Information Fusion 41 (2018) 161–175

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

A tutorial review on entropy-based handcrafted feature extraction for information fusion

Rodrigo Capobianco Guido

Instituto de Biociências, Letras e Ciências Exatas, Unesp - Univ Estadual Paulista São Paulo State University, Rua Cristóvão Colombo 2265, Jd Nazareth, São José do Rio Preto - SP, 15054-000, Brazil

ARTICLE INFO

Article history:
Received 19 June 2017
Revised 27 August 2017
Accepted 2 September 2017
Available online 12 September 2017

Keywords:
Entropy
Handcrafted feature extraction
Information fusion
Deep networks
Restricted-vocabulary speech recognition
Image synthesis

ABSTRACT

Entropy (H) is the main subject of this article, concisely written to serve as a tutorial introducing two feature extraction (FE) methods for usage in digital signal processing (DSP) and pattern recognition (PR). The theory, carefully exposed, is supplemented with numerical cases, augmented with C/C++ source-codes and enriched with example applications on restricted-vocabulary speech recognition and image synthesis. Complementarily and as innovatively shown, the ordinary calculation of H corresponds to the outcome of a partially pre-tuned deep neural network architecture which fuses important information, bringing a cutting-edge point-of-view for both DSP and PR communities.

1. Introduction

1.1. Objective and text structure

This is the third in a set of tutorials I have recently published with the same objective: innovative usage of humble and well-known concepts for the benefit of both digital signal processing (DSP) and pattern recognition (PR) communities. The preceding texts, [23] and [24], were respectively dedicated to the exploration of relevant aspects of *energy* by means of proposed methods A_1, A_2 and A_3, and *zero-crossing rates* (ZCRs), according to the techniques introduced as B_1, B_2 and B_3. Successfully, I employed those formulations for neurophysiological signal analysis, texture characterisation, text-dependent speaker verification, speech classification and segmentation, image border extraction and biomedical signal processing. Energy, that is used to express the potential to perform work, as well as ZCRs, which are commonly applied to elementary spectral content analysis, act disparately in correlation to *entropy* (H) [13,70], the feature explored in this article.

Despite the emerging deep learning (DL) technologies employed for automatic feature learning [22,34], handcrafted feature extraction (FE), i.e., the situation in which the system engineer chooses the appropriate features to be extracted from the signal under analysis, continues to play an important role in DSP and PR. Particularly, I demonstrate that H, by itself, obtained based on two proposed approaches for FE from both unidimensional (1D) and bidimensional (2D) data, has flagrant potential, as also evidenced in relevant scientific articles published last year [11,17,20,42,50,51,53,81,84,86] and a few years ago [6,15,21,37,56,57,62,63,73,75,83]. Similarly to the characterization of ZCRs as neurocomputing agents [24], H is shown to be the outcome of a specifically tuned deep neural network (DNN) that fuses important information, bringing an innovative point-of-view for both DSP and PR communities. Furthermore, experiments and applications on restricted-vocabulary speech recognition and image synthesis reassure the efficacy of the proposed techniques.

Compromised with a balance among *creativity*, *simplicity* and *accuracy*, exactly as in [23] and [24], this paper is organised as follows. A review on H accompanied by some of its recent applications is the theme of the next subsection. Section 2, oppositely, describes the proposed approaches in detail and my particular point-of-view about H for both 1D and 2D signals. Proceeding, Section 3 presents some numerical examples which complement the theoretical explanations, easing their comprehension. Illustrative experiments and applications involving 1D and 2D signals can be found in Section 4 and, lastly, I conclude the paper. Following my previous rationale, as indicated in [24]-pp.1 with respect to article [23], I emphasise the importance of those articles, suggesting their readings beforehand for a better understanding of the ideas discussed herein.

