

Article

## Tsallis Mutual Information for Document Classification

Màrius Vila \*, Anton Bardera, Miquel Feixas and Mateu Sbert

Institut d'Informàtica i Aplicacions, Universitat de Girona, Campus Montilvi, Girona 17071, Spain;  
E-Mails: anton.bardera@ima.udg.edu (A.B.); feixas@ima.udg.edu (M.F.); mateu@ima.udg.edu (M.S.)

\* Author to whom correspondence should be addressed; E-Mail: marius.vila@ima.udg.edu;  
Tel.: +34-972418823; Fax: +34-972418792.

Received: 1 August 2011; in revised form: 5 September 2011 / Accepted: 8 September 2011 /

Published: 14 September 2011

---

**Abstract:** Mutual information is one of the mostly used measures for evaluating image similarity. In this paper, we investigate the application of three different Tsallis-based generalizations of mutual information to analyze the similarity between scanned documents. These three generalizations derive from the Kullback–Leibler distance, the difference between entropy and conditional entropy, and the Jensen–Tsallis divergence, respectively. In addition, the ratio between these measures and the Tsallis joint entropy is analyzed. The performance of all these measures is studied for different entropic indexes in the context of document classification and registration.

**Keywords:** Tsallis entropy; mutual information; image similarity; document classification

---

### 1. Introduction

Based on the capability of scanners to transform a large amount of documents to digital images, the automatic processing of administrative documents is a topic of major interest in many office applications. Some examples are noise removal, image extraction, or background detection. Other processes, such as document clustering or template matching, require the definition of document similarity. Document clustering aims to classify similar documents in groups and template matching consists in finding the spatial correspondence of a given document with a template in order to identify the relevant fields of the document.

According to [1], the definition of the similarity between documents can be divided into two main groups based respectively on matching local features, such as the matching of recognized characters [2]

or different types of line segments [3], and extracting global layout information, such as the use of a spatial layout representation [4] or geometric features [5]. In this paper, instead of extracting specific pieces of information or analyzing the document layout, we propose to use global measures to evaluate the similarity between two image documents. The similarity between two images can be computed using numerous distance or similarity measures. In the medical image registration field, mutual information has become a standard image similarity measure [6]. In this paper we investigate three different generalizations of this measure based on Tsallis entropy. As it was previously noted in [7], the main motivation for the use of non-extensive measures in image processing is the presence of correlations between pixels of the same object in the image that can be considered as long-range correlations. Although our analysis can be extended to a wide variety of document types, in this paper we focus our attention on invoice classification. In our experiments, we show the good performance of some of the proposed measures using an invoice database composed by colored images.

This paper is organized as follows. Section 2 briefly reviews some previous work on information theory and its use in image registration and document classification. Section 3 presents three generalizations of mutual information that will be applied to document classification. Section 4 presents our general framework for document processing. Section 5 analyzes the obtained results in invoice classification and registration. Finally, Section 6 presents conclusions and future work.

## 2. Related Work

In this section, we review some basic concepts on information theory, image registration and document image analysis.

### 2.1. Information-Theoretic Measures

Let  $\mathcal{X}$  be a finite set, let  $X$  be a random variable taking values  $x \in \mathcal{X}$  with distribution  $p(x) = Pr[X = x]$ . Likewise, let  $Y$  be a random variable taking values  $y \in \mathcal{Y}$ . The *Shannon entropy*  $H(X)$  of a random variable  $X$  is defined by

$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \log p(x) \quad (1)$$

The Shannon entropy  $H(X)$  measures the average uncertainty of random variable  $X$ . If the logarithms are taken in base 2, entropy is expressed in bits. The *conditional entropy* is defined by

$$H(X|Y) = - \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x, y) \log p(x|y) \quad (2)$$

where  $p(x, y) = Pr[X = x, Y = y]$  is the joint probability and  $p(x|y) = Pr[X = x|Y = y]$  is the conditional probability. The conditional entropy  $H(X|Y)$  measures the average uncertainty associated with  $X$  if we know the outcome of  $Y$ . The *mutual information (MI)* between  $X$  and  $Y$  is defined by

$$I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \quad (3)$$

$MI$  is a measure of the shared information between  $X$  and  $Y$ .

An alternative definition of *MI* can be obtained from the definition of the *informational divergence* or *Kullback–Leibler distance (KL)*. The distance  $KL(p, q)$  between two probability distributions  $p$  and  $q$  [8,9], that are defined over the alphabet  $\mathcal{X}$ , is given by

$$KL(p, q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \quad (4)$$

The conventions that  $0 \log(0/0) = 0$  and  $a \log(a/0) = \infty$  if  $a > 0$  are adopted. The informational divergence satisfies the information inequality  $KL(p, q) \geq 0$ , with equality if and only if  $p = q$ . The informational divergence is not strictly a metric since it is not symmetric and does not satisfy the triangle inequality. Mutual information can be obtained from the informational divergence as follows [8]:

$$I(X; Y) = KL(p(x, y), p(x)p(y)) \quad (5)$$

Thus, mutual information can also be seen as the distance between the joint probability distribution  $p(x, y)$  and the distribution  $p(x)p(y)$ , *i.e.*, the distance of the joint distribution to the independence.

