Universitat de Girona

INFORMATION THEORY TECHNIQUES FOR MULTIMEDIA DATA CLASSIFICATION AND RETRIEVAL

Marius Vila Duran

Dipòsit legal: Gi. 1379-2015
http://hdl.handle.net/10803/302664

Universitat de Girona

DOCTORAL THESIS

Information theory techniques for multimedia data classification and retrieval

Marius VILA DURAN

2015

Universitat de Girona

DOCTORAL THESIS

Information theory techniques for multimedia data classification and retrieval

Author:
Marius VILA DURAN

2015

Doctoral Programme in Technology

Advisors:
Dr. Miquel FEIXAS FEIXAS
Dr. Mateu SBERT CASASAYAS

This manuscript has been presented to opt for the doctoral degree from the University of Girona

List of publications

Publications that support the contents of this thesis:

- "*Tsallis Mutual Information for Document Classification*", Marius Vila, Anton Bardera, Miquel Feixas, Mateu Sbert. Entropy, vol. 13, no. 9, pages 1694-1707, 2011.
- "*Tsallis entropy-based information measure for shot boundary detection and keyframe selection*", Marius Vila, Anton Bardera, Qing Xu, Miquel Feixas, Mateu Sbert. Signal, Image and Video Processing, vol. 7, no. 3, pages 507-520, 2013.
- "*Analysis of image informativeness measures*", Marius Vila, Anton Bardera, Miquel Feixas, Philippe Bekaert, Mateu Sbert. IEEE International Conference on Image Processing pages 1086-1090, October 2014.
- "*Image-based Similarity Measures for Invoice Classification*", Marius Vila, Anton Bardera, Miquel Feixas, Mateu Sbert. Submitted.

List of figures

List of tables

Agraïments

Primer de tot vull agrair el suport, les idees i l'ajuda rebuda per part dels meus directors Miquel Feixas i Mateu Sbert. La seva confiança m'ha ajudat a pensar que podia començar i acabar aquesta tesi amb èxit. També vull agrair especialment a l'Anton Bardera les seves explicacions, consells i contribucions, ja que han estat imprescindibles durant tots els anys que he dedicat a aquest treball. Cada cop que he estat davant un problema del que no trobava cap solució, ells em feien veure que sempre hi ha algun camí possible. També vull agrair a en Qing Xu i en Philippe Bekaert les seves contribucions al meu treball. Sens dubte he pogut formar part d'un gran equip. Gràcies també a l'Alex Brusi per la col·laboració que hem mantingut des de l'inici i per proporcionar-me els recursos necessaris per acabar la tesi un cop finalitzada la beca d'estudis. Agraeixo també tant l'ajuda rebuda com les bones estones que he passat amb els companys de despatx i de l'Escola Politècnica Superior: Anton, Ferran, Marc, Roger, Xavi, Pau, Ester, Yago, Marta, Francesc, Imma, Olga, June, Miquel, Nacho, Joan...

Finalment, agrair també a la família (pares, avis, germà, oncles, cosins, sogra, cunyats, nebots...) i als amics el paper tant important que han tingut durant aquests anys de doctorat. Primer de tot, gràcies a la meva parella Mafer pel recolzament, la confiança cega i els ànims que m'ha donat en tot moment, especialment, en aquells moments en que penses que no pots més, i pel gran nombre de sacrificis que ha fet per tal que jo pogués centrar-me en els meus objectius, espero poder-ho compensar a partir d'ara. Gràcies a la meva filla Maika, que sempre aconsegueix animar-me els dies rebent-me amb un gran somriure quan arribo a casa i acomiadant-se sempre amb el seu "que te portes bien en el trabajo!!". Gràcies als meus pares, germà, i avis per confiar en mi més del que hi confio jo mateix i per intentar facilitar-me sempre el camí. Tant amb les seves paraules com amb els seus actes, han estat sempre un exemple a seguir i m'han ensenyat que, per més difícil que sembli, amb força de voluntat es pot assolir qualsevol objectiu. Un record especial al meu avi "Poli", que tot i que no podrà veure el final d'aquest camí, va estar al meu costat la major part d'aquest. I gràcies als amics, Bruno, Lorena i al meu fillol Alex que més que amics són família i que han estat sempre pel que ha fet falta, tant en els bons com en els mals moments. I per acabar gràcies a totes aquelles persones que poder no veig tant sovint com m'agradaria però que en algun moment o altre s'han interessat per mi Aleix, Inma, Toni, Cristofer, Raquel, Luisito, Laura, Enric, Albert...

Acknowledgements

The work in this thesis has been supported by an FI grant of Generalitat de Catalunya (Catalan Government), TIN2010-21089-C03-01 and TIN2013-47276-C6-1-R of Spanish Government, the European Social Fund, and by grant number 2009-SGR-643 and 2014-SGR-1232 of Generalitat de Catalunya (Catalan Government).

Contents

Information theory techniques for multimedia data classification and retrieval

Abstract: We are in the information age where most data is stored in digital format. Thus, the management of digital documents and videos requires the development of efficient techniques for automatic analysis. Among them, capturing the similarity or dissimilarity between different document images or video frames are extremely important.

In this thesis, we first analyze for several image resolutions the behavior of three different families of image-based similarity measures applied to invoice classification. In these three set of measures, the computation of the similarity between two images is based, respectively, on intensity differences, mutual information, and normalized compression distance. As the best results are obtained with mutual information-based measures, we proceed to investigate the application of three different Tsallis-based generalizations of mutual information for different entropic indexes. These three generalizations derive respectively from the Kullback-Leibler distance, the difference between entropy and conditional entropy, and the Jensen-Shannon divergence.

In relation to digital video processing, we propose two different information-theoretic approaches based, respectively, on Tsallis mutual information and Jensen-Tsallis divergence to detect the abrupt shot boundaries of a video sequence and to select the most representative keyframe of each shot.

Finally, Shannon entropy has been commonly used to quantify the image informativeness. The main drawback of this measure is that it does not take into account the spatial distribution of pixels. In this thesis, we analyze four information-theoretic measures that overcome this limitation. Three of them (entropy rate, excess entropy, and erasure entropy) consider the image as a stationary stochastic process, while the fourth (partitional information) is based on an information channel between image regions and histogram bins.

Tècniques de la teoria de la informació per a la classificació i recuperació de dades multimèdia

Resum: Ens trobem a l'era de la informació on la majoria de les dades s'emmagatzemen en format digital. Per tant, la gestió de documents i vídeos digitals requereix el desenvolupament de tècniques eficients per a l'anàlisi automàtic. Entre elles, la captura de la similitud o dissimilitud entre diferents imatges de documents o fotogrames de vídeo és extremadament important.

En aquesta tesi, analitzem, a diverses resolucions d'imatge, el comportament de tres famílies diferents de mesures basades en similitud d'imatges i aplicades a la classificació de factures. En aquests tres conjunt de mesures, el càlcul de la similitud entre dues imatges es basa, respectivament, en les diferències d'intensitat, en la informació mútua, i en la distància de compressió normalitzada. Degut a que els millors resultats s'obtenen amb les mesures basades en la informació mútua, es procedeix a investigar l'aplicació de tres generalitzacions de la informació mútua basades en Tsallis en diferents índexs entròpics. Aquestes tres generalitzacions es deriven respectivament de la distància de Kullback-Leibler, la diferència entre l'entropia i entropia condicional, i la divergència de Jensen-Shannon.

En relació al processament de vídeo digital, proposem dos enfocaments diferents de teoria de la informació basats respectivament en la informació mútua de Tsallis i en la divergència de Jensen-Tsallis, per detectar els límits d'un pla cinematogràfic en una seqüència de vídeo i per seleccionar el fotograma clau més representatiu de cada pla.

Finalment, l'entropia de Shannon s'ha utilitzat habitualment per quantificar la informativitat d'una imatge. El principal inconvenient d'aquesta mesura és que no té en compte la distribució espacial dels píxels. En aquesta tesi, s'analitzen quatre mesures de teoria de la informació que superen aquesta limitació. Tres d'elles (*entropy rate*, *excess entropy* i *erasure entropy*) consideren la imatge com un procés estocàstic estacionari, mentre que la quarta (*partitional information*) es basa en un canal d'informació entre les regions d'una imatge i els intervals de l'histograma.

Técnicas de la teoría de la información para la clasificación y recuperación de datos multimedia

Resumen: Estamos en la era de la información donde la mayoría de los datos se almacenan en formato digital. Por lo tanto, la gestión de documentos y videos digitales requiere el desarrollo de técnicas eficientes para el análisis automático. Entre ellas, la captura de la similitud o disimilitud entre diferentes imágenes de documentos o fotogramas de vídeo es extremadamente importante.

En esta tesis, analizamos, a varias resoluciones de imagen, el comportamiento de tres familias diferentes de medidas basadas en similitud de imágenes y aplicadas a la clasificación de facturas. En estos tres conjunto de medidas, el cálculo de la similitud entre dos imágenes se basa, respectivamente, en las diferencias de intensidad, en la información mutua, y en la distancia de compresión normalizada. Debido a que los mejores resultados se obtienen con las medidas basadas en la información mutua, se procede a investigar la aplicación de tres generalizaciones de la información mutua basadas en Tsallis con diferentes índices entrópicos. Estas tres generalizaciones se derivan respectivamente de la distancia de Kullback-Leibler, la diferencia entre la entropía y entropía condicional, y la divergencia de Jensen-Shannon.

En relación al procesamiento de vídeo digital, proponemos dos enfoques diferentes de teoría de la información basados respectivamente en la información mutua de Tsallis y en la divergencia de Jensen-Tsallis, para detectar los límites de un plano cinematográfico en una secuencia de video y para seleccionar el fotograma clave más representativo de cada plano.

Por último, la entropía de Shannon se ha utilizado habitualmente para cuantificar la informatividad de una imagen. El principal inconveniente de esta medida es que no tiene en cuenta la distribución espacial de los píxeles. En esta tesis, se analizan cuatro medidas de teoría de la información que superan esta limitación. Tres de ellas (*entropy rate*, *excess entropy* y *erasure entropy*) consideran la imagen como un proceso estocástico estacionario, mientras que la cuarta (*partitional information*) se basa en un canal de información entre las regiones de una imagen y los intervalos del histograma.

CHAPTER 1

Introduction

Contents

1.1 Motivation

We are in the information age where most data is stored in digital format. Thus, multimedia databases management techniques became very popular. The development of intelligent systems capable of dealing with this kind of data efficiently and effectively has become an extremely important task. Therefore, it is absolutely necessary and critical to find good metrics in order to develop similarity measures for multimedia data classification and retrieval.

Based on the capability of scanners to transform large amounts of documents to digital images, big organizations and companies use information systems to deal with scanned images, which are usually stored in a database as image files. Some information of these images, such as, for instance, the provider, the date, or the total amount in an invoice, is integrated in the database via manual editing or OCR techniques. To automatize the postprocessing tasks, such as binarization and text extraction, the classification of these documents in different types can be very useful. Thus, the automatic classification of this type of documents with a geometric layout is a topic of major interest for many office applications.

Similar to digital images, in the last decades, the availability of digital video is growing at an exponential rate. In this context, video summarization constitutes one of the major goals of multimedia research. Video shot boundary detection, or the segmentation of a video sequence in its constituent shots, is a fundamental step in video data management. Another step is keyframe selection within each shot.

Also, an important and not very well studied issue is the information content of an image. But what is image information? How can we quantify this content? In this thesis, we analyze the performance of several techniques based on information theory, applied in the scopes of image informativeness, document classification, and video processing.

1.2 Objectives

The main goal of this thesis is to find good metrics based on information theory with the aim of developing robust similarity measures for multimedia data classification and retrieval.

To reach this objective we aim to

- Analyze for several image resolutions the behaviour of several families of image-based similarity measures applied to invoice classification.
- Investigate the application of several Tsallis-based generalizations of mutual information to analyze the similarity between scanned invoices.
- Analyze the behaviour of several information-theoretic measures to detect the video discontinuities and to extract the most representative keyframes.
- Investigate the application of several information theoretic measures to quantify the image informativeness.

1.3 Thesis outline

This dissertation is organized in seven chapters. Apart from this introduction, the thesis is divided into six chapters:

- Chapter 2: **Background**

  In this chapter, the background on document classification and video processing required for the comprehension of the main issues that are going to be analyzed in this thesis is introduced. The main concepts of information theory are also reviewed since they are the basis of most of our contributions.

- Chapter 3: **Image-based similarity measures for document classification**

  In this chapter, we analyze for several image resolutions the behaviour of three different families of image-based similarity measures applied to invoice classification. In these three groups of measures, the computation of the similarity between two images is based, respectively, on intensity differences, mutual information-based measures, and the normalized compression distance.

  The content of this chapter is presented in "*Image-based Similarity Measures for Invoice Classification*", Marius Vila, Anton Bardera, Miquel Feixas, Mateu Sbert. Submitted.

- Chapter 4: **Tsallis mutual information for document classification**

  In this chapter, 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-Shannon 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 invoice classification and registration.

The content of this chapter has been published in"*Tsallis Mutual Information for Document Classification*", Marius Vila, Anton Bardera, Miquel Feixas, Mateu Sbert. Entropy, vol. 13, no. 9, pages 1694-1707, 2011 [Vila 2011].

- Chapter 5: **Shot boundary detection and keyframe selection**

  In this chapter, we propose two different information-theoretic approaches to detect the abrupt shot boundaries of a video sequence. These approaches are, respectively, based on two information measures, Tsallis mutual information and Jensen-Tsallis divergence, that are used to quantify the similarity between two frames. Both measures are also used to find out the most representative keyframe of each shot. Several experiments analyze the behavior of the proposed measures for different color spaces (RGB, HSV, and Lab), regular binnings, and entropic indices.

  The content of this chapter has been published in "*Tsallis entropy-based information measure for shot boundary detection and keyframe selection*", Marius Vila, Anton Bardera, Qing Xu, Miquel Feixas, Mateu Sbert. Signal, Image and Video Processing, vol. 7, no. 3, pages 507-520, 2013 [Vila 2013].

