedure followed to form a *VarioGram* is illustrated. A sequence is first sampled, quantized and subdivided into frames. Each frame is then subsampled with 3 different rates. For each subsampling rate applied, differences between consecutive amplitude levels are encoded into a vertical array as +1. These vertical arrays are further min-max normalized to 0-1 range, and combined as a RGB-slice representing a single frame. Vertical RGB slices derived from each frame are then tiled horizontally to form a *VarioGram*.

The resulting *VarioGram* images captures both time (horizontal) and convexity (vertical) information of the corresponding time series. Optimal values of the parameters used in conversion procedure are given in Table 1. Having applied the VarioGram converter function to the sound samples in ESC-10 dataset, a set of image representations are generated. Sample VarioGram images together with the original sound plot and mel-spectrogram images are given in Fig. 2 to provide visual comparison to the reader.

As presented in Fig. 2, *VarioGram* representations resemble the shape of the original sound, holding the amplitude information. It also captures signal convexity information as the color range encoded inside the signal shape. By the way, a time-convexity based representation also including amplitude information is generated.

Table 1. *VarioGram* parameters together with their optimal values used in experiments.

Parameter	Value	Definition
bit-depth	6	The original signal sequence is quantized into 2^6 positive integer levels. This value corresponds to vertical slices of size $2^7 \times 1$ since differences between neighboring amplitude levels can vary between -63 and +63 for this setup.
windowSize	2	Differences between each samples neighboring in forward direction are encoded into the vertical variance arrays. This neighboring is applied for 1 and 2-distant samples.
sr	22050	Sampling rate applied to the original sound
jump	[3,5,7]	Undersampling rate for the 3 layers of the variance arrays
winLength	1024	Width of the sound frames
hopLength	512	Distance between the consecutive frames, corresponding to 50% overlap