Mutual information can be also expressed as a *Jensen–Shannon divergence*. Since Shannon entropy is a concave function, from Jensen’s inequality, we can obtain the Jensen–Shannon inequality [10]:

$$JS(\pi_1, \dots, \pi_n; p_1, \dots, p_n) = H\left(\sum_{i=1}^n \pi_i p_i\right) - \sum_{i=1}^n \pi_i H(p_i) \geq 0 \quad (6)$$

where  $JS(\pi_1, \dots, \pi_n; p_1, \dots, p_n)$  is the Jensen–Shannon divergence of probability distributions  $p_1, p_2, \dots, p_n$  with prior probabilities or weights  $\pi_1, \pi_2, \dots, \pi_n$ , fulfilling  $\sum_{i=1}^n \pi_i = 1$ . The JS-divergence measures how ‘far’ are the probabilities  $p_i$  from their likely joint source  $\sum_{i=1}^n \pi_i p_i$  and equals zero if and only if all  $p_i$  are equal. Jensen–Shannon’s divergence coincides with  $I(X; Y)$  when  $\{\pi_i\}$  is equal to the marginal probability distribution  $p(x)$  and  $\{p_i\}$  are equal to the rows  $p(Y|x_i)$  of the probability conditional matrix of the information channel  $X \rightarrow Y$ . Then, *MI* can be redefined as

$$I(X; Y) = JS(p(x_1), \dots, p(x_n); p(Y|x_1), \dots, p(Y|x_n)) \quad (7)$$

A generalization of the Shannon entropy was given by Tsallis in [11]:

$$H_\alpha^T(X) = \frac{1}{\alpha - 1} \left( 1 - \sum_{x \in \mathcal{X}} p(x)^\alpha \right) \quad (8)$$

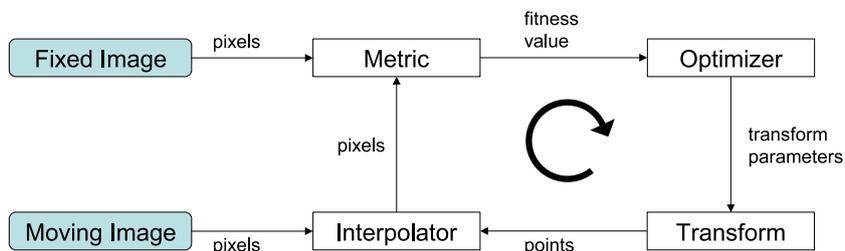
where  $\alpha > 0$  and  $\alpha \neq 1$ .  $H_\alpha^T(X)$  is a concave function of  $p$  for  $\alpha > 0$  and  $H_\alpha^T(X) = H(X)$  when  $\alpha \rightarrow 1$  (and natural logarithms are taken in the definition of the Shannon entropy).

## 2.2. Image Registration

Image registration is treated as an iterative optimization problem with the goal of finding the spatial mapping that will bring two images into alignment. This process is composed of four elements (see Figure 1). As input, we have both fixed  $\mathbf{X}$  and moving  $\mathbf{Y}$  images. The *transform* represents the spatial mapping of points from the fixed image space to points in the moving image space. The *interpolator* is used to evaluate the moving image intensity at non-grid positions. The *metric* provides a measure of how well the fixed image is matched by the transformed moving one. This measure forms the

quantitative criterion to be optimized by the *optimizer* over the search space defined by the parameters of the transform.

**Figure 1.** Main components of the registration process.



The crucial point of image registration is the choice of a metric. One of the simplest measures is the *sum of squared differences (SSD)*. For  $N$  pixels in the overlap domain  $\Omega_{A,B}$  of images  $A$  and  $B$ , this measure is defined as

$$SSD = \frac{1}{N} \sum_{i \in \Omega_{A,B}} |A(i) - B(i)|^2 \tag{9}$$

where  $A(i)$  and  $B(i)$  represent the intensity at a pixel  $i$  of the images  $A$  and  $B$ , respectively, and  $N$  the number of overlapping pixels. When this measure is applied, we assume that the image values are calibrated to the same scale. This measure is very sensitive to a small number of pixels that have very large intensity differences between images  $A$  and  $B$ . Another common image similarity measure is the *correlation coefficient (CC)*, which is defined as

$$CC = \frac{\sum_{i \in \Omega_{A,B}} (A(i) - \bar{A})(B(i) - \bar{B})}{[\sum_{i \in \Omega_{A,B}} (A(i) - \bar{A})^2 \sum_{i \in \Omega_{A,B}} (B(i) - \bar{B})^2]^{\frac{1}{2}}} \tag{10}$$

where  $\bar{A}$  is the mean pixel value in image  $A|_{\Omega_{A,B}}$  and  $\bar{B}$  is the mean of  $B|_{\Omega_{A,B}}$ . While the *SSD* makes the implicit assumption that the images differ only by Gaussian noise, the *CC* assumes that there is a linear relationship between the intensity values in the images [12].

From the information theory perspective, the registration between two images  $X$  and  $Y$  (associated with the random variables  $X$  and  $Y$ , respectively) can be represented by an information channel  $X \rightarrow Y$ , where its marginal and joint probability distributions are obtained by simple normalization of the corresponding intensity histograms of the overlap area of both images [13]. The most successful automatic image registration methods are based on the maximization of *MI*. This method, almost simultaneously introduced by Maes *et al.* [13] and Viola *et al.* [14], is based on the conjecture that the correct registration corresponds to the maximum *MI* between the overlap areas of the two images. Later, Studholme *et al.* [15] proposed a normalization of mutual information defined by

$$NMI(X; Y) = \frac{I(X; Y)}{H(X, Y)} \tag{11}$$

which is more robust than *MI*, due to its greater independence of the overlap area. Another theoretical justification of its good behavior is that *NMI* is a true distance. Different measures derived from the Tsallis entropy have also been applied to image registration [16–19].