- Chapter 6: **Image informativeness**

  In this chapter, we analyze the performance of four information-theoretic measures when the image informativeness is quantified. Three of them (entropy rate, excess entropy, and erasure entropy) consider the image as a stationary stochastic process, while the fourth (partitional information) is based on an information channel between image regions and histograms bins.

  The content of this chapter has been published in "*Analysis of image informativeness measures*", Marius Vila, Anton Bardera, Miquel Feixas, Philippe Bekaert, Mateu Sbert. IEEE International Conference on Image Processing pages 1086-1090, October 2014 [Vila 2014].

- Chapter 7: **Conclusions and future work**

  In this chapter, both the conclusions and the future work of the thesis are presented, along with a summary of the publications related with this thesis.

CHAPTER 2

Background

Contents

2.1 Introduction

In this chapter, we review the basic concepts of information theory used in this thesis, together with some previous work on document classification and video processing.

This chapter is structured as follows. Section 2.2 presents the concepts of information theory that will be used in this thesis. Section 2.3 briefly summarizes some of the most important aspects to be taken into account to solve the document classification problems, referring several approaches presented by different authors. Section 2.4 provides an overview on video processing where shot boundary detection and keyframe selection concepts are defined.

2.2 Information theory tools

In 1948, Claude Shannon published a paper entitled "A mathematical theory of communication" [Shannon 1948] which marks the beginning of information theory. In this paper, Shannon defined measures such as entropy and mutual information[1], and introduced the fundamental laws of data compression and transmission.

[1] In Shannon's paper, the mutual information is called rate of transmission.

Information theory deals with the transmission, storage, and processing of information and is used in fields such as physics, computer science, mathematics, statistics, economics, biology, linguistics, neurology, learning, image processing, and computer graphics.

In information theory, *information* is simply the outcome of a selection among a finite number of possibilities and an information source is modelled as a random variable or a random process. The classical measure of information, Shannon entropy, expresses the information content or the uncertainty of a single random variable. It is also a measure of the dispersion or diversity of a probability distribution of observed events. For two random variables, their mutual information is a measure of the dependence between them. Mutual information plays an important role in the study of a *communication channel*, a system in which the output depends probabilistically on its input [Cover 1991, Verdú 1998, Yeung 2008].

This section presents Shannon's information measures (entropy, conditional entropy, and mutual information) and their most basic properties. The information bottleneck method is also introduced. Good references of information theory are the books by Cover and Thomas [Cover 1991], and Yeung [Yeung 2008].

2.2.1 Entropy

Let X be a discrete random variable with alphabet $\mathcal{X}$ and probability distribution $\{p(x)\}$, where $p(x) = \Pr[X = x]$ and $x \in \mathcal{X}$. In this thesis, $\{p(x)\}$ will be also denoted by $p(X)$ or simply p. This notation will be extended to two or more random variables.

The entropy $H(X)$ of a discrete random variable X is defined by

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log p(x), \tag{2.1}$$

where the summation is over the corresponding alphabet and the convention $0\log 0 = 0$ is taken.

In this thesis, logarithms are taken in base 2 and, as a consequence, entropy is expressed in bits. The convention $0\log 0 = 0$ is justified by continuity since $x\log x \to 0$ as $x \to 0$. The term $-\log p(x)$ represents the information content (or uncertainty) associated with the result x. Thus, the entropy gives us the average amount of information (or uncertainty) of a random variable. Note that the entropy depends only on the probabilities. We can use interchangeably the notation $H(X)$ or $H(p)$ for the entropy, where p stands for the probability distribution $p(X)$.

Some relevant properties [Shannon 1948] of the entropy are:

- $0 \le H(X) \le \log|\mathcal{X}|$.
  - $H(X) = 0$ if and only if all the probabilities except one are zero, this one having the unit value, i.e., when we are certain of the outcome.

  - $H(X) = \log|\mathcal{X}|$ when all the probabilities are equal, i.e., we have maximum uncertainty.

- If the probabilities are equalized, entropy increases.

The binary entropy (Fig. 2.1) of a random variable X with alphabet $\{x_1, x_2\}$ and probability distribution $\{p, 1-p\}$ is given by

$$H(X) = -p\log p - (1-p)\log(1-p). \tag{2.2}$$

Note that the maximum entropy is $H(X) = 1$ bit when $p = \frac{1}{2}$.

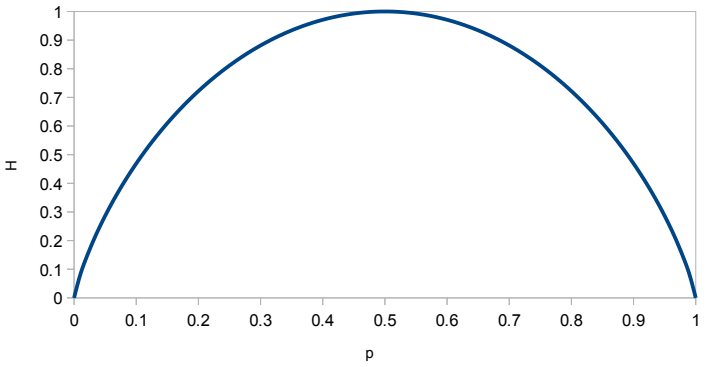


Figure 2.1: Plot of binary entropy.

The definition of entropy is now extended to a pair of random variables. The joint entropy $H(X,Y)$ of a pair of discrete random variables X and Y with a joint probability distribution $p(X,Y) = \{p(x,y)\}$ is defined by

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y), \tag{2.3}$$

where $p(x,y) = \Pr[X = x, Y = y]$ is the joint probability of x and y.

The conditional entropy $H(Y|X)$ of a random variable Y given a random variable X is defined as the expected value of the entropies of the conditional distributions:

$$\begin{aligned} H(Y|X) &= \sum_{x\in\mathcal{X}} p(x)H(Y|X=x) = \sum_{x\in\mathcal{X}} p(x)\left(-\sum_{y\in\mathcal{Y}} p(y|x)\log p(y|x)\right) \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x), \end{aligned} \tag{2.4}$$

where $p(y|x) = \Pr[Y = y|X = x]$ is the conditional probability of y given x.

The Bayes theorem relates marginal probabilities $p(x)$ and $p(y)$, conditional probabilities $p(y|x)$ and $p(x|y)$, and joint probabilities $p(x,y)$:

$$p(x,y) = p(x)p(y|x) = p(y)p(x|y). \tag{2.5}$$

If X and Y are independent, then $p(x,y) = p(x)p(y)$. Marginal probabilities can be obtained from $p(x,y)$ by summation: $p(x) = \sum_{y\in\mathscr{Y}} p(x,y)$ and $p(y) = \sum_{x\in\mathscr{X}} p(x,y)$.

The conditional entropy can be thought of in terms of a communication or *information channel* $X \to Y$ whose output Y depends probabilistically on its input X. This information channel is characterized by a transition probability matrix which determines the conditional distribution of the output given the input [Cover 1991]. Hence, $H(Y|X)$ corresponds to the uncertainty in the channel output from the sender's point of view, and vice versa for $H(X|Y)$. Note that in general $H(Y|X) \neq H(X|Y)$. In this thesis, the conditional probability distribution of Y given x will be denoted by $p(Y|x)$ and the transition probability matrix (i.e., the matrix whose rows are given by $p(Y|x)$) will be denoted by $p(Y|X)$.

The following properties hold:

- $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$.
- $H(X,Y) \leq H(X) + H(Y)$.
- $H(X) \geq H(X|Y) \geq 0$.
- If X and Y are independent, then $H(Y|X) = H(Y)$ since $p(y|x) = p(y)$ and, consequently, $H(X,Y) = H(X) + H(Y)$ (i.e., entropy is additive for independent random variables).

2.2.2 Kullback-Leibler divergence and mutual information

We now introduce two new measures, Kullback-Leibler divergence and mutual information, which quantify the distance between two probability distributions and the shared information between two random variables, respectively.

The relative entropy or Kullback-Leibler divergence [Kullback 1951] $D_{\mathrm{KL}}(p\|q)$ between two probability distributions p and q, that are defined over the alphabet $\mathscr{X}$, is defined by

$$D_{\mathrm{KL}}(p\|q) = \sum_{x\in\mathscr{X}} p(x) \log \frac{p(x)}{q(x)}. \tag{2.6}$$

The conventions that $0 \log \frac{0}{0} = 0$ and $a \log \frac{a}{0} = \infty$ if $a > 0$ are adopted. The Kullback-Leibler divergence satisfies the information inequality

$$D_{\mathrm{KL}}(p\|q) \geq 0, \tag{2.7}$$

with equality if and only if $p = q$. The Kullback-Leibler divergence is also called information divergence [Csiszár 2004] or informational divergence [Yeung 2008], and it is not strictly a metric[2] since it is not symmetric and does not satisfy the

[2] A metric between x and y is defined as a function $d(x,y)$ that fulfills the following properties: (1) non-negativity: $d(x,y) \geq 0$, (2) identity: $d(x,y) = 0$ if and only if $x = y$, (3) symmetry: $d(x,y) = d(y,x)$, and (4) triangle inequality: $d(x,y) + d(y,z) \geq d(x,z)$.

triangle inequality. The Kullback-Leibler divergence is "a measure of the inefficiency of assuming that the distribution is q when the true distribution is p" [Cover 1991].

The mutual information $I(X;Y)$ between two random variables X and Y is defined by

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{2.8}$$

$$= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = \sum_{x\in\mathcal{X}} p(x) \sum_{y\in\mathcal{Y}} p(y|x) \log \frac{p(y|x)}{p(y)}. \tag{2.9}$$

Mutual information represents the amount of information that one random variable, the input of the channel, contains about a second random variable, the output of the channel, and vice versa. That is, mutual information expresses how much the knowledge of Y decreases the uncertainty of X, and vice versa. $I(X;Y)$ is a measure of the shared information or dependence between X and Y. Thus, if X and Y are independent, then $I(X;Y) = 0$. Note that the mutual information can be expressed as the relative entropy between the joint distribution and the product of marginal distributions:

$$I(X;Y) = D_{\mathrm{KL}}(p(X,Y) \| p(X)p(Y)). \tag{2.10}$$

Mutual information $I(X;Y)$ fulfills the following properties:

- $I(X;Y) \geq 0$ with equality if and only if X and Y are independent
- $I(X;Y) = I(Y;X)$
- $I(X;Y) = H(X) + H(Y) - H(X,Y)$
- $I(X;Y) \leq \min\{H(X), H(Y)\}$
- $I(X;X) = H(X)$

The relationship between Shannon's information measures can be expressed by a Venn diagram, as shown in Fig. 2.2[3]. The correspondence between Shannon's information measures and set theory is discussed in [Yeung 2008].

The normalized mutual information NMI can be defined as

$$NMI(X;Y) = \frac{I(X;Y)}{H(X,Y)}, \tag{2.11}$$

where NMI takes values in the range [0,1].

2.2.3 Inequalities

In this section, we introduce some inequalities that are essential in the study of information theory.

[3]The information diagram does not include the universal set as in a usual Venn diagram.

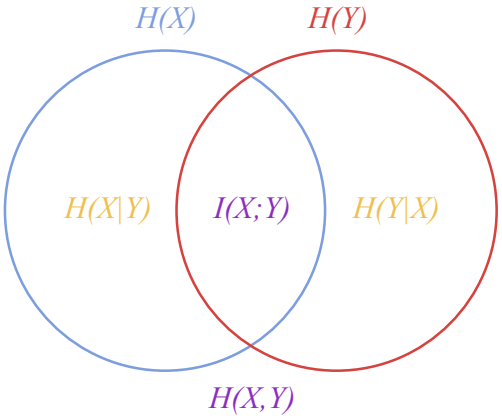


Figure 2.2: The information diagram represents the relationship between Shannon's information measures. Observe that $I(X;Y)$ and $H(X,Y)$ are represented, respectively, by the intersection and the union of the information in X (represented by $H(X)$) with the information in Y (represented by $H(Y)$). $H(X|Y)$ is represented by the difference between the information in X and the information in Y, and vice versa for $H(Y|X)$.

2.2.3.1 Jensen's inequality

In this section, we introduce the concepts of convexity and concavity. Many important inequalities and results in information theory are obtained from the concavity of the logarithmic function.

A function $f(x)$ is convex over an interval $[a,b]$ (the graph of the function lies below any chord) if for every $x_1, x_2 \in [a,b]$ and $0 \leq \lambda \leq 1$,

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2). \tag{2.12}$$

A function is strictly convex if equality holds only if $\lambda = 0$ or $\lambda = 1$.

A function $f(x)$ is concave (the graph of the function lies above any chord) if $-f(x)$ is convex.

For instance, x^2 and $x \log x$ (for $x > 0$) are strictly convex functions, and $\log x$ (for $x > 0$) is a strictly concave function.

Jensen's inequality can be expressed as follows. If f is a convex function on the interval $[a,b]$, then

$$\sum_{i=1}^{n} \lambda_i f(x_i) - f\left(\sum_{i=1}^{n} \lambda_i x_i\right) \geq 0, \tag{2.13}$$

where $0 \leq \lambda \leq 1$, $\sum_{i=1}^{n} \lambda_i = 1$, and $x_i \in [a,b]$. If f is a concave function, the inequality is reversed. A special case of this inequality is when $\lambda_i = \frac{1}{n}$ because then

$$\frac{1}{n}\sum_{i=1}^{n} f(x_i) - f\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \geq 0, \tag{2.14}$$

that is, the value of the function at the mean of the x_i is less or equal than the mean of the values of the function at each x_i.

Jensen's inequality can also be expressed in the following way: if f is convex on the range of a random variable X, then

$$f(\mathrm{E}[X]) \leq \mathrm{E}[f(X)], \tag{2.15}$$

where E denotes expectation (i.e., $\mathrm{E}[f(X)] = \sum_{x\in\mathcal{X}} p(x)f(x)$). Observe that if $f(x) = x^2$ (convex function), then $\mathrm{E}[X^2]-(\mathrm{E}[X])^2 \geq 0$. Thus, the variance is always positive.