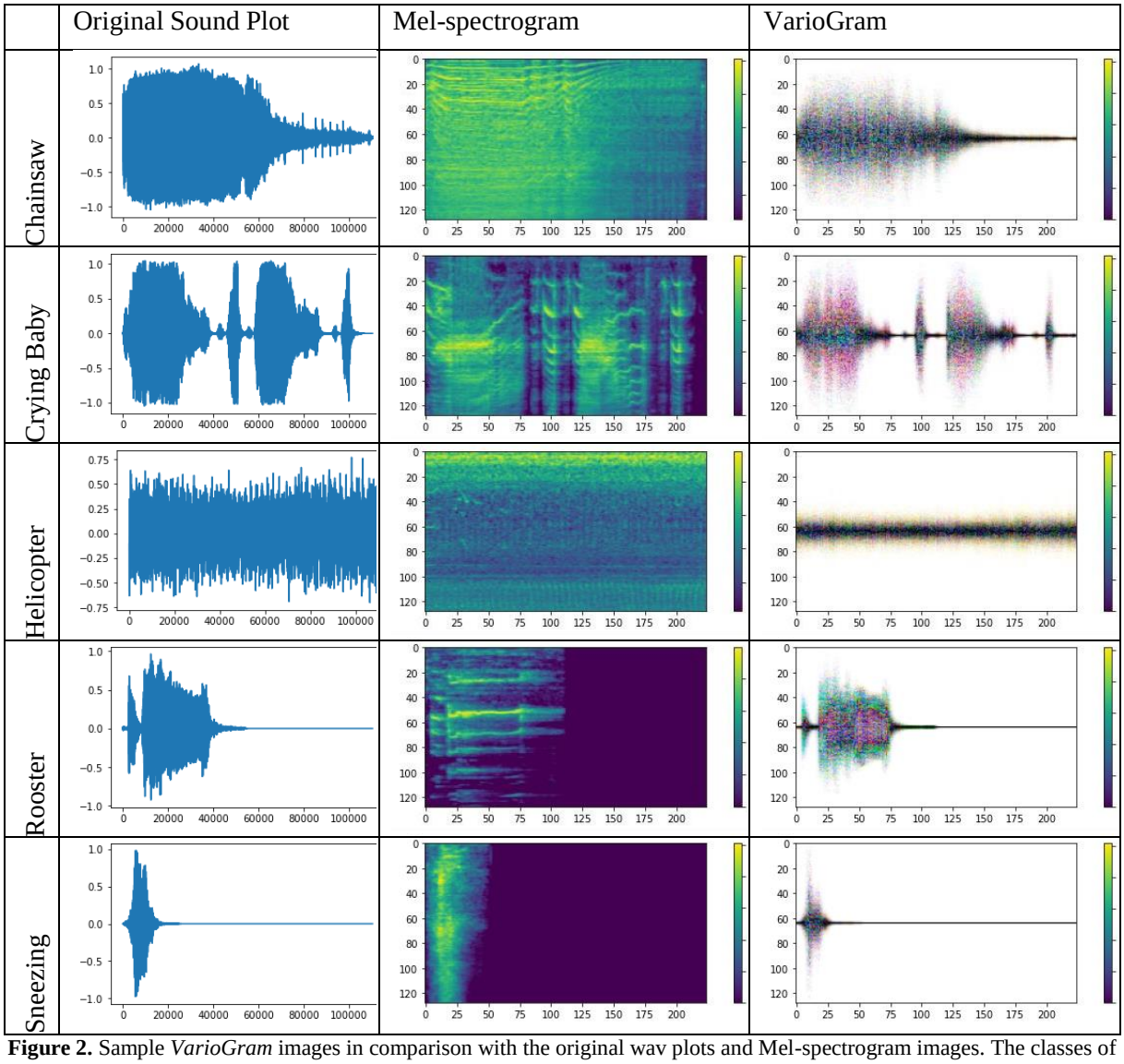

Figure 2. Sample *VarioGram* images in comparison with the original wav plots and Mel-spectrogram images. The classes of the sample sound waves are given in the row headers, while the type of the plot is given in the column headers.

Deep Learning with *VarioGrams*

To assess the representation capability of the *VarioGrams*, we tested these representations through a residual network deep learning architecture (ResNet). This model was first introduced by Microsoft researchers in 2015 and resembles the pyramidal cells of cerebral cortex of human brain. The main property of ResNets is the residual connection shortcutting between consecutive convolutional layers (Ismail Fawaz et al., 2019).

ESC dataset is presented with a fixed 5-fold structure, where sound excerpts from the same original long-duration source are placed in the same fold to preserve the hardness of the problem. Therefore, it is recommended to obey these fold assignments due to avoiding inflation of the classification success (Piczak, 2015a). To strictly satisfy this rule, we also kept the augmented sounds in the same fold.

Due to the nature of the ResNet architecture inheriting the transfer learning parameters, all VarioGram images are resized as 224x224 pixel RGB images to provide optimal match for this model. During the experiments, these VarioGram images are used either in its original RGB form or fused with the mel-spectrogram images of the same source. If they are combined with spectrograms, both *VarioGrams* and mel-spectrograms are converted into grayscale and these grayscale images are used as the single channel of the resulting fusion images. As a result, *VarioGrams* (vario) and mel-spectrograms (mels) derived with same windowing parameters stand for the synchronous channels of the resulting RGB image. We refer these fusion images as [vario+vario+vario] (standalone usage) or [mels+mels+vario] (fusion) notation in the results section.

Findings

To better understand the representation capacity of the *VarioGram* images, we performed classification tasks with 5-fold cross-validation procedure using standalone *VarioGram* images first, followed by standalone mel-spectrogram images and lastly with fusion images as their combination. The classification tasks are performed 10 times for each set, while the best results achieved are presented in Table 2.

Table 2. Classification results for standalone images or combinations of mel-spectrograms (mels) and *VarioGrams* (vario). In combination cases, each representation is first converted to grayscale to form a single channel of the resulting RGB fusion image.