### 2.3. Document Image Similarity

In the context of document image analysis, image similarity is mainly used for classification purposes in order to index, retrieve, and organize specific document types. Nowadays, this task is especially important because huge volumes of documents are scanned to be processed in an automatic way. Some automatic solutions based on optical character recognition (OCR), bank check reader, postal address reader and signature verifier, have already been proposed but a lot of work has still to be done to classify other types of documents such as tabular forms, invoices, bills, and receipts [20]. Chen and Blostein [21] presented an excellent survey on document image classification.

Many automatic classification techniques of image documents are based on the extraction of specific pieces of information from the documents. In particular, OCR software is especially useful to extract relevant information in applications that are restricted to a few specific models where the information can be located precisely [22]. However, many applications require to deal with a great variety of layouts, where relevant information is located in different positions. In this case, it is necessary to recognize the document layout and apply the appropriate reading strategy [23]. Several strategies have been proposed to achieve an accurate document classification based on the layout analysis and classification [1,4,5,23–25].

An invoice is a commercial document issued by a seller, containing details about the seller, the buyer, products, quantities, prices, etc., and usually a logo and tables. Hamza *et al.* [20] identify two main research directions in invoice classification. The first one concerns data-based systems and the second one concerns model-based systems. Data-based systems are usually used in heterogeneous document flows and extract different information from documents, such as tables [26], graphical features such as logos and trademarks [27], or the general layout [23]. On the contrary, model-based systems are used in homogeneous document flows, where similar documents arrive generally one after the other [28–31].

In this paper, we focus our attention on capturing visual similarity between different document images using global measures that do not require the analysis of the document layout. In the literature of document image classification, different measures of similarity have been used. Appiani *et al.* [23] design a criterion to compare the structural similarity between trees that represent the structure of a document. Shin and Doermann [24] use a similarity measure that considers spatial and layout structure. This measure quantifies the relatedness between two objects, combining structural and content features. Behera *et al.* [32] propose to measure the similarity between two images by computing the distance between their respective kernel density estimation of the histograms using the Minkowski distance or the intersection of the histograms.

## 3. Generalized Mutual Information

We review here three different mutual information generalizations based on the Kullback–Leibler distance, the difference between entropy and conditional entropy, and the Jensen–Tsallis divergence, respectively.

### 3.1. Mutual Information

From Equation (5), we have seen that mutual information can be expressed as the Kullback–Leibler distance between the joint probability distribution  $p(x, y)$  and the distribution  $p(x)p(y)$ . On the other hand, Tsallis [33] generalized the Kullback–Leibler distance in the following form:

$$KL_\alpha^T(p, q) = \frac{1}{\alpha - 1} \left( 1 - \sum_{x \in \mathcal{X}} \frac{p(x)^\alpha}{q(x)^{\alpha-1}} \right) \tag{12}$$

Thus, from Equations (5) and (12), Tsallis mutual information can be defined [33,34] as

$$\begin{aligned} MI_\alpha^T(X; Y) &= KL_\alpha^T(p(x, y), p(x)p(y)) \\ &= \frac{1}{1 - \alpha} \left( 1 - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \frac{p(x, y)^\alpha}{p(x)^{\alpha-1} p(y)^{\alpha-1}} \right) \end{aligned} \tag{13}$$

Although a simple substitution of  $MI$  for  $MI_\alpha^T$  can be used as an absolute similarity measure between two images, we focus our interest on a relative one. Such a decision is motivated by the better behavior of  $NMI$  with respect to  $MI$  [15]. Then, the generalization of  $NMI$  can be given by

$$NMI_\alpha^T(X; Y) = \frac{MI_\alpha^T(X; Y)}{H_\alpha^T(X, Y)} \tag{14}$$

Although  $NMI_\alpha^T(X; Y)$  is a normalized measure for  $\alpha \rightarrow 1$ , this is not true for other  $\alpha$  values as  $NMI^T$  can take values greater than 1. This measure is always positive and symmetric.

### 3.2. Mutual Entropy

Another way of generalizing mutual information is the so-called Tsallis mutual entropy [35]. The Tsallis mutual entropy is defined for  $\alpha > 1$  as

$$\begin{aligned} ME_\alpha^T(X; Y) &= H_\alpha^T(X) - H_\alpha^T(X|Y) = H_\alpha^T(Y) - H_\alpha^T(Y|X) \\ &= H_\alpha^T(X) + H_\alpha^T(Y) - H_\alpha^T(X, Y) \end{aligned} \tag{15}$$

This measure is positive and symmetric and Tsallis joint entropy  $H_\alpha^T(X, Y)$  is an upper bound [35]. Tsallis mutual entropy represents a kind of correlation between  $X$  and  $Y$ .

As in [35], the normalized Tsallis mutual entropy can be defined as

$$NME_\alpha^T(X; Y) = \frac{ME_\alpha^T(X; Y)}{H_\alpha^T(X, Y)} \tag{16}$$

Normalized mutual entropy takes values in the interval  $[0..1]$ , taking the value 0 if and only if  $X$  and  $Y$  are independent and  $\alpha = 1$ , and taking the value 1 if and only if  $X = Y$  [35].