2.2.3.2 Log-sum inequality

The log-sum inequality can be obtained from Jensen's inequality (Equation (2.13)). For non-negative numbers $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$, the log-sum inequality is expressed as

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} - \left(\sum_{i=1}^{n} a_i\right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \geq 0, \tag{2.16}$$

with equality if and only if $\frac{a_i}{b_i}$ is constant for all i. The conventions that $0\log 0 = 0$, $0\log\frac{0}{0} = 0$, and $a\log\frac{a}{0} = \infty$ if $a > 0$ are again adopted.

From this inequality, it can be proved that $H(X)$ is a concave function of p [Cover 1991].

From this inequality, the following properties can be proved [Cover 1991]:

- $D_{\mathrm{KL}}(p\|q)$ is convex in the pair (p, q).
- $H(X)$ is a concave function of p.
- If X and Y have the joint distribution $p(x, y) = p(x)p(y|x)$, then $I(X;Y)$ is a concave function of $p(x)$ for fixed $p(y|x)$ and a convex function of $p(y|x)$ for fixed $p(x)$.

2.2.3.3 Jensen-Shannon inequality

The Jensen-Shannon divergence, derived from the concavity of entropy, is used to measure the dissimilarity between two probability distributions and has the important feature that a different weight can be assigned to each probability distribution. The Jensen-Shannon (JS) divergence is defined by

$$JS(\pi_1, \pi_2, \ldots, \pi_n; p_1, p_2, \ldots, p_n) = H\left(\sum_{i=1}^{n} \pi_i p_i\right) - \sum_{i=1}^{n} \pi_i H(p_i), \tag{2.17}$$

where $p_1, p_2, \ldots, p_n$ are a set of probability distributions defined over the same alphabet with prior probabilities or weights $\pi_1, \pi_2, \ldots, \pi_n$, fulfilling $\sum_{i=1}^{n} \pi_i = 1$,

and $\sum_{i=1}^{n} \pi_i p_i$ is the probability distribution obtained from the weighted sum of the probability distributions $p_1, p_2, \ldots, p_n$.

From the concavity of entropy (Section 2.2.3.2), the Jensen-Shannon inequality [Burbea 1982] is obtained:

$$JS(\pi_1, \pi_2, \ldots, \pi_n; p_1, p_2, \ldots, p_n) \geq 0. \tag{2.18}$$

The JS-divergence measures how far the probabilities p_i are from their mixing distribution $\sum_{i=1}^{n} \pi_i p_i$, and equals zero if and only if all the p_i are equal. It is important to note that the JS-divergence is identical to the mutual information $I(X;Y)$ when $\pi_i = p(x_i)$ (i.e., $\{\pi_i\}$ corresponds to the marginal distribution $p(X)$), $p_i = p(Y|x_i)$ for all $x_i \in \mathcal{X}$ (i.e., p_i corresponds to the conditional distribution of Y given x_i), and $n = |\mathcal{X}|$ [Burbea 1982, Slonim 2000b].

2.2.3.4 Data processing inequality

The data processing inequality is expressed as follows. If $X \rightarrow Y \rightarrow Z$ is a Markov chain[4], then

$$I(X;Y) \geq I(X;Z). \tag{2.19}$$

This result proves that no processing of Y, deterministic or random, can increase the information that Y contains about X. In particular, if $Z = f(Y)$, then $X \rightarrow Y \rightarrow f(Y)$ and, consequently, $I(X;Y) \geq I(X;f(Y))$ [Cover 1991].

2.2.4 Entropy rate

Using the property $H(X_1,X_2) = H(X_1) + H(X_2|X_1)$ (Section 2.2.1) and the induction on n [Yeung 2008], it can be proved that the joint entropy of a collection of n random variables $X_1, \ldots, X_n$ is given by

$$H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_1, \ldots, X_{i-1}). \tag{2.20}$$

We now introduce the entropy rate that quantifies how the entropy of a sequence of n random variables increases with n. The entropy rate or entropy density h_x of a stochastic process[5] $\{X_i\}$ is defined by

$$h_x = \lim_{n \rightarrow \infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n) \tag{2.21}$$

when the limit exists.

[4] For random variables X, Y, and Z, $X \rightarrow Y \rightarrow Z$ forms a Markov chain if $p(x,y,z) = p(x)p(y|x)p(z|y)$. That is, the probability of the future state depends on the current state only and is independent of what happened before the current state.

[5] A stochastic process or a discrete-time information source $\{X_i\}$ is an indexed sequence of random variables characterized by the joint probability distribution $p(x_1, x_2, \ldots, x_n) = \Pr[(X_1, X_2, \ldots, X_n) = (x_1, x_2, \ldots, x_n)]$ with $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ for $n \geq 1$ [Cover 1991, Yeung 2008].

Entropy rate represents the average information content per symbol in a stochastic process. For a stationary stochastic process[6], the entropy rate exists and is equal to

$$h_x = \lim_{n\to\infty} h_x(n), \tag{2.22}$$

where $h_x(n) = H(X_1,\ldots,X_n) - H(X_1,\ldots,X_{n-1}) = H(X_n|X_{n-1},\ldots,X_1)$. Entropy rate can be seen as the uncertainty associated with a given symbol if all the preceding symbols are known. It can also be interpreted as the irreducible randomness in sequences produced by an information source [Feldman 1998].

An alternative notation, inspired by the work of Feldman and Crutchfield [Crutchfield 2003], is also used here to define the entropy rate. Given a chain $\ldots X_{-2}X_{-1}X_0X_1X_2\ldots$ of random variables X_i taking values in $\mathcal{X}$, a block of L consecutive random variables is denoted by $X^L = X_1\ldots X_L$. The probability that the particular L-block x^L occurs is denoted by $p(x^L)$. The joint entropy of length-L sequences or *L-block entropy* is now denoted by

$$H(X^L) = -\sum_{x^L\in\mathcal{X}^L} p(x^L)\log p(x^L), \tag{2.23}$$

where the sum runs over all possible L-blocks. Thus, the *entropy rate* can be rewritten as

$$h_x = \lim_{L\to\infty} \frac{H(X^L)}{L} = \lim_{L\to\infty} h_x(L), \tag{2.24}$$

where $h_x(L) = H(X_L|X_{L-1},X_{L-2},\ldots,X_1)$ is the entropy of a symbol conditioned on a block of $L-1$ adjacent symbols.

2.2.5 Entropy and coding

In this section, we review other interpretations of the Shannon entropy:

- As we have seen in Section 2.2.1, $-\log p(x)$ represents the information associated with the result x. The value $-\log p(x)$ can also be interpreted as the surprise associated with the outcome x. If $p(x)$ is small, the surprise is large; if $p(x)$ is large, the surprise is small. Thus, entropy (Equation (2.1)) can be seen as the expectation value of the surprise [Feldman 2002].

- A fundamental result of information theory is the Shannon source coding theorem, which deals with the encoding of information in order to store or transmit it efficiently. This theorem can be formulated in the following ways [Cover 1991, Feldman 2002]:

[6] A stochastic process $\{X_i\}$ is stationary if two subsets of the sequence, $\{X_1,X_2,\ldots,X_n\}$ and $\{X_{1+l},X_{2+l},\ldots,X_{n+l}\}$, have the same joint probability distribution for any $n,l \geq 1$: $\Pr[(X_1,\ldots,X_n) = (x_1,x_2,\ldots,x_n)] = \Pr[(X_{1+l},X_{2+l},\ldots,X_{n+l}) = (x_1,x_2,\ldots,x_n)]$. That is, the statistical properties of the process are invariant to a shift in time. At least, H_X exists for all stationary stochastic processes.

– Given a random variable X, $H(X)$ fulfills

$$H(X) \leq \overline{\ell} < H(X) + 1, \tag{2.25}$$

where $\overline{\ell}$ is the expected length of an optimal binary code for X. An example of an optimal binary code is the Huffman instantaneous coding[7].

– If we optimally encode n identically distributed random variables X with a binary code, the Shannon source coding theorem can be enunciated in the following way:

$$H(X) \leq \overline{\ell}_n < H(X) + \frac{1}{n}, \tag{2.26}$$

where $\overline{\ell}_n$ is the expected codeword length per unit symbol. Thus, by using large block lengths, we can achieve an expected codelength per symbol arbitrarily close to the entropy [Cover 1991].

– For a stationary stochastic process, we have

$$\frac{H(X_1, X_2, \ldots, X_n)}{n} \leq \overline{\ell}_n < \frac{H(X_1, X_2, \ldots, X_n)}{n} + 1 \tag{2.27}$$

and, from the definition of entropy rate H_X (Equation (2.21)),

$$\lim_{n \to \infty} \overline{\ell}_n \to H_X. \tag{2.28}$$

Thus, the entropy rate is the expected number of bits per symbol required to describe the stochastic process.

2.2.6 Information bottleneck method

The information bottleneck method, introduced by Tishby et al. [Tishby 1999], is a technique that extracts a compact representation of the variable X, denoted by $\widehat{X}$, with minimal loss of mutual information with respect to another variable Y (i.e., $\widehat{X}$ preserves as much information as possible about the control variable Y). Thus, given an information channel between X and Y, the information bottleneck method tries to find the optimal tradeoff between accuracy and compression of X when the bins of this variable are clustered.

Soft [Tishby 1999] and hard [Slonim 2000a] partitions of X can be adopted. In the first case, every $x \in \mathcal{X}$ can be assigned to a cluster $\hat{x} \in \widehat{\mathcal{X}}$ with some conditional probability $p(\hat{x}|x)$ (soft clustering). In the second case, every $x \in \mathcal{X}$ is assigned to only one cluster $\hat{x} \in \widehat{\mathcal{X}}$ (hard clustering).

In this thesis, we consider hard partitions and we focus our attention on the agglomerative information bottleneck method [Slonim 2000a]. Given a cluster $\hat{x}$

[7] A code is called a prefix or instantaneous code if no codeword is a prefix of any other codeword. Huffman coding uses a specific algorithm to obtain the representation for each symbol. The main characteristic of this code is that the most common symbols use shorter strings of bits than the ones used by the less common symbols.

defined by $\hat{x} = \{x_1, \ldots, x_l\}$, where $x_k \in \mathscr{X}$ for all $k \in \{1, \ldots, l\}$, and the probabilities $p(\hat{x})$ and $p(y|\hat{x})$ defined by

$$p(\hat{x}) = \sum_{k=1}^{l} p(x_k), \tag{2.29}$$

$$p(y|\hat{x}) = \frac{1}{p(\hat{x})} \sum_{k=1}^{l} p(x_k, y) \quad \forall y \in \mathscr{Y}, \tag{2.30}$$

the following properties are fulfilled:

- The decrease in the mutual information $I(X;Y)$ due to the merge of $x_1, \ldots, x_l$ is given by
$$\delta I_{\hat{x}} = p(\hat{x}) JS(\pi_1, \ldots, \pi_l; p_1, \ldots, p_l) \geq 0, \tag{2.31}$$
where the weights and probability distributions of the JS-divergence are given by $\pi_k = \frac{p(x_k)}{p(\hat{x})}$ and $p_k = p(Y|x_k)$ for all $k \in \{1, \ldots, l\}$, respectively. An optimal clustering algorithm should minimize $\delta I_{\hat{x}}$.
- An optimal merge of l components can be obtained by $l-1$ consecutive optimal merges of pairs of components.

2.2.7 Generalized entropies

Rényi [Rényi 1961] and Harvda and Charvát [Harvda 1967] introduced, respectively, two generalized definitions of entropy which includes the Shannon entropy as a particular case.

The *Rényi entropy* $H_\alpha^R(X)$ of a random variable X is defined by

$$H_\alpha^R(X) = \frac{1}{1-\alpha} \log \sum_{x \in \mathscr{X}} p(x)^\alpha, \tag{2.32}$$

where $\alpha > 0$ and $\alpha \neq 1$. When $\alpha \to 1$, $H_\alpha^R(X) = H(X)$. $H_\alpha^R(X)$ is a concave function of p if $\alpha \leq 1$, but neither concave nor convex if $\alpha > 1$.

Tsallis [Tsallis 1988] used the Harvda-Charvát entropy in order to generalize the Boltzmann entropy in statistical mechanics. The introduction of this entropy responds to the objective of generalizing the statistical mechanics to non-extensive systems[8]. In this thesis, we only use the Harvda-Charvát entropy.

The Harvda-Charvát-Tsallis entropy $H_\alpha(X)$ of a discrete random variable X is defined by

$$H_\alpha(X) = k \frac{1 - \sum_{x \in \mathscr{X}} p(x)^\alpha}{\alpha - 1}, \tag{2.33}$$

[8] An extensive system fulfills that quantities like energy and entropy are proportional to the system size. Similarly to Shannon entropy, a fundamental property of the Boltzmann entropy is its additivity. That is, if we consider a system composed by two probabilistically independent subsystems X and Y (i.e., $p(x,y) = p(x)p(y)$), then H(X,Y) = H(X) + H(Y). This property ensures the extensivity of the entropy but strongly correlated systems present non-extensive properties that require another type of entropy fulfilling non-additivity.

where k is a positive constant (by default $k = 1$) and $\alpha \in \mathbb{R}\backslash\{1\}$ is called entropic index. This entropy recovers the Shannon entropy (calculated with natural logarithms) when $\alpha \to 1$ and fulfils the properties of non-negativity and concavity (for $\alpha > 0$). In this thesis, the Harvda-Charvát-Tsallis entropy is also called Tsallis entropy.

If X and Y are independent, then the Harvda-Charvát-Tsallis entropy fulfills the non-additivity property:

$$H_\alpha(X,Y) = H_\alpha(X) + H_\alpha(Y) + (1-\alpha)H_\alpha(X)H_\alpha(Y), \tag{2.34}$$

hence, superextensivity, extensivity or subextensivity occurs when $\alpha < 1$, $\alpha = 1$ or $\alpha > 1$, respectively [Tsallis 2002].

The *Tsallis conditional entropy* $H_\alpha(Y|X)$ is defined by

$$\begin{aligned} H_\alpha(Y|X) &= \sum_{x\in\mathscr{X}} p(x)^\alpha H_\alpha(Y|x) \\ &= \sum_{x\in\mathscr{X}} p(x)^\alpha \frac{1-\sum_{y\in\mathscr{Y}} p(y|x)^\alpha}{\alpha-1}, \end{aligned} \tag{2.35}$$

where $H_\alpha(Y|x)$ is the Tsallis entropy of Y known x.