E-mail address: guido@ieee.org
URL: http://www.sjrp.unesp.br/~guido/

http://dx.doi.org/10.1016/j.inffus.2017.09.006
1566-2535/

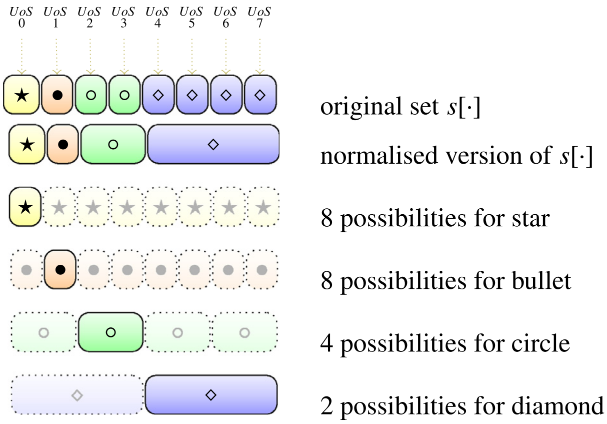

Fig. 1. The way normalisations should be interpreted for the example set $s[\cdot] = \{\star, \bullet, \circ, \circ, \diamond, \diamond, \diamond, \diamond\}$. *UoS* means "unit of space".

1.2. A review on entropy and its applications

The records I found on the *Web of Science* website show that the first published article about H is dated from 1900 [82]. As science advances, H has been at the forefront of research in a diversity of fields such as general physics [39], thermodynamics [10], astrophysics [8], statistical mechanics [67], genetics [65], economics [46] and arts [1]. In order to understand its essence, focusing particularly on the field of Information Theory, where the DSP and PR communities find utility [19,49], the readers are first requested to reflect on the meaning of *information* [39]-pp.117, as follows. The exposition hereafter is inspired by the traditional article [68] published in 1948 by Claude Shannon[1], who is considered the father of Information Theory, and by additional respected bibliographical materials.

Let $p_i = \frac{\alpha_i}{M} = \frac{1}{(M/\alpha_i)}$, $(0 \le i < K)$ and $(0 \le \alpha_i < M)$, be the *probability* [59] of the ith distinct *datum*, i.e., symbol, in a set $s[\cdot] = \{s_0, s_1, s_2, \ldots, s_{M-1}\}$ of size M with K distinct symbols. Consequently, there are α_i symbols within M matching the ith expectation or, equivalently, there is one among $\frac{1}{(M/\alpha_i)}$. Specifically, the denominator (M/α_i) represents the normalised number of possibilities for the ith symbol, with the normalisation interpreted in such a way that each subset of repeated elements is converted into an unique size representative of size equal to that of the entire subset. Furthermore, (M/α_i) written based on a certain alphabet β produces words for which **the length** corresponds to what we know as *information*.

Assume, as an example, the 8-sample long set $s[\cdot] = \{\star, \bullet, \circ, \circ, \diamond, \diamond, \diamond, \diamond\}$. The probabilities of stars, bullets, circles and diamonds are, respectively, $p_0 = \frac{1}{(M/\alpha_0)} = \frac{1}{(8/1)} = \frac{1}{8}$, $p_1 = \frac{1}{(M/\alpha_1)} = \frac{1}{(8/1)} = \frac{1}{8}$, $p_2 = \frac{1}{(M/\alpha_2)} = \frac{1}{(8/2)} = \frac{1}{4}$ and $p_3 = \frac{1}{(M/\alpha_3)} = \frac{1}{(8/4)} = \frac{1}{2}$. Thus, the corresponding normalised number of possibilities for each star, bullet, circle and diamond is $(8/1) = 8$, $(8/1) = 8$, $(8/2) = 4$ and $(8/4) = 2$. Fig. 1 illustrates the physical meaning of the normalisations. Regarding the star, only one unit of space among eight is required for its placement; thus, there are eight possibilities to place it. The same holds true for the bullet. In relation to both circles in the original set, the normalisation converts them in only one double-length circle and forces it to occupy two original units of space among eight; thus, there are four possible placements for the bigger circle. Lastly, the four original diamonds are converted, due to the normalisation, into only one larger diamond which occupies four original units of space, implying that there are only two possible placements for this enlarged symbol.