### 3.3. Jensen–Tsallis Information

Since Tsallis entropy is a concave function for  $\alpha > 0$ , the Jensen–Shannon divergence can be extended to define the *Jensen–Tsallis divergence*:

$$JT_\alpha(\pi_1, \dots, \pi_n; p_1, \dots, p_n) = H_\alpha^T \left( \sum_{i=1}^n \pi_i p_i \right) - \sum_{i=1}^n \pi_i H_\alpha^T(p_i) \tag{17}$$

As we have seen in Equation (7), Jensen–Shannon divergence coincides with  $I(X; Y)$  when  $\{\pi_1, \dots, \pi_n\}$  is the marginal probability distribution  $p(x)$ , and  $\{p_1, \dots, p_n\}$  are the rows  $p(Y|x)$  of the probability conditional matrix of the channel. Then, for the channel  $X \rightarrow Y$ , a generalization of mutual information, which we call *Jensen–Tsallis Information* ( $JTI^\alpha$ ) can be expressed by

$$\begin{aligned} JTI_\alpha^T(X \rightarrow Y) &= JT_\alpha(p(x); p(Y|x)) = H_\alpha^T\left(\sum_{x \in \mathcal{X}} p(x)p(Y|x)\right) - \sum_{x \in \mathcal{X}} p(x)H_\alpha^T(Y|x) \\ &= H_\alpha^T(Y) - \sum_{x \in \mathcal{X}} p(x)H_\alpha^T(Y|x) \end{aligned} \tag{18}$$

For the reverse channel  $Y \rightarrow X$ , we have

$$JTI_\alpha^T(Y \rightarrow X) = JT_\alpha(p(x); p(Y|x)) = H_\alpha^T(X) - \sum_{y \in \mathcal{Y}} p(y)H_\alpha^T(X|y) \tag{19}$$

This measure is positive and, in general, non-symmetric with respect to the reversion of the channel. Thus,  $JTI_\alpha^T(X \rightarrow Y) \neq JTI_\alpha^T(Y \rightarrow X)$ . An upper bound of this measure is given by the Tsallis joint entropy:  $JTI_\alpha^T \leq H_\alpha^T(X, Y)$ . The Jensen–Tsallis divergence and its properties have been studied in [17,36].

Similar to the previous measures, a normalized version of  $JTI_\alpha^T$  can be defined as

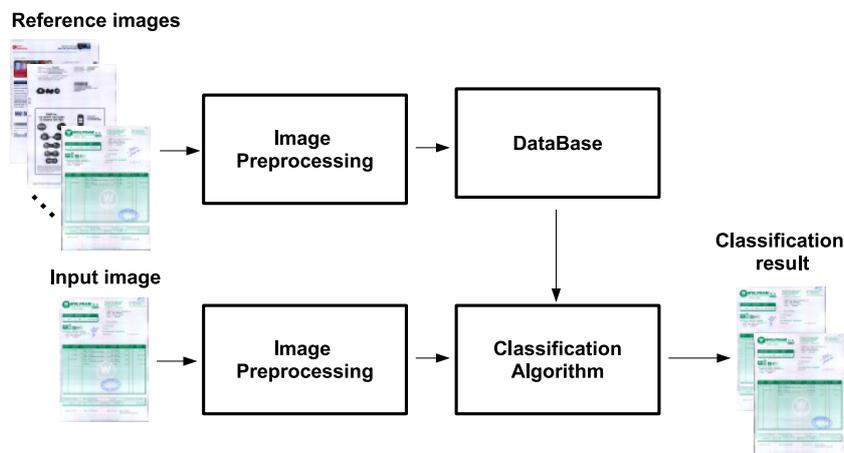
$$NJTI_\alpha^T(X \rightarrow Y) = \frac{JTI_\alpha^T(X \rightarrow Y)}{H_\alpha^T(X, Y)} \tag{20}$$

This measure will also take values in the interval  $[0, 1]$ .

#### 4. Overview

Large organizations and companies deal with a large amount of documents, such as invoices and receipts, which are usually scanned and stored in a database as image files. Then, some information of these images, such as the seller, the date, or the total amount of the invoice, is integrated in the database via manual editing or OCR techniques.

A critical issue for document analysis is the classification of similar documents. The documents of the same class can share some interesting information such as the background color, the document layout, the position of the relevant information on the image, or metadata, such as the seller. Once one document is grouped into a class, an specific processing for extracting the desired information can be designed depending on these features [23]. A simple way to define a class consists in taking a representative image. Then, we can create a database with the representative images and every new entry in the database is grouped into the class that the similarity between the new image and the representative image is maximum. A general scheme of our framework is represented in Figure 2. There are two different groups of images. The first one, formed by the *reference images*, is given by a document set where all documents are different between them and where each document represents a document type that identifies a class. This group of documents forms the document database. The second one, composed by the *input images*, is given by a set of documents that we want to use as classifier input with the aim of finding their corresponding class within the database of the reference images. Note that each input image has one, and only one, reference image, and different input images can have the same reference image.

**Figure 2.** Document classification pipeline.

The main goal of this paper is to analyze the application of the Tsallis-based generalizations of mutual information presented in the previous section to the document classification process. In the experiments of document classification carried out in this paper, we do not apply any spatial transform to the images as we assume that they are approximately aligned.

Another objective of this paper is to analyze the performance of the Tsallis-based generalizations of mutual information in aligning two documents. This is also a critical point since it allows us to find the spatial correspondence between an input document and a template. The registration framework used in this paper is represented in Figure 1.

## 5. Results and Discussion

To evaluate the similarity between two document images, the similarity measures presented in Section 3 have been implemented in Visual C++ .NET. In our experiments, we have dealt with a color invoice database, where 24-bits per pixel (8-bits for each RGB color channel) are used. These images usually present a complex layout, including pictures, logos, and highlighted areas. The database is composed by 51 reference invoices and 95 input invoices to be classified. It is required that each input invoice has one and only one reference invoice of the same type in the database. This reference invoice is similar (*i.e.*, from the same supplier) but not identical to the input invoice. In our first experiment on invoice classification, we assume that the images to be compared are fairly well aligned.