From Equation (2.10), 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 $p(x)p(y)$. Tsallis [Tsallis 1998] generalized the Kullback–Leibler distance in the following form:

$$KL_\alpha(p,q) = \frac{1}{1-\alpha}\left(1 - \sum_{x\in\mathscr{X}} \frac{p(x)^\alpha}{q(x)^{\alpha-1}}\right). \tag{2.36}$$

Thus, from Equations (2.10) and (2.36), Tsallis mutual information can be defined [Taneja 1988, Tsallis 1998] as

$$\begin{aligned} I_\alpha(X;Y) &= KL_\alpha\big(p(x,y), p(x)p(y)\big) \\ &= \frac{1}{1-\alpha}\left(1 - \sum_{x\in\mathscr{X}}\sum_{y\in\mathscr{Y}} \frac{p(x,y)^\alpha}{p(x)^{\alpha-1}p(y)^{\alpha-1}}\right). \end{aligned} \tag{2.37}$$

In the context of Tsallis entropy, the normalized mutual information NMI (Equation (2.11)) can be generalized as

$$NMI_\alpha(X;Y) = \frac{I_\alpha(X;Y)}{H_\alpha(X,Y)}, \tag{2.38}$$

where $H_\alpha(X,Y)$ is the Harvda-Charvát-Tsallis entropy defined in Section 2.2.7. Although $NMI_\alpha(X;Y)$ is a normalized measure for $\alpha \to 1$, this is not true for other α values as NMI_α can take values greater than 1. This measure is always positive

and symmetric.

2.3 Document classification

The automated classification of scanned documents has become an essential task in document image processing and it is used in many applications due to the advances in communication and information technology:

- Office automation. Document classification allows the automatic distribution or archiving of documents. For example, after classification of business letters according to the sender and message type (such as order, offer, or inquiry), the letters are sent to the appropriate departments for processing.
- Digital libraries. Document classification improves the indexing efficiency in digital library construction. For example, the classification of documents into table of contents page or title page can narrow the set of pages from which to extract specific meta-data, such as the title or table of contents of a book.
- Image retrieval. Document classification plays an important role in document image retrieval. For example, consider a document image database containing a large heterogeneous collection of document images. Users have many retrieval demands, such as retrieval of papers from one specific journal, or retrieval of document pages containing tables or graphics. Classification of documents based on visual similarity helps to limit the search and improves retrieval efficiency and accuracy.
- Other document image analysis applications. Document classification facilitates higher-lever document analysis. Due to the complexity of document understanding, most high-level document analysis systems rely on domain-dependent knowledge to obtain high accuracy. Many available information extraction systems are specially designed for a specific type of document, such as forms processing or postal address precessing, to achieve high speed and performance. To process a broad range of documents, it is necessary to classify the documents first, so that a suitable document analysis system for each specific document type can be adopted.

Electronic documents have many advantages over paper documents due to their compact storage, easy maintenance, and efficient retrieval. The huge amount of electronic documents creates the need of automatic classification based on its content or visual similarity with minimal human intervention. According to Peng et al. [Peng 2003], 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 [Lopresti 2000] or different types of line segments [Tseng 1997], and on extracting global layout information, such as the use of a spatial layout representation [Hu 1999b] or geometric features [Shin

2001]. Chen and Blostein [Chen 2007] present an excellent survey on document classification.

Most published document image classification systems rely on either OCR [Trier 1996, Lopresti 2000], so that the documents can be categorized depending on the textual content, or on some form of layout analysis to produce an "appearance signature" that can be compared to category-prototypes [Hu 1999b, Shin 2001]. Classifying documents by appearance is a useful part of document indexing, retrieval, organization, sorting, and workflow routing. This is especially true if OCR is inaccurate, expensive, or relies on prior knowledge about document type, or if the text content is too variable or insufficient for classification [Gupta 2007].

Due to the wide variety of classification problems that can be raised, there is a huge diversity of document classifiers with significant differences between them. This fact involves the need to clearly specify the problem before starting the design of a classifier. Some of the most important aspects to be taken into account are the level of detail, the document features, and the classification techniques. Next, we briefly summarize these aspects:

- *Level of detail*. Bagdanov and Worring [Bagdanov 2001] defined two levels of detail: coarse-grained and fine-grained. While the coarse-grained classification is used to classify documents with different types [Eglin 2003, Maderlechner 1997, Shin 2001], such as invoices versus journal pages, the fine-grained classification is used to classify documents with the same typology [Appiani 2001, Bagdanov 2001, Baumann 1997, Diligenti 2003, Sako 2003], such as invoices from different suppliers.

- *Document features*. In order to classify electronic documents, it is necessary to extract a set of document features. These can be grouped into four groups: image, physical layout, logical layout, and textual features. *Image features*, that are directly extracted from the image, can be divided into two subgroups: global and local image features. Global features [Bagdanov 2001, Fan 2010, Fränti 2000] are computed on the whole image (e.g., the density of black pixels), whereas local features [Bagdanov 2001, Diligenti 2003, Meshesha 2008, Tseng 1997] are computed on regions of the image (e.g., the number of horizontal lines in a segmented region). These features are usually limited to a coarse-grained level classification. *Physical* [Esposito 2000, Baldi 2003, Bagdanov 2003] and *logical* [Rangoni 2011, Duygulu 2002, Liang 2002] *layout features* are used to classify documents of the same type with structural variations [Nagy 2000, Mao 2003, Haralick 1994]. The use of these structural features allows the classifier to perform a fine-grained level classification. Finally, *textual features* can be extracted directly from the images [Shin 2001, Spitz 1999] or computed from the OCR output [Baumann 1997, Maderlechner 1997]. These features allow that the classifier assigns one or multiple labels to a document based on its content.

- *Classification techniques*. There is a huge variety of classification techniques.

The most used ones, can be grouped into four basic categories: statistical, structural, knowledge-based, and template matching techniques. *Statistical techniques* [Duda 2001, Jain 2000] are relatively mature and include the use of tools such as nearest neighbor [Héroux 1998, Alippi 2005], decision trees [Shin 2001], neural networks [Rangoni 2011], or hidden Markov models [Hu 1999a]. These techniques do not usually capture the document structure and therefore they are not suitable for fine-grained classification. *Structural techniques* are computationally complex and include the use of tools such as decision trees [Shin 2001], hidden tree Markov models [Diligenti 2003], or graph matching techniques [Bunke 2000, Liang 2002]. *Knowledge-based techniques* require a significant effort to acquire, maintain, and update the knowledge base [Lam 1993], although there are systems that learn rules automatically from labeled training samples [Esposito 2000, Wenzel 2001]. Finally, *template matching techniques* [Byun 2000, Kochi 1999] are used to match a document image with a template document.These techniques are most commonly applied in cases where document images have a fixed geometric configuration and are recommended for coarse-grained classification.

It is interesting to note that the above techniques are usually combined to improve the classification performance [Ho 2001, Héroux 1998].

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.

In our work, 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 [Hajnal 2001]. Although our analysis can be extended to a wide variety of document types, we focus our attention on invoice classification.

2.4 Video processing

Video processing constitutes one of the major areas of multimedia research due to the exponential growth of digital video generation that has taken place in the last decades. Shot boundary detection and keyframe selection are fundamental tasks for video processing applications such as content-based video retrieval and video summarization applications [Money 2008, Peng 2010]. In video processing, a shot may be defined as a sequence of frames that was continuously captured from a single camera at a time, and it can encompass pans, tilts, or zooms. Usually, a shot

is a group of frames that have consistent visual characteristics such as color, texture, and motion. A video sequence normally contains a large number of shots, which are connected with each other through different video editing methods. A shot boundary is the gap between two shots and its identification enables us to index the video sequence, facilitating the fast browsing and retrieval of shots of interest to the user [Cotsaces 2006, Grana 2007, Urhan 2006, Yuan 2007]. The main difficulties of automatic shot boundary detection are related to the object motion and illumination variations, which can be easily confused with a shot boundary.

The transitions between shots can be classified into two basic groups: hard cuts and gradual transitions. A hard cut is an abrupt shot change that occurs between two continuous frames, i.e., when the last frame in one shot is followed by the first frame in the next shot. A comparison of existing cut detector methods is presented by Lienhart [Lienhart 1999] and Browne et al. [Browne 1999]. Gradual transitions occur over multiple frames and the most typical are fade-in, fade-out, dissolve, and wipe, but many other types of gradual transition are possible [Lienhart 2001, Hanjalic 2002, Yuan 2007].

Shot boundary detection techniques can be divided into two basic categories depending on the use of compressed video data or not. The methods that work in the compressed domain usually use motion differences encoded in the MPEG standard and are faster than the ones in the uncompressed domain since we assume that the video is given in a compressed format and, therefore, the decompression step is unnecessary [Meng 1995, Lelescu 2003]. However, these algorithms often show a lower performance due to the limited features that can be extracted from the compressed video [Lee 2006]. Thus, the proposed methods in the uncompressed domain are usually more accurate but also more computationally expensive than the ones in the compressed domain due to the decompression step [Hanjalic 2002]. The most common methods that work in the uncompressed domain are based on pixel differences [Nagasaka 1992, Lienhart 1999], statistical differences [Hanjalic 2002, Yoo 2006], histogram comparison [Lienhart 1999, Gargi 2000], edge differences [Zabih 1995], motion vector [Dhawale 2008, Tardini 2005, Hesseler 2006], and information-theoretic measures [Butz 2001, Cernekova 2006, Xu 2010].

Keyframes provide a suitable video summarization and a framework for video indexing, browsing, and retrieval [Günsel 1998, Nagasaka 1992, Wolf 1996]. The use of keyframes greatly reduces the amount of data required in video indexing and provides a framework to deal with the video content. The simplest proposed methods choose only one frame for each shot (usually the first one), regardless of the complexity of visual content. The more sophisticated approaches take into account visual content, motion analysis, and shot activity [Zhuang 1998]. These approaches either do not effectively capture the major visual content or are computationally expensive. Ciocca and Schettini [Ciocca 2006] propose an approach for keyframe selection by analyzing the differences between two consecutive frames using different frame descriptors. Peng and Xiao-Lin [Peng 2010] introduce an adaptive keyframe extraction method based on the visual attention model.

CHAPTER 3

Image-based similarity measures for document classification

3.1 Introduction

In this chapter, we analyze a set of global measures to evaluate the similarity between two invoices instead of extracting specific pieces of information or analyzing the invoice layout. The use of global similarity measures is in general very demanding to be applied to large databases, however we propose to downscale the scanned images and, thus, to compute the similarity on low resolution images. This makes our approach completely feasible, achieving more robustness and accuracy.

We investigate three types of measures, based respectively on intensity differences, mutual information, and normalized compression distance. While the first group is based on the intensity difference between the corresponding pixels, the second group uses the joint probability distribution of intensities to take into account the correlation between the structures of the document. In the third group, the normalized compression distance utilizes different compressors to compute the dissimilarity between document images. A number of experiments analyze the performance of some of the proposed measures applied to two testing invoice databases composed by colored and black-and-white images, respectively, and to a real-world invoice database.

The content of this chapter is presented in "*Image-based Similarity Measures for Invoice Classification*", Marius Vila, Anton Bardera, Miquel Feixas, Mateu Sbert. Submitted.

3.2 Previous work

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 [Hamza 2008]. Chen and Blostein [Chen 2007] 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 [Trier 1996]. 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 [Appiani 2001]. Several strategies have been proposed to achieve an accurate document classification based on the layout analysis and classification [Hu 1999b, Appiani 2001, Shin 2001, Peng 2003, Shin 2006, Gupta 2007].

In the literature of document image classification, different measures of similarity have been used. Appiani *et al*. [Appiani 2001] design a criterion to compare the structural similarity between trees that represent the structure of a document. Shin and Doermann [Shin 2006] 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*. [Behera 2005] 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.

In this work, we focus our attention on invoice classification. Invoices are commercial document issued by a seller, containing details about the seller, the buyer, products, quantities, prices, etc., and usually a logo and tables. Several approaches for invoice classification have been presented by Alippi et al. [Alippi 2005], Silva et al. [Silva 2006], and Hamza et al. [Hamza 2008]. Hamza *et al*. [Hamza 2008] 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 frequently used in heterogeneous document flows and to extract different information from documents, such as tables [Silva 2006], graphical features such as logos and trademarks [Alippi 2005], or the general layout [Appiani 2001]. On the contrary, model-based systems are used in homogeneous document flows, where similar documents arrive generally one after the other [Arai 1997, Cesarini 1998, Tang 1997, Duygulu 2002].

In this chapter, we focus our attention on capturing visual similarity between different invoice images using global measures that do not require the analysis of the invoice layout.

3.3 Similarity measures

In this section, we present three types of similarity measures based on intensity differences, the joint probability distribution of intensities, and image compressibility, respectively. The performance of these measures to evaluate the invoice similarity will be analyzed in the next section.

The following similarity measures are computed in the overlapping domain $\Omega_{A,B}$ of images A and B, which is defined as the intersection area of both images when these are aligned at the coordinate origin. Note that, since the images have been rescaled to a fixed height, $\Omega_{A,B}$ is given by the image height times the minimum width of the images A and B.

3.3.1 Intensity-based measures

We present here three intensity-based measures to compute the similarity between two images [Studholme 1997, Hajnal 2001]. Specifically, the similarity is calculated from the local difference between the intensities corresponding to the pairs of matching pixels. We analyze the following three measures:

- *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

$$SSD = \frac{1}{N} \sum_{i \in \Omega_{A,B}} |\, A(i) - B(i) \,|^2, \tag{3.1}$$

  where $A(i)$ and $B(i)$ represent the intensity at a pixel i of the images A and B, respectively. Note that the images can have a different size and, therefore, the similarity measures are only computed on the overlap area $\Omega_{A,B}$ between both images. It is assumed that the image values are calibrated to the same scale. It can be shown that SSD is the optimal measure when two images only differ by Gaussian noise [Viola 1995]. A drawback of this measure is that it is very sensitive to a small number of pairs of pixels that have very large intensity differences.