Model	Accuracy
[mels, mels, mels]	92.91 %
[vario, vario, vario]	82.08 %
[mels, mels, vario]	93.33 %

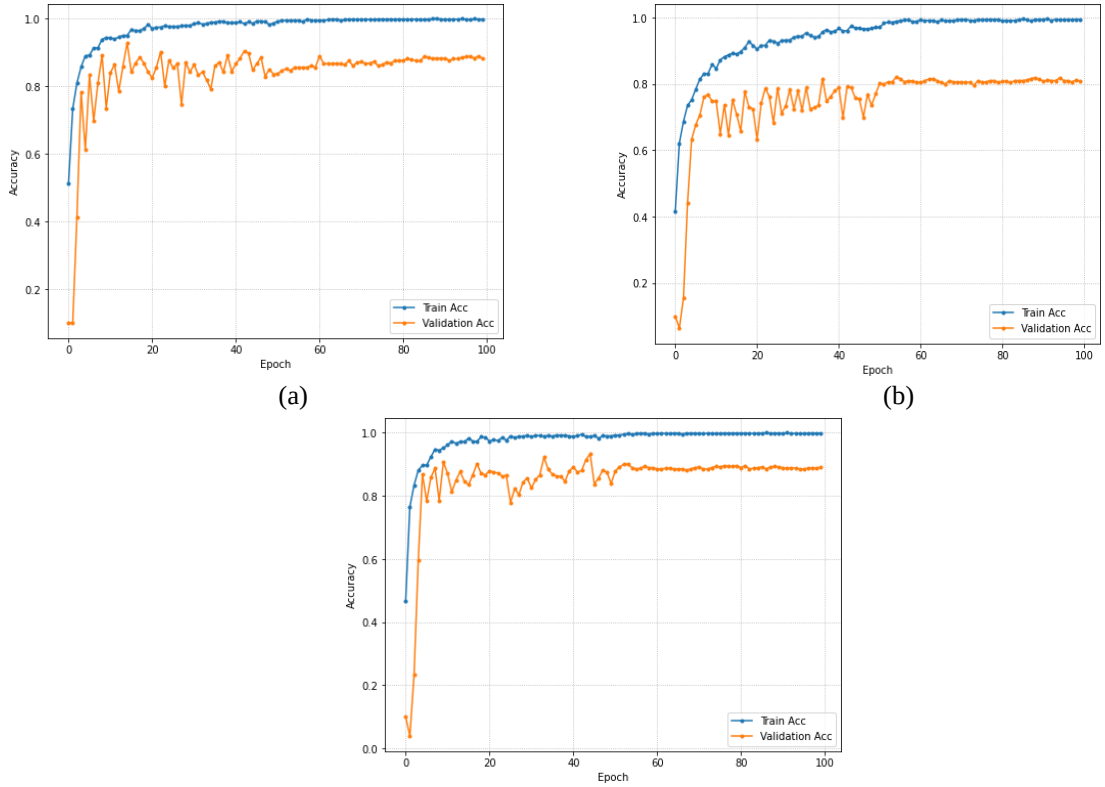

Figure 3. Classification curves for the best case achieved with ResNet50 architecture using the image representations of sounds as (a) [mels, mels, mels], 92.91%, (b) [vario, vario, vario], 82.08%, (c) [mels, mels, vario], 93.33% accuracy.

As presented in Table 2 and Fig. 3, standalone *VarioGram* representations are not as successful as mel-spectrogram representations, yielding approximately 10% lower accuracy scores. However, a combination of mel-spectrogram and *VarioGram* images can outperform the best standalone representation by ~0.4%. This is an indicator that *VarioGrams* have a boosting capability while used in combination with mel-spectrograms, and they cannot be recommended to be used solely for high-accuracy classification tasks. Since the window parameters used for both representors are the same, the resulting images have synchronous properties in horizontal (time) axis, making them complementary representations those are more powerful while used in hybrid fashion. We can also note that the classification performance that VarioGram represents, being not as competitive as the current classification studies, is better than the original study (80%) presented by Piczak in 2015 (Piczak, 2015a).