The reference and input invoices have been preprocessed using the method presented in [37] with the aim of correcting the skew error introduced during the scanning process. Although the skew error is corrected, they still present small translation errors between them. Preliminary experiments have shown that the best classification results are obtained for resolutions with height between 100 and 200 pixels. Note that this fact greatly speeds up the computation process as computation time is proportional to image resolution. In our experiments, all images have been scaled from the original scanning resolution (around  $2500 \times 3500$  pixels) to a height of 100 pixels, conveniently adjusting the image width to keep the aspect ratio of the images.

Let us remember that the main objective is to calculate the degree of similarity between each input invoice and all reference invoices. In this way, an ordered list of reference invoices, called *similarity list*,

can be obtained from the degree of similarity (from the highest to the lowest) between both the input and the reference invoices. Thus, we interpret that the first reference invoice of the list is the class assigned to the input invoice.

Next, two performance measures are considered for comparison purposes: the *percentage of success* and the *classification error*. The percentage of success is given by the number of correctly classified input invoices (*i.e.*, the corresponding reference image of the input invoice has been set to the first place in the similarity list) over the total number of inputs. Given an input invoice, the classification error is determined by the position of the corresponding reference invoice in the similarity list. If the reference invoice is chosen properly, this will be located at position 0 of the list.

Table 1 shows the two performance values for each measure and different  $\alpha$  values. Note that the values for the  $ME^T$  and  $NME^T$  measures are not shown for  $\alpha < 1$  since these measures are only defined for  $\alpha > 1$ . For  $\alpha = 1$ , the corresponding Shannon measures are considered in all cases. The first parameter represents the classification success in percentage and the second, in parentheses, represents the mean of classification error of the misclassified input invoices. As it can be seen, we can observe that the measures have a different behavior with respect to the  $\alpha$  values. While  $MI^T$  and  $NMI^T$  achieve the best classification success for  $\alpha$  values between 0.4 and 1.2, the rest of the measures ( $ME^T$ ,  $NME^T$ ,  $JTI^T$ ,  $NJTI^T$ ) perform better for  $\alpha$  values between 1.0 and 1.4. For these values, the normalized measures classify correctly all the documents. In general, the normalized measures perform much better than the corresponding non normalized ones. We have also tested the performance of  $SSD$  and  $CC$  measures and we have obtained a classification success of 70.53% and 88.42%, respectively. Note that these results are worse than the ones obtained using the proposed Tsallis-based measures.

**Table 1.** The percentage of classification success and the mean of classification error of the misclassified input invoices (in parentheses) for different measures and  $\alpha$  values.

| $\alpha$ values | $MI^T$ |        | $NMI^T$ |        | $ME^T$ |         | $NME^T$ |        | $JTI^T$ |        | $NJTI^T$ |        |
|-----------------|--------|--------|---------|--------|--------|---------|---------|--------|---------|--------|----------|--------|
| 0.2             | 96.84  | (1.67) | 92.63   | (1.29) |        |         |         |        | 71.58   | (6.26) | 69.47    | (9.00) |
| 0.4             | 98.95  | (2.00) | 100.0   | (0.00) |        |         |         |        | 80.00   | (2.58) | 81.05    | (3.39) |
| 0.6             | 98.95  | (1.00) | 100.0   | (0.00) |        |         |         |        | 90.53   | (1.67) | 89.47    | (1.30) |
| 0.8             | 98.95  | (1.00) | 100.0   | (0.00) |        |         |         |        | 94.74   | (1.40) | 94.74    | (1.00) |
| 1.0             | 98.95  | (2.00) | 100.0   | (0.00) | 98.95  | (2.00)  | 100.0   | (0.00) | 98.95   | (2.00) | 100.0    | (0.00) |
| 1.2             | 98.95  | (2.00) | 100.0   | (0.00) | 87.37  | (2.58)  | 100.0   | (0.00) | 97.89   | (1.00) | 100.0    | (0.00) |
| 1.4             | 97.89  | (1.50) | 97.89   | (1.00) | 78.95  | (6.40)  | 100.0   | (0.00) | 97.89   | (1.50) | 100.0    | (0.00) |
| 1.6             | 94.74  | (1.40) | 94.74   | (1.00) | 72.63  | (8.27)  | 97.89   | (1.50) | 96.84   | (1.33) | 97.89    | (1.00) |
| 1.8             | 89.47  | (2.10) | 90.53   | (1.56) | 67.37  | (9.65)  | 93.68   | (2.50) | 96.84   | (1.33) | 97.89    | (1.00) |
| 2.0             | 87.37  | (2.33) | 86.32   | (2.15) | 63.16  | (10.66) | 91.58   | (4.86) | 96.84   | (1.33) | 97.89    | (1.00) |
| 2.2             | 75.79  | (2.13) | 77.89   | (2.10) | 54.74  | (10.23) | 88.42   | (6.64) | 96.84   | (1.33) | 97.89    | (1.00) |
| 2.4             | 67.37  | (2.61) | 70.53   | (2.50) | 52.63  | (10.78) | 86.32   | (8.31) | 96.84   | (1.33) | 97.89    | (1.00) |
| 2.6             | 65.26  | (3.06) | 66.32   | (2.97) | 46.32  | (10.55) | 85.26   | (9.50) | 96.84   | (1.33) | 97.89    | (1.00) |
| 2.8             | 64.21  | (3.50) | 64.21   | (3.38) | 42.11  | (10.60) | 81.05   | (8.67) | 96.84   | (1.33) | 97.89    | (1.00) |
| 3.0             | 63.16  | (3.80) | 64.21   | (3.79) | 38.95  | (10.93) | 77.89   | (8.67) | 97.89   | (1.50) | 100.0    | (0.00) |