- *Sum of absolute differences (SAD)*: This measure is defined as

$$SAD = \frac{1}{N} \sum_{i \in \Omega_{A,B}} |\, A(i) - B(i) \,|\,. \tag{3.2}$$

  Since the differences are not squared, the negative effects of SSD on an small number of large intensity differences are reduced by using SAD.

- *Correlation coefficient (CC)*: This measure is defined as

$$CC = \frac{\sum_{i \in \Omega_{A,B}} (A(i) - \bar{A})(B(i) - \bar{B})}{[\sum_{i \in \Omega_{A,B}} (A(i) - \bar{A})^2 \sum_{i \in \Omega_{A,B}} (B(i) - \bar{B})^2]^{\frac{1}{2}}}, \tag{3.3}$$

  where $\bar{A}$ and $\bar{B}$ are, respectively, the mean intensity values in images A and B in the overlap domain $\Omega_{A,B}$. While SSD makes the implicit assumption that the images differ only by Gaussian noise, CC assumes that there is a linear relationship between the intensity values in the images [Hill 2001].

3.3.2 Information-theoretic measures

In this section, we present two similarity measures, mutual information and normalized mutual information, which are based on the joint probability distribution of two matching images. This fact confers to these measures a better behavior than the intensity-based measures presented in the previous section. Mutual information and normalized mutual information have been widely used in numerous papers in the multimodal registration field [Hajnal 2001]. These measures are based on the probabilities of the intensities instead of the intensities themselves. To define these measures, an *information channel* $X \rightarrow Y$ is created, where X and Y stand for the two images A and B, respectively. In this channel, their marginal probability distributions, $p(x)$ and $p(y)$, and the joint probability distribution, $p(x,y)$, are obtained by simple normalization of the marginal and joint intensity histograms of the overlapping area of both images. To define these measures, we remember here the definition of entropy for one and two random variables. For more details, see Section 2.2.1 and the books by Cover and Thomas [Cover 1991], and Yeung [Yeung 2008].

The most basic information measure is the *Shannon entropy* $H(X)$, where the random variable X represent the intensity bins of an image A. The Shannon entropy quantifies the average uncertainty of a random variable X with probability distribution $\{p(x)\}$, and is defined by

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x), \tag{3.4}$$

where $\mathcal{X}$ is the alphabet of X. In this case, the alphabet $\mathcal{X}$ is given by the set of intensity bins.

The *joint entropy* $H(X,Y)$ of a pair of discrete random variables X and Y with joint probability distribution $\{p(x,y)\}$ is defined by

$$H(X,Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log p(x,y), \tag{3.5}$$

where $\mathcal{X}$ and $\mathcal{Y}$ are, respectively, the alphabets of X and Y, given by the intensity bins of images A and B, respectively. $H(X,Y)$ measures the average uncertainty of the pair (X,Y).

Mutual information and normalized mutual information are defined as follows:

- *Mutual information (I)*: The mutual information between two random variables X and Y is defined as

$$\begin{aligned} I(X;Y) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\ &= H(X) + H(Y) - H(X,Y) \end{aligned} \tag{3.6}$$

  and represents the shared information between X and Y, where X and Y

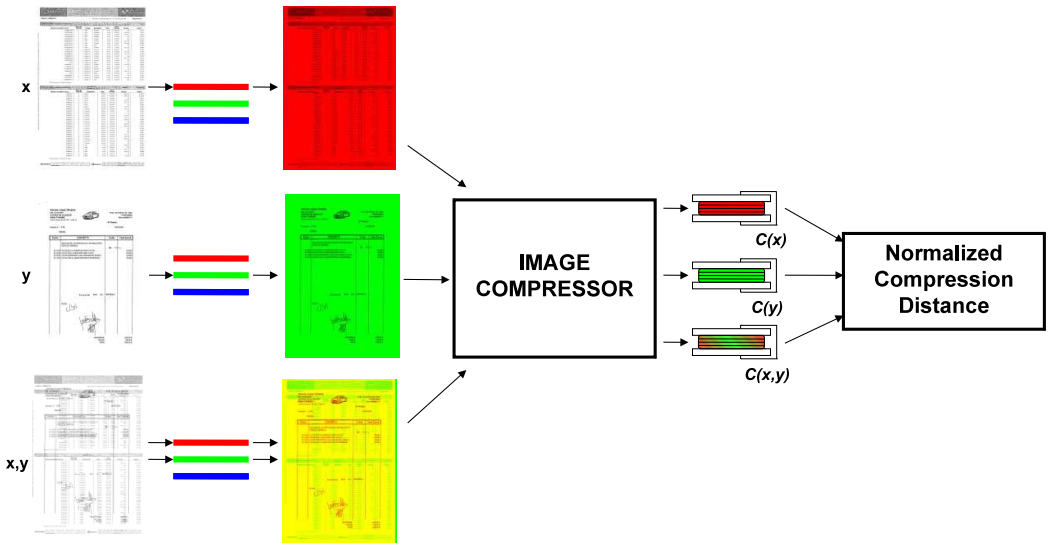


Figure 3.1: Computation of the normalized compression distance using an image compressor.

represent the intensity bins of images A and B, respectively. $I(X;Y)$ was introduced by Viola [Viola 1995] and Maes et al. [Maes 1997] as a similarity measure for image registration.

- *Normalized mutual information (NMI)*: The normalized mutual information is defined as

$$NMI(X;Y) = \frac{I(X;Y)}{H(X,Y)}. \tag{3.7}$$

The joint entropy $H(X,Y)$ is an upper bound of mutual information and, hence, normalizes the measure between $[0,1]$. In the context of image registration, Studholme et al. [Studholme 1997] showed that this measure is more robust than the mutual information $I(X;Y)$, due to its greater independence of the overlap area. Another theoretical justification of its good behavior is that the normalized mutual information NMI is a true distance.

It is important to remark that, while the intensity-based measures seen in the previous section are very sensitive to changes in the intensity values, the mutual information-based measures are not directly based on these values but on their co-occurrences. This fact helps to capture with greater precision the structural similarity between two images.

3.3.3 Compression-based measures

In this section, we present a dissimilarity measure between two images based on the Kolmogorov complexity. The *Kolmogorov complexity* $K(x)$ of a string x is the length of the shortest program to compute x on an appropriate universal computer.

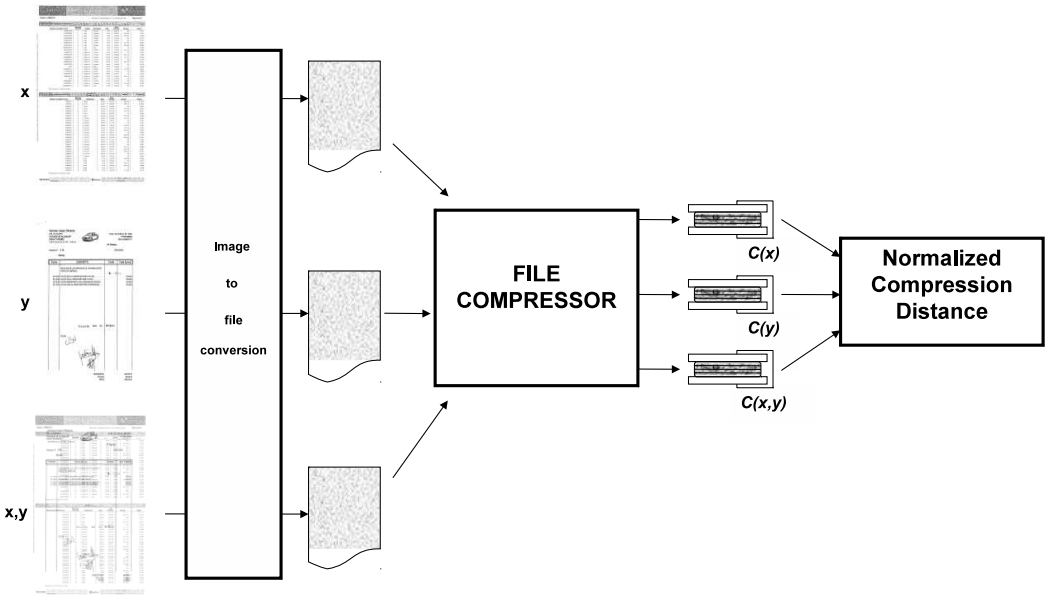


Figure 3.2: Computation of the normalized compression distance using a file compressor.

Essentially, the Kolmogorov complexity of a string is the length of the ultimate compressed version of the string. Moreover, the joint complexity $K(x,y)$ represents the length of the shortest program for the pair (x,y) [Li 2004, Li 2008].

Li et al. [Li 2004] presented the *normalized information distance* (NID), called also *the similarity metric*, as a universal metric which quantifies the distance between two strings x and y as the length of the shortest program that computes x from y and y from x in a normalized way. NID is defined by

$$NID(x,y) = \frac{K(x,y) - \min\{K(x),K(y)\}}{\max\{K(x),K(y)\}} \tag{3.8}$$

and takes values in $[0,1]$. This metric is universal in the sense that if two strings are similar according to the particular feature described by a particular normalized admissible distance (not necessarily metric), then they are also similar in the sense of the normalized information metric [Cilibrasi 2005].

The Kolmogorov complexity K is a non-computable measure in the Turing sense [Li 2008] and, therefore, for real-world applications, we will need an approximation of it. An upper bound of K is the length $C(x)$ (or $C(y)$) of the compressed string x (or y) generated by a compression algorithm. The better the compression algorithm, the better the approximation to K [Li 2008]. Then, a feasible version of the normalized information distance (Equation (3.8)), called the

normalized compression distance (NCD), is defined as

$$NCD(x,y) = \frac{C(x,y) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}, \tag{3.9}$$

where $C(x)$ (or $C(y)$) represents the length of the compressed string x (or y) and $C(x,y)$ the length of the compressed pair (x,y) [Cilibrasi 2005]. Thus, NCD is computed from the lengths of compressed data files and, therefore, NCD approximates NID by using standard real-world compressors. Bardera et al. [Bardera 2010] have studied the performance of this measure in the image registration field.

In our experiments, we use two different strategies to compress the images:

- *Image compressors*: In this approach, three standard real-world image compressors (JPEG, JPEG2000, and PNG) are used to compute NCD between two images. For each image, the values $C(x)$ and $C(y)$ can be easily computed by compressing the original images and taking the size of the resulting file. The problem arises with the computation of $C(x,y)$, since compressors are designed to deal with a single image. To overcome this limitation, we propose to use the three channels of the RGB representation. First, both images are converted to grayscale and are fused in a single image using the red channel for one image, the green channel for the other, and assigning null values to the blue one. In this way, we have a single image that can be compressed with the same compressor that has been used with the original images. This method is represented in Figure 3.1. For more details, see Bardera et al. [Bardera 2010].

- *File compressors*: In this approach, all the intensity values of the image are written in a binary file and compressed with a standard file compressor (GZIP and BZIP2). In order to encode the image in the file, each intensity value is represented as a byte. The size of the compressed files is given by $C(x)$ and $C(y)$, respectively. To compute $C(x,y)$, a binary file composed by the intensity values of both images is generated and compressed. This file has been obtained by taking alternately the intensity values of the pixels of both superimposed images. Figure 3.2 shows the different steps of this approach. For more details, see Bardera et al. [Bardera 2010].

In theory, NID takes into account all the characteristics of the images to obtain the dissimilarity between them. However, the approximation of that measure by NCD introduces a certain degree of inaccuracy due to the limited resources used by a compressor and makes this measure more inaccurate than it could be expected.

3.4 Methodology

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, as textual content, can be designed depending on these features [Appiani 2001].

As we have previously mentioned, we focus our attention on the classification of invoices. In our framework, each image class represents a set of different invoices of the same provider. As we can see in Figure 3.3, different invoices of the same class usually share a similar document layout. However, the document layouts can be very different from one class to another and they can differ, for instance, in font type, font size, background color, and the presence of logos, tables, and images (see Figure 3.3). This variety of document images makes it difficult to define a general scheme for OCR or layout analysis that produces accurate results.

A simple way to define a class consists in taking one or more representative images. Then, we can create a database with the representative images and every new entry in the classifier 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 3.4. As it can be seen, two different groups of images are defined. The first group, formed by the *reference images* (or template images), is constituted by a set of invoice images where each image belongs to a class. This group of images forms the invoice database. Note that one class can contain more than one image. The second group, composed by the *input images*, is constituted by a set of invoice images that is used as the input to the classifier with the aim of finding their corresponding class within the database of the reference images. When an input image does not correspond to any class, this image is considered as the reference image of a new class and extends the size of the reference image database.

Our classification framework is based on global image similarity measures between the images to be classified and the class templates, and not on extracting image features. It can be considered that our method is basically a template matching technique and that, due to the fact that each class represents a set of invoices of the same provider, it applies a fine-grained classification.

As all the invoice images are digitized using a scanner we can have both translation and skew errors. These errors should be corrected to obtain a more reliable evaluation of the image similarity. In our experiments, we assume that the used scanning protocol produces quasi-aligned images. The skew error is also corrected using the method presented by Gatos et al. [Gatos 1997]. Moreover, the images are downscaled to analyze the behavior of the classification process for different image resolutions. Note that this downscaling process contributes to minimize the small alignment errors. After these preprocessing steps, the similarity

Figure 3.3: Two document class samples.

between the input image and all the reference images is computed, and the input image is assigned to the class of the reference image for which the similarity value is maximum and greater than a given threshold. If this maximum value is lower than the threshold, we consider that the input invoice does not correspond to any existing class, and the new class is added to the database.

The main goal of this chapter is to evaluate the performance of different image similarity measures in the invoice classification process using different image resolutions.

3.5 Results

This section is split into two parts. In the first part, we present the experiments carried out with two testing databases (a black-and-white invoice database and a color invoice database) with the aim of studying the behaviour of the presented measures (see Section 3.3) at different image resolutions. In the second part, we analyze the behaviour of these measures with a real-world database. The proposed measures have been implemented in Visual C++ .NET and their performance has been measured on a system with an Intel Core 2 Duo CPU.