Conclusion

The proposed time-convexity representation named VarioGram, having roots from network science, offers a time-dependent representation for time series capturing both its envelope and varying intensity in its resulting shape and color. Although not performing as good as a well-known mel-spectrogram images in c classification task, it increases the sole performance of them while combined, and also forms an alternative way that should be consulted for different AI-related tasks. We recommend studying the representation performance of *VarioGrams* in medical time series data such as ECG or EMG, for which temporal convexity characteristics play a vital role in diagnosing the anomalies.

Genişletilmiş Özet

Giriş

Tıp, finans, havacılık, telekomünikasyon, endüstri, hava durumu, multimedya gibi çeşitli alanlardan zamansal veri setlerinin artan kullanılabilirliği ile birlikte zaman serilerinin sınıflandırma ve temsil tekniklerine olan bilimsel ilgi artmıştır (Bao et al., 2017; Canizo et al., 2019; Dafna et al., 2018; Gharehbaghi & Lindén, 2017; Pourbabaee et al., 2018; Soares et al., 2018). Bu kullanılabilir veri setleri, sınıflandırma, kümeleme, segmentasyon, anomalilik algılama vb. dahil olmak üzere çeşitli veri odaklı görevleri kolaylaştırır (Ares et al., 2016; Canizo et al., 2019; Kanani & Padole, 2020; Ruiz et al., 2021). Verileri temsil etmek için farklı teknikleri geliştirmedeki temel motivasyon, çoğunlukla verilerin yüksek boyutluluğundan kaynaklanan karmaşıklığı azaltmaktır. Ek olarak, bilim adamları zaman serisi verilerini çeşitli temsillere dönüştürürken, sınıflandırma doğruluğunu iyileştirmeye, gürültüyü ortadan kaldırmaya ve genel görevi hızlandırmaya odaklanır (Deng et al., 2016).

Zaman serilerini daha yapılandırılmış bir formatta temsil etmek için alternatif bir çerçeve de, karmaşık olarak birbirine bağlı nesnelerden oluşan sistemlerin bir ağ formatında temsil edilebilmesidir (Baydilli et al., 2017; Demir & Türker, 2021; Türker et al., 2016; Türker & Sulak, 2018). Bu yöntem kullanılarak zaman serileri, nicelenmiş genlik seviyelerinin düğümler ve ardışık seviyeler arasındaki komşulukların ise kenarlar olarak değerlendirildiği graflar şeklinde temsil edilebilir (Lacasa et al., 2015). Önceden tanımlanmış bir komşulukta başka bir seviyenin üstünden görülebilen nicelenmiş genlik seviyeleri arasında bağlantılar kurularak sıralı örneklerin dışbükeylik bilgileri ile görünürlük grafları inşa edilir (Lacasa et al., 2015). Bu çalışmalardan birinde Türker ve Aksu, zaman serilerinin graf gösterimi temeline dayalı olarak geliştirdikleri *ConnectoGram* temsili ile ESC-10 veri seti üzerindeki sınıflandırma çalışmalarında 2-3% oranında bir başarı artışı sağlanmışlardır (Türker & Aksu, 2022).

Yöntem

Veri ve ön işleme

Bu çalışmada, gerçek ortamlardan elde edilmiş bazı ses sınıflarını tanımakla ilgilenen bir Çevresel Ses Sınıflandırması (ÇSS) görevine odaklanılmıştır. Ses işlemenin bir alt alanı olarak çevresel seslerin sınıflandırılması işlemi; helikopter, deniz dalgaları, ağlayan bebek, çatırdayan ateş gibi önceden tanımlanmış seslerden oluşan çeşitli ses olaylarının tespit edilmesini içeren karmaşık bir görevdir (Zhang Zhichao and Xu, 2018). Bu alanda en yaygın kullanılan kütüphaneler arasında ESC-10, ESC-50 (Piczak, 2015b) ve UrbanSound8k (Salamon et al., 2014) veri setlerini sayabiliriz. ESC-50 veri seti 5 ana kategoriye ayrılmış 50 sınıftan oluşan 2000 sesten, ESC-10 ise 5 ana kategoride 10 sınıftan oluşan 400 sesten oluşmaktadır (Zhang Zhichao and Xu, 2018). Seslerin özelliklerini belirlemek için önerilen yönteme geçmeden önce tüm örneklere aynı ön-işleme prosedürü uygulanmıştır. Bu aşamada ses dalgaları 22050 Hz'de örneklenmiş, 60 dB'in altındaki bileşenler filtrelenmiştir. Bu işlemelerin ardından sinyallerin minimum ve maksimum değerleri [-1, 1] aralığında normalleştirilmiştir. Derin öğrenme yöntemlerinin daha başarılı sonuçlar vermesi için sinyal seviyesinde veri büyütme teknikleri uygulanmış ve 400 olan ses örneği sayısı 2400'e çıkartılmıştır. Grafa dayalı imajlar oluşturulduktan sonra ise *Python* dilinde yazılmış olan *Tensorflow* kütüphanesi ile ayrıca imaj düzeyinde döndürme, yatay ve dikey kaydırma, parlaklık, kırpma, yakınlaştırma gibi veri büyütme teknikleri uygulanmıştır (Mushtaq et al., 2021).