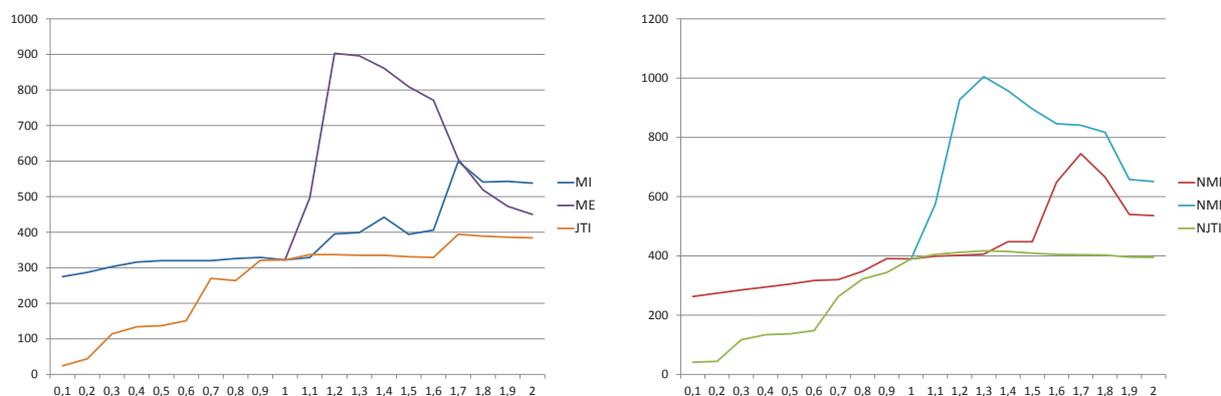
The classification error (shown between parentheses in Table 1) allows us to evaluate to what extent the classification is wrong when an invoice is misclassified. If this value is low, the system could suggest

a short list of class candidates and the user could select the correct one, while if this value is high that is not recommendable. From the results, we can conclude that, for a high range of  $\alpha$  values, methods identify the correct class in the first positions (for almost all cases the mean classification error is lower than 5). Thus, the short list can be taken into account for the final user interface design. The classification error obtained using *SSD* and *CC* measures is 20.25 and 6.73, respectively. Note also that Tsallis-based measures clearly outperform *SSD* and *CC* measures.

Our second experiment analyzes the capability of the Tsallis-based proposed measures to align two similar documents in the same spatial coordinates. In this case, two different features, robustness and accuracy, have been studied.

First, the robustness has been evaluated in terms of the partial image overlap. This has been done using the parameter AFA (Area of Function Attraction) introduced by Capek *et al.* [38]. This parameter evaluates the range of convergence of a registration measure to its global maximum, counting the number of pixels (*i.e.*,  $x - y$  translations in image space) from which the global maximum is reached by applying a maximum gradient method. Note that this global maximum may not necessarily be the optimal registration position. The AFA parameter represents the robustness with respect to the different initial positions of the images to be registered and with respect to the convergence to a local maximum of the similarity measure that leads to an incorrect registration. The higher the AFA, the wider the attraction basin of the measure. In this experiment, the images have been scaled to a height of 200 pixels, conveniently adjusting the width to keep the aspect ratio. In Figure 3, the left plot represents the results for the  $MI^T$ ,  $ME^T$ , and  $JTI^T$  measures with different  $\alpha$  values and the right plot represent the results for their corresponding normalized measures. As it can be seen, the best results are achieved for  $\alpha$  values greater than 1 for all the measures, being the mutual entropy the one that reaches the best results. As in the previous experiment, the normalized measures also perform better than the non normalized ones.

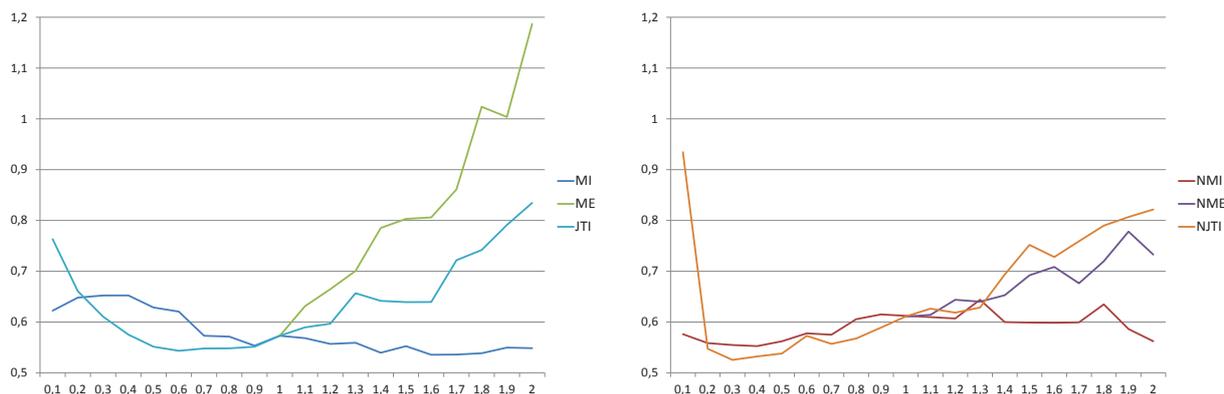
**Figure 3.** AFA parameter values with respect to the  $\alpha$  value for the  $MI^T$ ,  $ME^T$ , and  $JTI^T$  measures (left) and the corresponding normalized measures (right). AFA parameter evaluates the range of convergence of a registration measure to its global maximum.