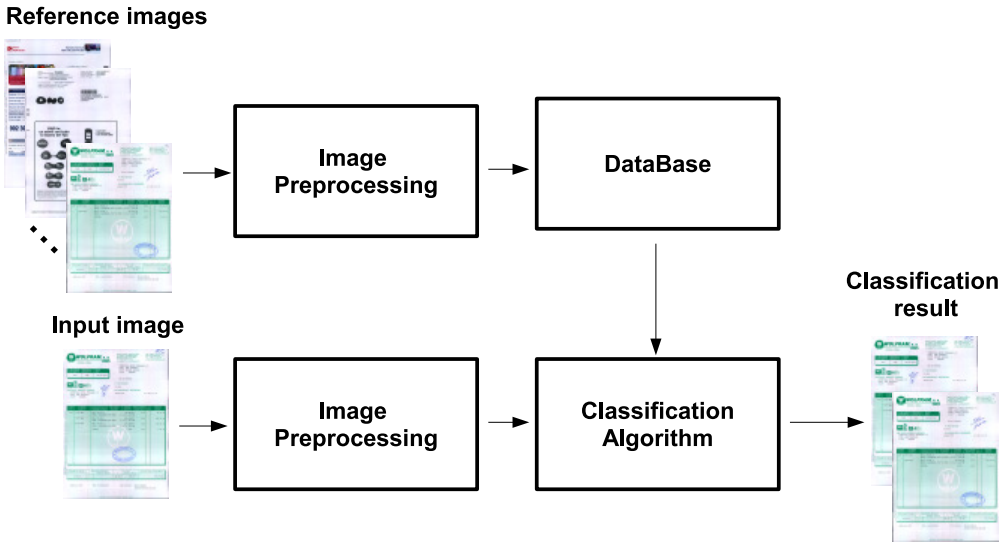


Figure 3.4: Invoice classification pipeline.

3.5.1 Experiments with two testing databases

In our first experiment we have used two different testing databases: a black-and-white invoice database and a color invoice database. First, we study the behaviour of the measures presented in Section 3.3 at different image resolutions. Second, we analyze how the image similarity-based classification is sensitive to the choice of the reference image of the class.

While the invoices from the first database only have two intensity values (black and white), the invoices from the second database use 24-bits per pixel (8-bits for each RGB color channel) and, usually, present a more complex layout, including pictures, logos, and highlighted areas. The black-and-white database is composed by 204 reference invoices and 154 input invoices, and the color database is composed by 51 reference invoices and 95 input invoices. In this evaluation experiment, for the sake of simplicity, each class is represented by a single invoice and a input invoice always correspond to an existing class, and thus will be assigned to the most similar reference image. In spite that invoice images of the same class come from the same scanning process and have a similar document layout, they might present a high variability due to changes in the textual content, different number of items, different figures, or stamp position (see two examples in Figure 3.3). Due to the design of this experiment, there are only two possible classification results: true positive (correct classification) or false positive (incorrect classification).

In our experiments, we assume that the images to be compared are fairly well aligned. With the aim of correcting the skew error introduced during the scanning process, all the reference and input invoices have been preprocessed using the method presented by Gatos et al. [Gatos 1997]. Although the skew error is corrected, they still present small translation errors among them. It is important to remark

(a) 800 (b) 400 (c) 200

(d) 100 (e) 50 (f) 25

Figure 3.5: An invoice sample with resolutions from (a) 800 to (f) 25 pixels. Resolution is specified by the number of pixels of image height.

that the invoices of the color database are not as well aligned as the invoices of the black-and-white database because they have been digitized using a less accurate scanning protocol.

To study the behavior of the similarity measures with respect to the image resolution, all images have been scaled to several sizes. This allows us to evaluate an approximated optimal resolution for each one of the proposed measures. Initially, the invoice images are acquired at a resolution of around 2500 pixels width × 3500 pixels height, however the height of the images is reduced to 800, 400, 200, 100, 50, and 25 pixels, conveniently adjusting the image width to keep the aspect ratio of the images. From now on, the resolution values of the images are only specified by the image height. It is important to note that the black-and-white images are transformed into grayscale images when they are downsampled to lower resolutions. That is, the intensity values of the low resolution pixels are produced from an average of the original black-and-white pixels. An invoice example with the resolutions used in this study can be seen in Figure 3.5. In these experiments, the information-theoretic measures, I and NMI, have been computed using only 8 intensity bins. In spite of the reduction of the number of pixels due to the image downscaling, the bin reduction avoids that the joint intensity histogram becomes too sparse.

As the main objective of our experiments is to calculate the degree of similarity between each input invoice and all reference invoices, an ordered list of reference invoices, called *similarity list*, is obtained from the similarity (from the highest to the lowest) between both the input and the reference invoices. Thus, it is interpreted that the first reference invoice of the list corresponds to the class assigned to the input invoice.

Next, we analyze the results obtained using both black-and-white and color invoice databases. For each database, each measure, and each different resolution, we obtain two types of results: accuracy, and classification error. The *accuracy* is given by the number of correctly classified input invoices over the total number of inputs. We consider that an invoice is correctly classified when the corresponding reference invoice occupies the first place in the similarity list.

We also define the *classification error* as the class position mean minus 1 of the misclassified invoices, where, given an input invoice, the class position is determined by the position of the corresponding reference invoice in the similarity list. If the reference invoice is chosen properly, this is located at position 1 of the list and, thus, the class position is 1. The classification error ranges from 0 to the number of classes minus 1.

Tables 3.1 and 3.2 present the results obtained by the proposed measures applied to the black-and-white and color databases, respectively. These tables show two values for each resolution and each measure, where the first represents the accuracy and the second represents the classification error. As it can be seen, most of the measures perform best at resolutions of 200, 100, and 50 pixels. In part, it can be considered that the reduction of the number of pixels allows us, on the one hand, to preserve the most relevant information and structure, and, on the other hand, to reduce noise. However, if we reduce too much the resolution, we lose significant information and, therefore, the accuracy decreases.

We now focus our attention on the black-and-white database (see Table 3.1). From the accuracy values, we can see that the best performance corresponds to the information-theoretic measures. Observe that NMI performs slightly better than I. This is due to the fact that the latter measure is very sensitive to the entropy of the reference image while NMI is normalized by the joint entropy. With respect to the intensity-based measures, CC performs better than both SSD and SAD, although the behavior of SAD and SSD is in some cases better at a resolution of 25 pixels. The behavior of the measures based on the compressors PNG, GZIP, and BZIP2 is worse than I, NMI, and CC. As we could expect, PNG and GZIP compressors have a similar behavior since they are based on the same compression algorithm, called Deflate. The better behavior of PNG against JPEG at a low resolution is due to the higher compression capacity of PNG when the images contain text, lines, and, in general, sharp transitions and large areas of solid color. Thus, NCD based on PNG is a better approximation of NID than NCD based on JPEG. In the JPEG case, the error decreases with the increase of resolution, achieving better results than PNG at the highest resolutions. This behavior could be explained by the fact that we use a lossy JPEG compression which suppresses redundant information at high resolution and

	25		50		100	
Measure	**A**	**E**	**A**	**E**	**A**	**E**
SSD	94.16	72.33	*95.45*	84.71	92.21	60.75
SAD	***96.10***	46.33	95.45	40.57	92.21	42.25
CC	94.16	7.78	97.40	3.50	97.40	2.25
I	92.86	49.00	**98.05**	2.67	***99.35***	2.00
NMI	94.16	55.44	98.05	5.00	*99.35*	3.00
PNG	85.06	25.52	96.10	26.00	*96.75*	24.60
JPEG	7.14	57.50	29.87	32.03	82.47	17.96
JPEG2000	*68.18*	21.14	60.39	18.79	64.94	20.33
GZIP	85.71	28.14	92.21	24.00	94.81	49.50
BZIP2	93.51	36.80	*95.45*	20.00	95.45	38.86

	200		400		800	
Measure	**A**	**E**	**A**	**E**	**A**	**E**
SSD	82.47	43.70	73.38	39.73	61.69	46.51
SAD	79.22	37.63	69.48	42.89	55.84	48.35
CC	*97.40*	2.00	96.75	4.40	96.75	3.60
I	**98.05**	2.00	96.10	2.50	96.10	2.83
NMI	**98.05**	2.00	96.75	3.60	96.10	3.00
PNG	95.45	25.71	95.45	36.00	92.86	69.00
JPEG	95.45	22.86	**97.40**	12.00	***98.05***	2.33
JPEG2000	65.58	21.21	64.94	25.76	59.09	34.38
GZIP	*98.05*	21.67	96.75	13.60	93.51	31.50
BZIP2	94.81	51.25	88.31	53.56	64.29	64.16

Table 3.1: For the black-and-white invoice database, the accuracy (A) and the classification error (E) for different image heights (25, 50, 100, 200, 400, and 800 pixels). Bold and italic numbers indicate, respectively, the best measure for each resolution and the best resolution for each measure.

	25		50		100	
Measure	**A**	**E**	**A**	**E**	**A**	**E**
SSD	*77.89*	26.67	76.84	23.91	70.53	20.25
SAD	77.89	16.29	*81.05*	16.22	76.84	14.41
CC	86.32	6.38	*89.47*	8.00	88.42	6.73
I	91.58	8.63	96.84	3.33	*98.95*	3.00
NMI	**92.63**	4.71	**97.89**	2.00	***100.00***	0.00
PNG	57.89	13.50	89.47	5.50	*94.74*	4.80
JPEG	7.37	13.49	10.53	11.80	58.95	7.31
JPEG2000	34.73	9.84	32.63	10.03	*35.79*	10.56
GZIP	73.68	12.08	90.53	5.00	*93.68*	7.67
BZIP2	84.21	6.53	*94.74*	3.40	92.63	5.00

	200		400		800	
Measure	**A**	**E**	**A**	**E**	**A**	**E**
SSD	72.63	21.54	66.32	19.13	57.89	17.25
SAD	71.58	13.26	65.26	12.58	58.95	12.87
CC	87.37	4.25	88.42	3.82	88.42	3.55
I	*98.95*	3.00	93.68	3.17	86.32	3.23
NMI	***100.00***	0.00	**94.74**	2.40	**91.58**	3.00
PNG	88.42	5.27	82.11	14.65	70.53	15.71
JPEG	77.89	12.57	*82.11*	15.53	82.11	20.35
JPEG2000	32.63	11.61	26.32	14.7	21.05	16.91
GZIP	86.32	10.15	77.89	11.71	74.74	13.00
BZIP2	77.89	4.43	57.89	7.48	32.63	11.52

Table 3.2: For the color invoice database, the accuracy (A) and the classification error (E) for different image heights (25, 50, 100, 200, 400, and 800 pixels). Bold and italic numbers indicate, respectively, the best measure for each resolution and the best resolution for each measure.

relevant information at low resolution. Finally, the JPEG2000 compressor presents a similar behaviour for all resolutions, but, in general, it achieves poor results.

The classification error, that allows us to evaluate to what extent the classification is wrong when an invoice is misclassified, is also shown in Table 3.1. If this value is low, the system could suggest a short list of class candidates and the user could select the correct one, but this procedure is not recommendable when this value is high. From the results obtained, we can see that the mutual information-based measures and correlation coefficient have a better performance than the other measures with respect to the classification error. Note that these measures also achieve better results in accuracy.

For the color database (see Table 3.2), the behavior of the measures is similar to the one observed in the case of the black-and-white database, although, in general, the accuracy is lower. Note that I and NMI clearly perform better than the rest of the measures. Observe in Table 3.2 that the performance of CC has notably decreased with respect to I and NMI. It is important to emphasize the robustness of the mutual information-based measures taking into account that, as we have mentioned, the color database invoice images are not as well aligned as the invoices of the black-and-white database.

An important issue in our framework is the selection of the reference image of each class. In our approach, the first analyzed image of a class is taken as the representative one. The next experiment analyzes how the image similarity-based classification is sensitive to the choice of the reference image of the class. In this experiment, we have computed the similarity between each input image and all the images of its class and we have used as the reference image in the database the one that obtains the lowest similarity value. In this way, for each input image, the worst case is now considered with respect to the reference image of the correct class. For this study, we analyze all presented measures with image resolutions of 100 and 200 pixels. The results are summarized in Table 3.3. As it can be seen, the current results are slightly worse than the ones obtained in the previous experiment (see Tables 3.1 and 3.2), as it could be expected since the worst case is considered. Observe that the best performance is also achieved by I and NMI, obtaining an accuracy around 95% in almost all the cases. As we can observe, for the other measures, the accuracy decreases more significantly than for I and NMI. This shows that, I and NMI are less sensitive to the choice of the reference image of the class. The classification error has also similar values to the previous experiment, except for the cases where the presence of outliers decrease the quality of the results. For instance, observe the high increase of the classification error for the file compressors measures applied to the black-and-white database.

3.5.2 Experiment with a real-world database

In this experiment, the similarity measures presented in Section 3.3 have been applied on a real-world black-and-white database using only two image resolutions. Although, in the first experiment, the best results have been obtained with image

	Black-and-white database			
	100		200	
Measure	**A**	**E**	**A**	**E**
SSD	85.71	53.18	70.78	46.11
SAD	83.12	45.42	64.29	65.30
CC	95.45	3.29	93.51	7.10
I	96.10	2.83	**95.45**	6.86
NMI	**96.10**	2.67	95.45	9.43
PNG	86.36	30.90	88.96	45.53
JPEG	65.58	26.17	83.77	22.68
JPEG2000	49.35	34.19	50.00	35.57
GZIP	87.01	94.65	93.51	118.40
BZIP2	90.91	129.57	86.36	97.48

	Color database			
	100		200	
Measure	**A**	**E**	**A**	**E**
SSD	53.68	29.95	50.53	28.00
SAD	61.05	21.51	56.84	20.51
CC	84.21	8.13	81.05	4.72
I	94.74	2.20	90.53	2.11
NMI	**96.84**	2.00	**94.74**	2.00
PNG	85.26	6.50	76.84	6.55
JPEG	41.05	14.38	61.05	15.43
JPEG2000	20	15.79	21.05	22.23
GZIP	76.84	8.45	57.89	10.58
BZIP2	74.74	7.50	54.74	6.44

Table 3.3: For the black-and-white and color invoice database, the accuracy (A) and the classification error (E) for different image heights (100 and 200 pixels) when, for each input image, the worst reference image is considered. Bold numbers indicate the best measure for each resolution.

resolution of 50, 100, and 200 pixels, in the current experiment, we only consider image resolutions of 50 and 100 pixels in order to reduce the computational cost, which is linearly proportional to the number of pixels.