Ses Sinyallerinin VarioGram'a Dönüştürülmesi

Ardışık sinyallerin örüntüsünü oluşturacak şekilde graf tabanlı bir diziye dönüştürme işlemini gerçekleştirmek için her bir çerçeve 1024 farklı örnek içerecek şekilde ele alınır. Bu amaçla;

- İlk olarak bit tabanlı sayısallaştırma yapmak üzere ölçekleme seviyesi varsayılan olarak 6 bit derinliği seçilir ve $2^6 = 64$ sinyal seviyesi oluşturulur. Böylece [-1,1] aralığı [0,63] aralığına dönüştürülmüş olur.
- Bir *varyans dizisi*, sonuçta yüksekliği $2 \times 2^6 = 2^7 = 128$, genişliği 1, ve derinliği 3 olan bir dizidir.
    - **Yükseklik:** Kodlanan komşu genlik seviyeleri arasındaki fark. Bu fark $[0-63]\ to\ [-63-0] \rightarrow [-63,63]$ aralığına karşılık gelir. Böylece bu ölçeği göz önünde bulundurulduğunda boyut 2^7 olur.
    - **Genişlik:** Ardışık ses segmentlerinden elde edilmiş olan ve *VarioGram*'ı oluşturmak üzere yatay olarak döşenecek dizinin genişliği 1 pikseldir.
    - **Derinlik:** Dönüştürme işlemi her bir ses sinyalinden [3,5,7] olmak üzere 3 farklı atlama seviyesi ile gerçekleştirilir ve böylece RGB imajlarının katmanlarını oluşturacak şekilde 3 farklı $2^7 \times 1$ boyutunda dizi elde edilmiş olur.
- Her bir *varyans dizisi*, varsayılan atlama değerleri [3,5,7] olarak tutan bir atlama parametresi ile elde edilen 3 farklı alt örnekleme oranı ile hesaplanır. Böylece her bir örnekleme için renklendirilmiş *varyans dizisi'*ni oluşturacak ***R***, ***G*** ve ***B*** katmanları elde edilmiş olur.
- Ayrıca her ses karesi için uygulanan bu prosedür, *VarioGram* olarak adlandırılan zaman-varyanslı görüntüleri oluşturmak için yatay olarak döşenen ardışık kare sayısı kadar dikey RGB dizileri üretir.

VarioGram ile Derin Öğrenme

VarioGram'ların temsil kabiliyetini değerlendirmek için, ResNet derin öğrenme modeli kullanılmıştır. Bu model ilk olarak 2015 yılında Microsoft araştırmacıları tarafından tanıtılmış ve insan beyninin serebral korteksinin piramidal hücrelerinin araştırılmasında kullanılmıştır. ResNets'in ana özelliği, ardışık evrişim katmanları arasındaki artık bağlantı kısayollarıdır (Ismail Fawaz et al., 2019).