The second feature that we will analyze for the alignment experiment is the accuracy. In this case, the general registration scheme of Figure 1 has been applied, where we have used the Powell’s method optimizer [39], a rigid transform (which only considers translation and rotation, but not scaling), and a linear interpolator. The registration process is applied to 18 images of the same class that are aligned with respect to a common template (scaling them to a height of 800 pixels and keeping the aspect

ratio). For each image with its original resolution (around  $2500 \times 3500$  pixels), 14 points have been manually identified and converted to the scaled space of a height of 800 pixels. The same process has been done with the template image. In order to quantify the registration accuracy, the points of each image have been moved using the final registration transform. The mean error, given by the average Euclidean distance between these moved points and the corresponding points in the template, has also been computed. In Figure 4, for each measure and each  $\alpha$  value, the mean error is plotted. In this case, we can not derive a general behavior.  $MI^T$  performs better for  $\alpha = 1.6$ , while  $NMI^T$  for  $\alpha = 0.4$ . In this case, the non normalized measure performs better than the normalized one. Both  $ME^T$  and  $NME^T$  do not outperform the corresponding Shannon measures ( $\alpha = 1$ ). Finally, Jensen–Tsallis information have a minimum in  $\alpha = 0.6$  and the accuracy diminishes when the  $\alpha$  value increases. Among all measures, the normalized Jensen–Tsallis information achieves the best results, obtaining the minimum error (and thus the maximum accuracy) for  $\alpha = 0.3$ .

**Figure 4.** Mean error at the final registration position for different measures and  $\alpha$  values for the  $MI^T$ ,  $ME^T$ , and  $JTI^T$  measures (left) and the corresponding normalized measures (right).



As a conclusion, for document classification, the best results have been obtained by the normalized measures, using  $\alpha$  values between 0.4 and 1.2 for  $NMI^T$  and between 1 and 1.4 for  $NME^T$  and  $NJTI^T$ . For document registration, the most robust results have been obtained by  $NME^T$  with  $\alpha = 1.3$  and the most accurate ones have been achieved by  $NJTI^T$  with  $\alpha = 0.3$ .

## 6. Conclusions

In this paper, we have analyzed the behavior of different similarity measures based on Tsallis entropy applied to document processing. Three different generalizations of mutual information, based respectively on Kullback–Leibler distance, the difference between entropy and conditional entropy, and the Jensen–Tsallis divergence, and their ratio with the Tsallis joint entropy have been tested. Two types of experiments have been carried out. First, the proposed measures have been applied to invoice classification, showing different behavior depending on the measure and the entropic index. Second, the document registration has been studied in terms of robustness and accuracy. While the highest robustness

is achieved for entropic indices higher than 1, the highest accuracy has been obtained for entropic indices clearly lower than 1.

In our future work, we will analyze the performance of the measures analyzed for different typologies of documents, such as scientific papers or journal pages, and further tests will be conducted on larger databases.

## Acknowledgements

This work has been funded in part with Grant Numbers TIN2010-21089-C03-01 from the Spanish Government and 2009-SGR-643 from the Catalan Government.

## References

1. Peng, H.; Long, F.; Chi, Z. Document image recognition based on template matching of component block projections. *IEEE Trans. Pattern Anal. Mach. Intell.* **2003**, *25*, 1188–1192.
2. Lopresti, D.P. String techniques for detecting duplicates in document databases. *IJDAR* **2000**, *2*, 186–199.
3. Tseng, L.Y.; Chen, R.C. The Recognition of Form Documents Based on Three Types of Line Segments. In *Proceedings of the 4th International Conference on Document Analysis and Recognition, ICDAR'97*, Ulm, Germany, 18–20 August 1997; pp. 71–75.
4. Hu, J.; Kashi, R.S.; Wilfong, G.T. Document Image Layout Comparison and Classification. In *Proceedings of the Fifth International Conference on Document Analysis and Recognition, ICDAR'99*, Bangalore, India, 20–22 September 1999; pp. 285–288.
5. Shin, C.; Doermann, D.S.; Rosenfeld, A. Classification of document pages using structure-based features. *IJDAR* **2001**, *3*, 232–247.
6. Hajnal, J.; Hawkes, D.; Hill, D. *Medical Image Registration*; CRC Press Inc.: Boca Raton, FL, USA, 2001.
7. Portes de Albuquerque, M.; Esquef, I.; Gesualdi Mello, A.; Portes de Albuquerque, M. Image thresholding using Tsallis entropy. *Pattern Recognit. Lett.* **2004**, *25*, 1059–1065.
8. Cover, T.M.; Thomas, J. *Elements of Information Theory*; John Wiley and Sons Inc.: Hoboken, NJ, USA, 1991.
9. Yeung, R.W. *Information Theory and Network Coding*; Springer: Berlin, Heidelberg, Germany, 2008.
10. Burbea, J.; Rao, C.R. On the convexity of some divergence measures based on entropy functions. *IEEE Trans. Inf. Theory* **1982**, *28*, 489–495.
11. Tsallis, C. Possible generalization of Boltzmann-Gibbs statistics. *J. Stat. Phys.* **1988**, *52*, 479–487.
12. Hill, D.L.G.; Batchelor, P.G.; Holden, M.; Hawkes, D.J. Medical image registration. *Phys. Med. Biol.* **2001**, *46*, R1–R45.
13. Maes, F.; Collignon, A.; Vandermeulen, D.; Marchal, G.; Suetens, P. Multimodality image registration by maximization of mutual information. *IEEE Trans. Med. Imaging* **1997**, *16*, 187–198.