Our real-world database is composed by 2177 reference black-and-white invoices divided into 557 classes, and by 3673 input black-and-white invoices to be classified. In contrast to the previous experiments (Section 3.5.1), each input invoice can belong or not to a class and a class can have one or several reference invoices. Therefore, four different classification results can be obtained: *true positive* (TP), when the input invoice belongs to an existing class and the method found it; *true negative* (TN), when the input invoice does not belong to an existing class and the method did not found any class; *false positive* (FP), when the method classifies the input invoice in a class and the invoice does not belong to this class (either the invoice belongs to another class or it does not belong to any existing class); and *false negative* (FN), when the input invoice belongs an existing class and the method did not found any class. The first two types (TP and TN) are considered as correct classification results, while the FP and FN are considered as wrong ones.

In order to carry out this classification, it is necessary to introduce a threshold to discriminate when a input image belongs or not to a class. When the maximum similarity is lower than the threshold, we consider that the input invoice does not belong to any existing class. For comparison purposes with the previous experiment, the results are also presented without a threshold, where the input invoice is always classified in the most similar class. When the threshold is used, we consider all possible results: TP, TN, FP, and TN. On the contrary, when the threshold is not used, all input images are classified on a class and, therefore, only TP and FP (only when the invoice belongs to another class) results are possible.

The results without the use of a threshold are summarized in Table 3.4, where the accuracy and the classification error are shown. Note that, in general, the performance of the global image measures is similar, or slightly better, to the one obtained in the first experiment. In particular, NMI still shows the best performance. In this experiment, we do not use the BZIP2 compressor due to its high computational cost.

The results with threshold are shown in Table 3.5, where four different measures have been used: accuracy $A = (TP + TN)/(TP + TN + FP + FN)$, precision $P = TP/(TP + FP)$, recall $R = TP/(TP + FN)$, and $F - measure = 2(P \times R)/(P + R)$. For each measure, the threshold that maximizes the F-measure has been used. Observe that NMI obtains the best results and is also the least affected measure by the use of a threshold. Note that, in general, the performance of most of the measures with image resolution of 50 pixels is similar to the one obtained with resolution of 100 pixels. Thus, we propose to use an image resolution of 50 pixels since its computational cost is four times lower than using a resolution of 100 pixels.

With respect to the measures analyzed in this experiment, observe the very good performance achieved by the mutual information-based measures, but also by the intensity-based measures and the NCD using the PNG compressor. This behaviour is, to a great extent, due to the fact that the invoice images of the real-world database

	50		100	
Measure	**A**	**E**	**A**	**E**
SSD	98.15	401.60	97.85	272.44
SAD	99.65	73.08	98.67	7.00
CC	99.40	17.55	99.40	14.05
I	99.67	8.75	99.65	12.38
NMI	**99.95**	44.50	**99.84**	22.50
PNG	96.95	68.63	98.75	23.99
JPEG	16.66	232.86	64.88	104.41
JPEG2000	51.18	61.99	50.45	37.94
GZIP	96.62	38.19	89.52	1016.02

Table 3.4: For the real-world invoice database, the accuracy (A) and the classification error (E) for different image heights (50 and 100 pixels). Bold numbers indicate the best measure for each resolution.

	50				100			
Measure	**A**	**P**	**R**	**F**	**A**	**P**	**R**	**F**
SSD	95.32	95.32	100	97.60	95.02	95.02	100	97.45
SAD	96.80	96.80	100	98.38	95.83	95.83	100	97.87
CC	98.67	99.55	99.07	99.31	98.94	98.99	99.92	99.45
I	98.77	99.41	99.33	99.37	99.02	99.44	99.55	99.49
NMI	99.84	99.92	99.92	**99.92**	99.51	99.66	99.83	**99.75**
PNG	94.12	94.12	100	96.97	96.60	98.46	97.99	98.23
JPEG	13.83	13.83	100	24.30	62.73	63.17	97.93	76.80
JPEG2000	48.35	48.35	100	65.19	47.62	47.70	99.66	64.51
GZIP	93.79	93.79	100	96.80	87.67	88.85	98.16	93.27

Table 3.5: For the real-world invoice database, the accuracy (A), precision (P), recall (R), and F-measure (F) for different image heights (50 and 100 pixels). Bold numbers indicate the best measure for each resolution.

are black-and-white and have been very well aligned in the scanning process. Let us remember that the results obtained by both the intensity-based measures and the NCD measures with the testing color database were notably worse (see Table 3.2). Thus, taking into account all the experiments carried out in this section, we can confirm the excellent performance and robustness of the mutual information-based measures, specially of NMI.

Finally, we compare these results with the ones obtained using an OCR-based method. This method captures the text of the invoice image using OCR, finds text candidates to be a tax ID code, and compares these candidates with the existing ID codes in the database. If the tax ID code coincides with a class ID code, the method classifies this invoice to belong to this class; on the contrary, a new class is created. In this case, if there is not any ID code candidate from the OCR process, the invoice can not be classified. With this method, the accuracy is 88.81%, the precision is 100%, the recall is 86.96%, and the F-measure is 93.02%. As we can see, the performance of

the OCR-based method is significantly lower than most of the global image methods. After analyzing the classification errors of the OCR-based method, they have been classified in the following types:

- Errors of OCR reading. Confusions of letters and/or numbers (63.74%).
- Invoice with some physical manipulation, such a hole or a stamp, that prevents the correct reading of the tax ID code (15.09%).
- Invoice without tax ID code (7.79%).
- Invoice with noise that prevents the correct reading of the tax ID code (6.81%).
- Invoice with an identifier instead of a tax ID code. OCR does not recognize the identifier as a tax ID code (4.38%).
- Invoice with an incomplete tax ID code. In international invoices may be missing letters of the country (2.19%).

In our current implementation, the average computational time to classify an input image in our real-world database using the NMI measure, with image resolution of 50 pixels, is between 4 and 5 seconds, and the time required by the OCR-based method is between 20 and 50 seconds.

3.6 Conclusions

In this chapter, we have analyzed the behavior of different similarity measures applied to invoice classification. Three types of measures, applied to invoice processing, have been tested, based respectively on the intensity differences (sum of squared differences, sum of absolute differences, and correlation coefficient), the shared information (mutual information and normalized mutual information), and the normalized compression distance between two images, calculated from image (PNG, JPEG, and JPEG2000) and file (GZIP and BZIP2) compressors. The experiments have been carried out on two testing databases and a real-world database. In both cases, low resolution images have been used to show the good performance of the mutual information-based measures, although an acceptable performance has also been obtained with the correlation coefficient and the normalized compression distance implemented using file and image compressors. We have demonstrated the suitability of global similarity measures for invoice image classification.

CHAPTER 4

Tsallis mutual information for document classification

4.1 Introduction

The definition of the similarity between documents [Peng 2003] can be divided into two main groups based respectively on matching local features, such as the matching of recognized characters [Lopresti 2000] or different types of line segments [Tseng 1997], and on extracting global layout information, such as the use of a spatial layout representation [Hu 1999b] or geometric features [Shin 2001]. In this chapter, 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 [Hajnal 2001].

In the previous chapter, we found that the best measures on invoice classification were based on mutual information. This motivates us to investigate in this chapter the performance of several extensions of mutual information. Thus, we will study the application of three different Shannon-based generalizations of mutual information to analyze the similarity between scanned invoices. These three generalizations derive from Kullback-Leiber distance, the difference between entropy and conditional entropy, and the Jensen-Shannon divergence, respectively. In addition, the ratio between these measures is studied for different entropic indexes in the context of invoice classification and registration. A number of experiments are carried out to study the performance of the proposed measures using to invoice databases.

The content of this chapter has been published in "*Tsallis Mutual Information for Document Classification*", Marius Vila, Anton Bardera, Miquel Feixas, Mateu Sbert. Entropy, vol. 13, no. 9, pages 1694-1707, 2011 [Vila 2011].

4.2 Background

In this section, we review three different definitions of mutual information and the basis of image registration. Note that the previous work on document image similarity has been already presented in Chapters 2 and 3.

4.2.1 Mutual information definitions

In this section, we review three different ways to define mutual information. First, as we have seen in Section 2.2.2, the *mutual information* between two random variables X and Y can be defined by

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X). \tag{4.1}$$

$I(X;Y)$ is a measure of the shared information between X and Y.

Second, as we have also seen in Section 2.2.2, an alternative definition of $I(X;Y)$ can be obtained from the definition of the *informational divergence* or *Kullback–Leibler distance* (KL) [Cover 1991] as follows:

$$I(X;Y) = KL(p(x,y), p(x)p(y)). \tag{4.2}$$

Third, mutual information can be expressed as a *Jensen–Shannon divergence*. As we have seen in Section 2.2.3.3, the Jensen–Shannon inequality [Burbea 1982] is defined by:

$$JS(\pi_1,\ldots,\pi_n;p_1,\ldots,p_n) = H\left(\sum_{i=1}^{n}\pi_i p_i\right) - \sum_{i=1}^{n}\pi_i H(p_i) \geq 0, \tag{4.3}$$

where $JS(\pi_1,\ldots,\pi_n;p_1,\ldots,p_n)$ is the Jensen–Shannon divergence of probability distributions $p_1,p_2,\ldots,p_n$ with prior probabilities or weights $\pi_1,\pi_2,\ldots,\pi_n$, 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 $X \to Y$. Then, mutual information can be redefined as

$$I(X;Y) = JS(p(x_1),\ldots,p(x_n);p(Y|x_1),\ldots,p(Y|x_n)). \tag{4.4}$$

4.2.2 Image registration

Image registration is a fundamental task in image processing used to match two or more images or volumes obtained at different times, from different devices or from different viewpoints. Basically, it consists in finding the geometrical transformation that enables us to align images into a unique coordinate space. In the scope of this thesis we will focus on 2D rigid registration techniques because only transformations that consider translations and rotations are allowed.

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 [Lavallee 1995]: the transformation, the interpolator, the metric, and the optimizer (see Figure 4.1).

As input, we have both fixed **X** and moving **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. Each of these components is now described in more detail.

1. **Spatial transformation**. The registration process consists in reading the input image, defining the reference space (i.e. its resolution, positioning, and orientation) for each of these images, and establishing the correspondence between them (i.e. how to transform the coordinates from one image to the coordinates of the other image). The spatial transformation defines the spatial relationship between both images. Basically, two groups of transformations can be considered:

   - *Rigid or affine transformations*. These transformations can be defined with a single global transformation matrix. Rigid transformations are defined as geometrical transformations that only consider translations and rotations, and, thus, they preserve all distances. Affine transformations also allow shearing transforms and they preserve the straightness of lines (and the planarity of surfaces) but not the distances.
   - *Nonrigid or elastic transformations*. These transforms are defined for each of the points of the images with a transformation vector. For simplification purposes, sometimes only some control points are considered and the transformation at the other points is obtained by interpolating the transformation at these control points. Using these kinds of transformations, the straightness of the lines are not ensured.

   In this thesis, rigid image registration is our reference point.

2. **Interpolation**. The interpolation strategy determines the intensity value of a point at a non-grid position. When a general transformation is applied to an image, the transformed points may not coincide with the regular grid. So, an interpolation scheme is needed to estimate the values at these positions. One of the main problem of registration appears when there is not a direct correspondence between the coordinates of the two models. In this situation certain criteria has to be fixed to determine how this point has to be approximated in the second model. Therefore, spatial transformation rely for their proper implementation on interpolation and image resampling. Interpolation is the process of intensity based transformation and resampling is the process where intensity values are assigned to the pixels in the transformed image. Several interpolation schemes have been introduced [Lehmann 1999]. The most common are:

- *Nearest neighbour interpolation*: the intensity of each point is given by the one of the nearest grid-point.

- *Linear interpolation*: the intensity of a point is obtained from the linearweighted combination of the intensities of its neighbors.

- *Splines*: the intensity of a point is obtained from the spline-weighted combination of a grid-point kernel [Unser 1999].

- *Partial volume interpolation*: the weights of the linear interpolation are used to update the histogram, without introducing new intensity values [Collignon 1995].

3. **Metric**. The metric evaluates the similarity (or disparity) between the two images to be registered. Several image similarity measures have been proposed. They can be classified depending on the used features which are:

   - *Geometrical features*. A segmentation process detects some features and, then, they are aligned. These methods do not have high computational cost. Nevertheless, there is a great dependence on the initial segmentation results.

   - *Correlation measures*. The intensity values of each image are analyzed and the alignment is achieved when a certain correlation measure is maximized. Usually, a priori information is used in these metrics.

   - *Intensity occurrence*. These measures depend on the probability of each intensity value and are based on information theory [Shannon 1948].

4. **Optimization**. The optimizer finds the maximum (or minimum) value of the metric varying the spatial transformation. For the registration problem, an analytical solution is not possible. Then, numerical methods can be used in order to obtain the global extreme of a non analytical function. The most used methods in the image registration context are Powell's method, simplex method, gradient descent, conjugate-gradient method, and genetic algorithms (such as one-plus-one evolutionary). The choice of a method will depend on the implementation criteria and the measure features (smoothness, robustness, etc.). A detailed description of several numerical optimization methods and their implementations can be found in Press et al. [Press 1992].

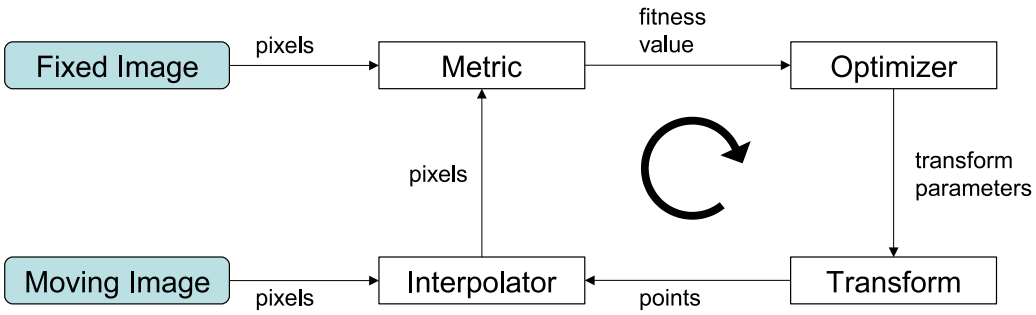


Figure 4.1: Main components of the registration process.