Çalışmanın veri kısmında ise 5 farklı ortamdan elde edilip 5 farklı klasörün her birine benzer sınıfa ait eşit uzunluktaki seslerin Dağıtıldığı ESC veri seti kullanılmıştır. Bu nedenle, sınıflandırma başarısındaki aşırı öğrenme gibi sorunlardan kaçınmak için eğitim ve test aşamalarında bu 5 klasöre ayrılmış ve bu seslerin karıştırılmadan kullanılması tavsiye edilmiştir (Piczak, 2015a).

Transfer öğrenme parametrelerini kullanan ResNet mimarisinin doğası gereği en iyi sonucu elde etmek üzere tüm *VarioGram* imajları RGB formatında 224x224 piksel olarak yeniden boyutlandırılmıştır. Eğer *VarioGram* imajlar mel-spectrogram imajlarla birleştirilecekse her iki imaj da öncelikle gri tonlamaya dönüştürülür ve bu gri tondaki imajlar birleştirilerek RGB formatının birer katmanı olarak ele alınır. Sonuç olarak, birleştirilmeden sadece *VarioGram*'ların olduğu imajlar [vario+vario+vario] ve mel-spectrogram'lar ile birleştirilen imajlar ise [mels+mels+vario] olarak eğitime alınmış olur.

Bulgular

VarioGram görüntülerinin temsil kapasitesini daha iyi belirlemek için, sınıflandırma işlemi 5-fold çapraz-doğrulama yöntemi ile gerçekleştirilmiştir. İlk olarak *VarioGram* ve mel-spectrogram imajlar ayrı ayrı hiçbir

birleştirme işlemi uygulanmadan eğitim için kullanılmıştır. Ardından *VarioGram* ve mel-spectrogram imajların birleştirilmiş hali ele alınmıştır. Eğitimden daha başarılı sonuçlar elde etmek için her sınıflandırma görevi 10 kez tekrarlanmış ve bu tekrarların ortalaması alınmıştır. Yapılan çalışmalardan elde edilen sonuçlar Tablo 2'de gösterilmiştir.

Sadece *VarioGram* imajların kullanılması durumunda mel-spectrogram imajlar kadar iyi sonuç elde edilememiş ve yaklaşık olarak 10% kadar daha düşük verimli sonuçlar ortaya çıkmıştır. Bununla beraber *VarioGram* ve mel-spektrogram imajların birleştirilmiş hali ile yaklaşık olarak ~0.4% gibi daha iyi bir sonuç elde edilmiştir. Bu durum, mel-spectrogram'larla birlikte kullanıldığında *VarioGram*'ların bir performans artırma kapasitesine sahip olduğunu göstermektedir. Her iki temsil yöntemi için kullanılan pencere parametreleri aynı olduğundan, ortaya çıkan görüntüler yatay (zaman) eksende senkronize özelliklere sahiptir. Bu da, bu iki yöntemin hibrit modda kullanılması halinde zaman serileri ile ilgili daha güçlü tamamlayıcı temsiller elde edilmesini sağlamaktadır. Mevcut sınıflandırma çalışmaları kadar rekabetçi olmamakla birlikte *VarioGram'*ın temsil ettiği sınıflandırma performansının, 2015 yılında Piczak tarafından sunulan orijinal çalışmadan (%80) daha iyi bir sonuç verdiği görülmektedir (Piczak, 2015a).

Sonuç

Bu çalışmada önerdiğimiz ve köklerini ağ biliminden alan *VarioGram*, zaman serileri için genlik seviyeleri arasındaki farkları ve yoğunlukları dikkate alarak bu değişimleri renk tonları ile ifade eden zamana bağlı yeni bir temsil metodu sunmaktadır. Sınıflandırma çalışmalarında her ne kadar bu yöntem bilinen ve sık kullanılan mel-spektrogram imajlarına dayalı yöntemler kadar iyi sonuç vermese de farklı yapay zeka çalışmalarında önerdiğimiz bu yöntemin imajlarla birlikte kullanılması halinde performans artışına katkı sağladığı görülmüştür. Sağlık sorunlarını belirlemek amaıyla geçici değişimlerin hayati önem taşıdığı ECG veya EMG verileri gibi tıbbi zaman serilerinin analizinde *VarioGram* temsil yönteminin kullanılmasını tavsiye etmekteyiz.