14. Viola, P.A. Alignment by Maximization of Mutual Information. PhD thesis, MIT Artificial Intelligence Laboratory (TR 1548), Cambridge, MA, USA, 1995.
15. Studholme, C. Measures of 3D Medical Image Alignment. PhD thesis, Computational Imaging Science Group, Division of Radiological Sciences, United Medical and Dental school's of Guy's and St Thomas's Hospitals, University of London, London, UK, 1997.
16. Wachowiak, M.P.; Smolikova, R.; Tourassi, G.D.; Elmaghraby, A.S. Similarity Metrics Based on Non-additive Entropies for 2D-3D Multimodal Biomedical Image Registration. In *Proceedings of SPIE Medical Imaging 2003: Image Processing*, San Diego, CA, USA, 15 May 2003; Volume 5032, pp. 1090–1100.
17. Bardera, A.; Feixas, M.; Boada, I. Normalized Similarity Measures for Medical Image Registration. In *Proceedings of Medical Imaging SPIE 2004: Image Processing*, San Diego, CA, USA, 12 May 2004; Volume 5370, pp. 108–118.
18. Mohamed, W.; Ben Hamza, A. Nonextensive Entropic Image Registration. In *Image Analysis and Recognition*; Springer: Berlin, Heidelberg, Germany, 2009; Volume 5627, pp. 116–125.
19. Khader, M.; Ben Hamza, A.; Bhattacharya, P. Multimodality Image Alignment Using Information-Theoretic Approach. In *Image Analysis and Recognition*; Springer: Berlin, Heidelberg, Germany, 2010; Volume 6112, pp. 30–39.
20. Hamza, H.; Belaïd, Y.; Belaïd, A.; Chaudhuri, B.B. An End-to-End Administrative Document Analysis System. In *Proceedings of the 2008 The 8th IAPR International Workshop on Document Analysis Systems, DAS'08*, Nara, Japan, 16–19 September 2008; pp. 175–182.
21. Chen, N.; Blostein, D. A survey of document image classification: Problem statement, classifier architecture and performance evaluation. *Int. J. Doc. Anal. Recognit.* **2007**, *10*, 1–16.
22. Trier, Ø.D.; Jain, A.K.; Taxt, T. Feature extraction methods for character recognition—A survey. *Pattern Recognit.* **1996**, *29*, 641–662.
23. Appiani, E.; Cesarini, F.; Colla, A.M.; Diligenti, M.; Gori, M.; Marinai, S.; Soda, G. Automatic document classification and indexing in high-volume applications. *IJDAR* **2001**, *4*, 69–83.
24. Shin, C.; Doermann, D.S. Document Image Retrieval Based on Layout Structural Similarity. In *Proceedings of the 2006 International Conference on Image Processing, Computer Vision and Pattern Recognition, IPCV'06*, Las Vegas, NV, USA, 26–29 June 2006; pp. 606–612.
25. Gupta, M.D.; Sarkar, P. A Shared Parts Model for Document Image Recognition. In *Proceedings of the 9th International Conference on Document Analysis and Recognition, ICDAR'07*, Curitiba, Brazil, 23–26 September 2007; Volume 2, pp. 1163–1172.
26. Costa e Silva, A.; Jorge, A.M.; Torgo, L. Design of an end-to-end method to extract information from tables. *IJDAR* **2006**, *8*, 144–171.
27. Alippi, C.; Pessina, F.; Roveri, M. An Adaptive System for Automatic Invoice-Documents Classification. In *Proceedings of the IEEE International Conference on Image Processing, ICIP'05*, Genova, Italy, 11–14 September 2005; Volume 2, pp. II-526–II-529.
28. Arai, H.; Odaka, K. Form Processing based on Background Region Analysis. In *Proceedings of the 4th International Conference on Document Analysis and Recognition, ICDAR'97*, Ulm, Germany, 18–20 August 1997; pp. 164–169.

29. Cesarini, F.; Gori, M.; Marinai, S.; Soda, G. INFORMys: A flexible invoice-like form-reader system. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 730–745.
30. Tang, Y.Y.; Liu, J. Information Acquisition and Storage of Forms in Document Processing. In *Proceedings of the 4th International Conference on Document Analysis and Recognition, ICDAR'97*, Ulm, Germany, 18–20 August 1997; pp. 170–174.
31. Duygulu, P.; Atalay, V. A hierarchical representation of form documents for identification and retrieval. *IJDAR* **2002**, *5*, 17–27.
32. Behera, A.; Lalanne, D.; Ingold, R. Combining color and layout features for the identification of low-resolution documents. *Int. J. Signal Process.* **2005**, *2*, 7–14.
33. Tsallis, C. Generalized entropy-based criterion for consistent testing. *Phys. Rev. E* **1998**, *58*, 479–487.
34. Taneja, I.J. Bivariate measures of type  $\alpha$  and their applications. *Tamkang J. Math.* **1988**, *19*, 63–74.
35. Furuichi, S. Information theoretical properties of Tsallis entropies. *J. Math. Phys.* **2006**, *47*, 023302.
36. Ben Hamza, A. Nonextensive information-theoretic measure for image edge detection. *J. Electron. Imaging* **2006**, *15*, 13011.1–13011.8.
37. Gatos, B.; Papamarkos, N.; Chamzas, C. Skew detection and text line position determination in digitized documents. *Pattern Recognit.* **1997**, *30*, 1505–1519.
38. Capek, M.; Mroz, L.; Wegenkittl, R. Robust and Fast Medical Registration of 3D-Multi-Modality Data Sets. In *Proceedings of Medicon 2001*, Pula, Croatia, 12–15 June 2001; pp. 515–518.
39. Press, W.; Teukolsky, S.; Vetterling, W.; Flannery, B. *Numerical Recipes in C*; Cambridge University Press: Cambridge, UK, 1992.

© 2011 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>.)