When choosing the binary basis [30], i.e., $= 2$, as being the alphabet, "0" and "1" are the only existing characters, known as bits, which compose the corresponding words. Particularly,

- "000", "001", "010", "011", "100", "101", "110" and "111" are the $2^3 = 8$ possibilities for placing stars, implying that **three** bits are needed to express such locations;
- "000", "001", "010", "011", "100", "101", "110" and "111" are also the $2^3 = 8$ possibilities for placing bullets, which consequently require **three** bits to express such locations;
- "00", "01", "10" and "11" are the $2^2 = 4$ possibilities for placing circles, which require **two** bits to express such locations;
- "0" and "1" are the $2^1 = 2$ possibilities for placing diamonds, which require only **one** bit to express such locations.

Instead of written and counted, the number of bits in each case may be easily calculated by means of the base $\beta = 2$ logarithms of the normalized possibilities ([30]-pp.8), i.e., $log_\beta(\frac{M}{\alpha_0}) = log_\beta(\frac{1}{p_0}) = log_2(8) = 3$, $log_\beta(\frac{M}{\alpha_1}) = log_\beta(\frac{1}{p_1}) = log_2(8) = 3$, $log_\beta(\frac{M}{\alpha_2}) = log_\beta(\frac{1}{p_2}) = log_2(4) = 2$ and $log_\beta(\frac{M}{\alpha_3}) = log_\beta(\frac{1}{p_3}) = log_2(2) = 1$. These values are the lengths, i.e., the number of bits, of the words in each case, which correspond to their **information**.

Generally, $log_\beta(\frac{1}{p_i})$ expresses information in terms of the number of elements required to write $\frac{1}{p_i}$ possibilities for placements based on alphabet β, i.e.,

the amount of information for the ith symbol is

$$\underbrace{log_\beta \overbrace{\left(\frac{1}{p_i}\right)}^{\substack{\text{normalised number of possibilities} \\ \text{for the } i^{th} \text{ symbol}}}}_{\substack{\text{information, i.e., number of elements required} \\ \text{to write } (1/p_i) \text{ using alphabet } \beta}}.$$

In order to obtain a **global quantification** for the information in the **entire** set $s[\cdot]$, the more natural procedure corresponds to the calculation of the weighted sum of the independent amounts of information, where the probabilities of occurrences are the respective weights. Thus,

$$H = \sum_{i=0}^{K-1} p_i \cdot log_\beta\left(\frac{1}{p_i}\right). \quad (1)$$

Alternatively and taking into account the property of logarithms which states that $log(\frac{1}{x}) = -log(x)$, $\forall x \in \Re^+$, Eq. (1) may be rewritten as

$$H = -\sum_{i=0}^{K-1} p_i \cdot log_\beta(p_i),$$

that is the most traditional way used to express entropy. In our previous example, $H = -\sum_{i=0}^{K-1} p_i \cdot log_\beta(p_i) = -\sum_{i=0}^{3} p_i \cdot log_2(p_i)$ $= -((p_0 \cdot log_2(p_0)) + (p_1 \cdot log_2(p_1)) + (p_2 \cdot log_2(p_2)) + (p_3 \cdot log_2(p_3)))$ $= -((\frac{1}{8} \cdot log_2(\frac{1}{8})) + (\frac{1}{8} \cdot log_2(\frac{1}{8})) + (\frac{1}{4} \cdot log_2(\frac{1}{4})) + (\frac{1}{2} \cdot log_2(\frac{1}{2})))$ $= \frac{3}{8} + \frac{3}{8} + \frac{1}{2} + \frac{1}{2} = \frac{7}{4}$ bits.

Deservedly also known as *Shannon's entropy*, H may alternatively be understood as a measure of unpredictability of information content, whereas it equals zero upon a concrete and fully predictable outcome, i.e., when $K = 1$ and $p_0 = 1$. Particularly concerning our example, the unpredictability of the normalised diamond is the lowest: there are only two possible placements for it, as seen in Fig. 1. In contrast, normalised circles have a higher unpredictability because there are four possible placements for them. Lastly, both normalised stars and bullets present the highest unpredictabilities amongst the symbols with eight possible placements.

[1] Had he not passed away in 2001, Dr Shannon would have celebrated his 100th birthday on April 30, 2016.