Yayın Etiği Bildirimi / Research Ethics

Araştırma ve yayın etiği konusunda bilimsel etik kaideleri göz önünde bulundurulmuştur. / Scientific ethical principles have been taken into consideration in research and publication ethics.

Araştırmacıların Katkı Oranı / Contribution Rate of Researchers

Birinci araştırmacı, makalenin genelinde kavramsallaştırma, metodoloji, formal analiz ve sorumlu yazar olarak görev alırken, ikinci araştırmacı, literatür taraması, metodoloji, yazılım geliştirme, test ve doğrulama konularında katkı sağlamıştır. / While the first researcher worked as the conceptualization, methodology, formal analysis and lead author throughout the article, the second researcher contributed to the literature review, methodology, software development, testing and validation.

Çıkar Çatışması / Conflict of Interest

Bu çalışmada herhangi bir çıkar çatışması bulunmamaktadır. / This study has no conflict of interest.

Fon Bilgileri / Funding

Bu çalışmada herhangi bir fon kullanılmamıştır. / No funds were used in this study.

Etik Kurul Onayı / The Ethical Commitee Approval

Bu araştırma makalesinin etik sorunu olmadığını beyan ederiz. / We hereby declare that this research article does not have an unethical problem.

Kaynakça/References

Ares, J., Lara, J. A., Lizcano, D., & Suarez, S. (2016). A soft computing framework for classifying time series based on fuzzy sets of events. *Information Sciences*, *330*, 125–144. https://doi.org/10.1016/J.INS.2015.10.014

Bao, W., Yue, J., & Rao, Y. (2017). A deep learning framework for financial time series using stacked autoencoders and long-short term memory. *PLoS ONE*, *12*.

Baydilli, Y. Y., Bayir, Ş., & Türker, I. (2017). A Hierarchical View of a National Stock Market as a Complex Network. *Economic Computation & Economic Cybernetics Studies & Research*, *51*(1).

Canizo, M., Triguero, I., Conde, A., & Onieva, E. (2019). Multi-head CNN–RNN for multi-time series anomaly detection: An industrial case study. *Neurocomputing*, *363*, 246–260. https://doi.org/10.1016/J.NEUCOM.2019.07.034

Chan, K.-P., & Fu, A. W.-C. (1999). Efficient time series matching by wavelets. *Proceedings 15th International Conference on Data Engineering (Cat. No.99CB36337)*, 126–133. https://doi.org/10.1109/ICDE.1999.754915

Chen, Z., Zuo, W., Hu, Q., & Lin, L. (2015). Kernel sparse representation for time series classification. *Information Sciences*, *292*, 15–26. https://doi.org/10.1016/J.INS.2014.08.066

Dafna, E., Tarasiuk, A., & Zigel, Y. (2018). Sleep staging using nocturnal sound analysis. *Scientific Reports*, *8*(1), 13474. https://doi.org/10.1038/s41598-018-31748-0

Demir, S., & Türker, İ. (2021). Arithmetic success and gender-based characterization of brain connectivity across EEG bands. *Biomedical Signal Processing and Control*, *64*, 102222. https://doi.org/10.1016/J.BSPC.2020.102222

Deng, W., Wang, G., & Xu, J. (2016). Piecewise two-dimensional normal cloud representation for time-series data mining. *Information Sciences*, *374*, 32–50. https://doi.org/10.1016/J.INS.2016.09.027

Gharehbaghi, A., & Lindén, M. (2017). A deep machine learning method for classifying cyclic time series of biological signals using time-growing neural network. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(9), 4102–4115.

Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., & Muller, P.-A. (2019). Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, *33*(4), 917–963. https://doi.org/10.1007/s10618-019-00619-1

Kanani, P., & Padole, M. (2020). ECG Heartbeat Arrhythmia Classification Using Time-Series Augmented Signals and Deep Learning Approach. *Procedia Computer Science*, *171*, 524–531. https://doi.org/10.1016/J.PROCS.2020.04.056

Lacasa, L., Luque, B., Ballesteros, F., Luque, J., & Nuño, J. C. (2008). From time series to complex networks: The visibility graph. *Proceedings of the National Academy of Sciences*, *105*(13), 4972–4975. https://doi.org/10.1073/PNAS.0709247105

Lacasa, L., Nicosia, V., & Latora, V. (2015). Network structure of multivariate time series. *Scientific Reports*, *5*(1), 15508. https://doi.org/10.1038/srep15508

Mushtaq, Z., Su, S. F., & Tran, Q. V. (2021). Spectral images based environmental sound classification using CNN with meaningful data augmentation. *Applied Acoustics*, *172*, 107581. https://doi.org/10.1016/J.APACOUST.2020.107581

Peng, Z., Dang, J., Unoki, M., & Akagi, M. (2021). Multi-resolution modulation-filtered cochleagram feature for LSTM-based dimensional emotion recognition from speech. *Neural Networks*, *140*, 261–273. https://doi.org/10.1016/J.NEUNET.2021.03.027

Piczak, K. J. (2015a). Environmental sound classification with convolutional neural networks. *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, 1–6.

Piczak, K. J. (2015b). ESC: Dataset for Environmental Sound Classification. *Proceedings of the 23rd ACM International Conference on Multimedia*, 1015–1018. https://doi.org/10.1145/2733373.2806390

Pourbabaee, B., Roshtkhari, M. J., & Khorasani, K. (2018). Deep Convolutional Neural Networks and Learning ECG Features for Screening Paroxysmal Atrial Fibrillation Patients. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *48*(12), 2095–2104. https://doi.org/10.1109/TSMC.2017.2705582

Ruiz, A. P., Flynn, M., Large, J., Middlehurst, M., & Bagnall, A. (2021). The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, *35*(2), 401–449. https://doi.org/10.1007/s10618-020-00727-3

Salamon, J., Jacoby, C., & Bello, J. P. (2014). A Dataset and Taxonomy for Urban Sound Research. *Proceedings of the 22nd ACM International Conference on Multimedia*, 1041–1044. https://doi.org/10.1145/2647868.2655045

Sharan, R. v, & Moir, T. J. (2015). Cochleagram image feature for improved robustness in sound recognition. *2015 IEEE International Conference on Digital Signal Processing (DSP)*, 441–444. https://doi.org/10.1109/ICDSP.2015.7251910

Soares, E., Costa, P., Costa, B., & Leite, D. (2018). Ensemble of evolving data clouds and fuzzy models for weather time series prediction. *Applied Soft Computing*, *64*, 445–453. https://doi.org/10.1016/J.ASOC.2017.12.032

Türker, İ., & Aksu, S. (2022). Connectogram – A graph-based time dependent representation for sounds. *Applied Acoustics*, *191*, 108660. https://doi.org/10.1016/J.APACOUST.2022.108660

Türker, İ., Şehirli, E., & Demiral, E. (2016). Uncovering the differences in linguistic network dynamics of book and social media texts. *SpringerPlus*, *5*(1), 864. https://doi.org/10.1186/s40064-016-2598-2

Türker, İ., & Sulak, E. E. (2018). A multilayer network analysis of hashtags in twitter via co-occurrence and semantic links. *International Journal of Modern Physics B*, *32*(04), 1850029. https://doi.org/10.1142/S0217979218500297

Yin, J., Liu, Z., Jin, Z., & Yang, W. (2012). Kernel sparse representation based classification. *Neurocomputing*, *77*(1), 120–128. https://doi.org/10.1016/J.NEUCOM.2011.08.018

Zhang Zhichao and Xu, S. and C. S. and Z. S. (2018). Deep Convolutional Neural Network with Mixup for Environmental Sound Classification. In C.-L. and C. X. and Z. J. and T. T. and Z. N. and Z. H. Lai Jian-Huang and Liu (Ed.), *Pattern Recognition and Computer Vision* (pp. 356–367). Springer International Publishing.