The crucial point of image registration is the choice of a metric. Some of these metrics, such as sum of squared differences, correlation coefficient or mutual information, are presented in Sections 3.3.1 and 3.3.2. Different measures derived from the Tsallis entropy have also been applied to image registration [Wachowiak 2003, Bardera 2004, Mohamed 2009, Khader 2010].

The main problems currently being addressed by image registration researchers are briefly summarized.

- **Robustness and accuracy**. To evaluate the behaviour of a registration method robustness and accuracy are the main parameters to be considered. The first parameter, robustness, refers to how the method behaves with respect to different initial states, i.e. different initial positions of the images, image noise, modality of the images, etc. The second parameter, accuracy, refers to how the final method solution is closer to the ideal solution. Constantly, new measures and new interpolation schemes appear trying to improve the robustness and the accuracy of the standard measures.

- **Artifacts**. In the registration process, the interpolator algorithm plays an important role, since usually the transformation brings the point to be evaluated into a non-grid position. This importance is greater when the grid size coincides in both images, since the interpolator pattern is repeated for each point. When the mutual information or its derivations, which are the most common measures used in multimodal image registration, are computed, their value is affected by both the interpolation scheme and the selected sampling strategy, limiting the accuracy of the registration. The fluctuations of the measure are called artifacts and are well studied by Tsao [Tsao 2003].

- **Speed-up**. One of main user requirements when using registration techniques is speed. Users desire results as fast as possible. The large amount of data acquired by current capture devices makes its processing difficult in terms of time. Therefore, the definition of strategies able to accelerate the registration process is fundamental. Several multiresolution frameworks have been proposed achieving better robustness and speeding up the process.

4.3 Generalized mutual information

We review here three different mutual information generalizations inspired by the three different forms of mutual information presented in Section 4.2.1: the Kullback–Leibler distance, the difference between entropy and conditional entropy, and the Jensen–Shannon divergence, respectively. These generalizations are based on the Harvda-Charvát-Tsallis entropy defined in Section 2.2.7.

4.3.1 Mutual information

As we have seen in Section 2.2.7, *Tsallis mutual information* can be defined [Taneja 1988, Tsallis 1998] from Equations (2.10) and (2.36) as

$$\begin{aligned} I_\alpha(X;Y) &= KL_\alpha\big(p(x,y),p(x)p(y)\big) \\ &= \frac{1}{1-\alpha}\left(1-\sum_{x\in\mathscr{X}}\sum_{y\in\mathscr{Y}}\frac{p(x,y)^\alpha}{p(x)^{\alpha-1}p(y)^{\alpha-1}}\right). \end{aligned} \tag{4.5}$$

In Section 2.2.7 we have also presented the generalization of NMI given by

$$NMI_\alpha(X;Y) = \frac{I_\alpha(X;Y)}{H_\alpha(X,Y)}. \tag{4.6}$$

Although $NMI_\alpha(X;Y)$ is a normalized measure for $\alpha \to 1$, this is not true for other α values, since NMI_α can take values greater than 1. This measure is always positive and symmetric.

4.3.2 Mutual entropy

Another way of generalizing mutual information is the so-called *Tsallis mutual entropy* [Furuichi 2006]. From Equation (2.8), the Tsallis mutual entropy is defined for $\alpha > 1$ as

$$\begin{aligned} ME_\alpha(X;Y) &= H_\alpha(X)-H_\alpha(X|Y)=H_\alpha(Y)-H_\alpha(Y|X) \\ &= H_\alpha(X)+H_\alpha(Y)-H_\alpha(X,Y). \end{aligned} \tag{4.7}$$

This measure is positive and symmetric, and the Tsallis joint entropy $H_\alpha(X,Y)$ is an upper bound of it [Furuichi 2006]. Tsallis mutual entropy represents a kind of correlation between X and Y.

As in Furuichi [Furuichi 2006], the normalized Tsallis mutual entropy can be defined as

$$NME_\alpha(X;Y) = \frac{ME_\alpha(X;Y)}{H_\alpha(X,Y)}. \tag{4.8}$$

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 $X = Y$ [Furuichi 2006].

4.3.3 Jensen–Tsallis information

Since Tsallis entropy is a concave function for $\alpha > 0$, the Jensen–Shannon divergence (see Equation (2.17)) can be extended to define the *Jensen–Tsallis divergence*:

$$JT_\alpha(\pi_1,\ldots,\pi_n;p_1,\ldots,p_n) = H_\alpha\left(\sum_{i=1}^{n}\pi_i p_i\right) - \sum_{i=1}^{n}\pi_i H_\alpha(p_i). \tag{4.9}$$

As we have seen in Equation (4.4), Jensen–Shannon divergence coincides with $I(X;Y)$ when $\{\pi_1,\ldots,\pi_n\}$ is the marginal probability distribution $p(x)$, and $\{p_1,\ldots,p_n\}$ are the rows $p(Y|x)$ of the probability conditional matrix of the channel. Then, for the channel $X \to Y$, a generalization of mutual information, which we call *Jensen–Tsallis Information* (JTI_α) can be expressed by

$$\begin{aligned} JTI_\alpha(X \to Y) &= JT_\alpha(p(x);p(Y|x)) = H_\alpha(\sum_{x\in\mathcal{X}} p(x)p(Y|x)) - \sum_{x\in\mathcal{X}} p(x)H_\alpha(Y|x) \\ &= H_\alpha(Y) - \sum_{x\in\mathcal{X}} p(x)H_\alpha(Y|x). \end{aligned} \tag{4.10}$$

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

$$JTI_\alpha(Y \to X) = JT_\alpha(p(x);p(Y|x)) = H_\alpha(X) - \sum_{y\in\mathcal{Y}} p(y)H_\alpha(X|y). \tag{4.11}$$

This measure is positive and, in general, non-symmetric with respect to the reversion of the channel. Thus, $JTI_\alpha(X \to Y) \neq JTI_\alpha(Y \to X)$. An upper bound of this measure is given by the Tsallis joint entropy: $JTI_\alpha \leq H_\alpha(X,Y)$. The Jensen–Tsallis divergence and its properties have been studied by Bardera et al. [Bardera 2004] and Hamza [Hamza 2006].

Similar to the previous measures, a normalized version of JTI_α can be defined as

$$NJTI_\alpha(X \to Y) = \frac{JTI_\alpha(X \to Y)}{H_\alpha(X,Y)}. \tag{4.12}$$

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

4.4 Methodology

The main goal of this chapter is to analyze the application of the Tsallis-based generalizations of mutual information (Section 4.3) to the invoice classification

process.

As the methodology followed in this chapter is the same that the one used in Chapter 3, we only remember here its main points:

- A database is constituted by classes, where each class is defined by one or more representative images.
- Two different groups of images are defined: reference images, where each image belongs to a class, and input images used as the input to the classifier.
- The used scanning protocol produces quasi-aligned images. Despite this, all images are preprocessed to obtain a more reliable evaluation of the image similarity. Skew errors are corrected and images are downscaled to analyze the behaviour of the classification process for different image resolutions.
- Similarity between the input image and all the reference images is computed, and the input image is assigned to the class of the reference image for which the similarity value is maximum and greater than a given threshold. If this maximum value is lower than the threshold, we consider that the input invoice does not correspond to any existing class, and the new class is added to the database.

A general scheme of our framework is represented in Figure 3.4.

Another objective of this chapter is to analyze the performance of the Tsallis-based generalizations of mutual information in aligning two invoices. This is also a critical point since it allows us to find the spatial correspondence between an input invoice and a template. The registration framework used in this chapter is represented in Figure 4.1.

4.5 Results

This section is split into two parts. In the first part, we present the experiments carried out with one testing database (same color invoice database used in Section 3.5.1) with the aim of studying the behaviour of the presented measures (see Section 4.3). In the second part, we analyze the behaviour of these measures with a real-world database (same real-world database used in Section 3.5.2). The proposed measures have been implemented in Visual C++ .NET and their performance has been measured on a system with an Intel Core 2 Duo CPU. For the full description of both databases (testing color database and real-world database) and methodological details, see Section 3.5.

4.5.1 Experiments with a testing database

In our first experiment, we analyze the behaviour of the mutual information generalizations presented in Section 4.3 using the color invoice database presented

in Section 3.5.1. We also analyze the performance of these Tsallis-based generalizations of mutual information in aligning two invoices.

Preliminary experiments (see Section 3.5) have shown that the best classification results are obtained for resolutions with height between 50 and 200 pixels. Note that this fact greatly speeds up the computation process as computation time is proportional to image resolution. In this experiment, all images have been scaled from the original scanning resolution (around 2500×3500 pixels) to a height of 100 pixels, conveniently adjusting the image width to keep the aspect ratio of the images.

Table 4.1 shows the two performance values for each measure and different α values. Note that the values for the ME_α and NME_α 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. As in the previous experiment the first parameter represents the accuracy and the second represents the classification error. We can observe that the measures have a different behavior with respect to the α values. While I_α and NMI_α achieve the best classification success for α values between 0.4 and 1.2, the rest of the measures (ME_α, NME_α, JTI_α, $NJTI_\alpha$) perform better for α values between 1.0 and 1.4. For these values, the normalized measures classify correctly all the invoices. In general, the normalized measures perform much better than the corresponding non normalized ones.

Finally, the last experiment analyzes the capability of the Tsallis-based proposed measures to align similar invoices 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.* [Capek 2001]. 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 4.2, the left plot represents the results for the I_α, ME_α, and JTI_α measures with different α values and the right plot represent the results for their corresponding normalized measures. As it can be seen, the best results are achieved for α 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.

The second feature that we will analyze for the alignment experiment is the accuracy. In this case, the general registration scheme of Figure 4.1 has been applied, where we have used the Powell's method optimizer [Press 1992], a rigid

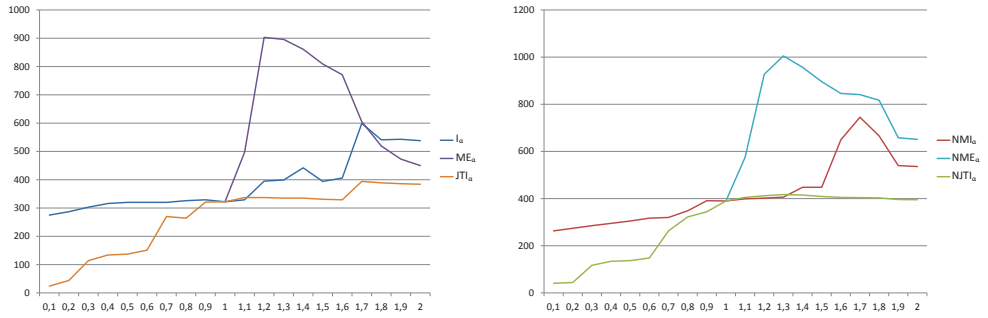


Figure 4.2: AFA parameter values with respect to the α value for the I_α, ME_α, and JTI_α measures (left) and the corresponding normalized measures (right). AFA parameter evaluates the range of convergence of a registration measure to its global maximum.

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 × 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 4.3, for each measure and each α value, the mean error is plotted. In this case, we can not derive a general behavior. I_α performs better for $\alpha = 1.6$, while NMI_α for $\alpha = 0.4$. In this case, the non normalized measure performs better than the normalized one. Both ME_α and NME_α 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 α value increases. Among all measures, the normalized Jensen–Tsallis information achieves the best results, obtaining the minimum error (and thus the maximum accuracy) for $\alpha = 0.3$.

As a conclusion, for invoice classification, the best results have been obtained by the normalized measures, using α values between 0.4 and 1.2 for NMI_α and between 1 and 1.4 for NME_α and $NJTI_\alpha$. For invoice registration, the most robust results have been obtained by NME_α with $\alpha = 1.3$ and the most accurate ones have been achieved by $NJTI_\alpha$ with $\alpha = 0.3$.

4.5.2 Experiment with a real-world database

In this second experiment, the similarity measures introduced in Section 4.3 have been applied on the black-and-white real-world database (presented in Section 3.5.2) using only two image resolutions (50 and 100 pixels).

In contrast to the previous experiments (Section 4.5.1), each input invoice can

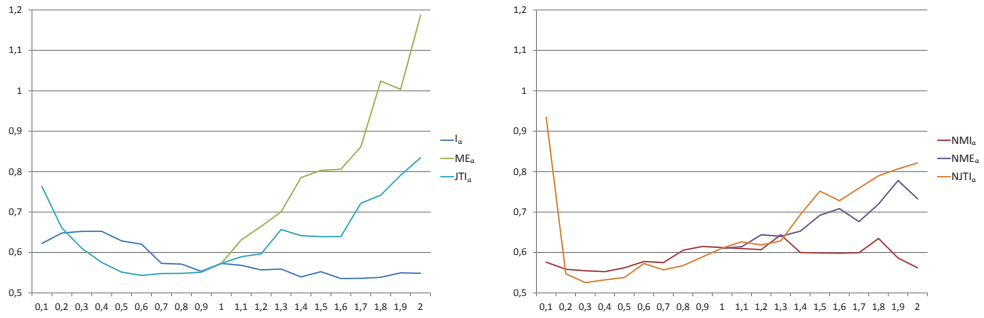


Figure 4.3: Mean error at the final registration position for different measures and α values for the I_α, ME_α, and JTI_α measures (left) and the corresponding normalized measures (right).

belong or not to a class and a class can have one or several reference invoices. As in Section 3.5.2 four different classification results can be obtained: *true positive* (TP), *true negative* (TN), *false positive* (FP), and *false negative* (FN). The first two types (TP and TN) are considered as correct classification results, while the FP and FN are considered as wrong ones.

In order to carry out the classification, we introduce a threshold to discriminate when a input image belongs or not to a class. When the maximum similarity is lower than the threshold, we consider that the input invoice does not belong to any existing class. For comparison purposes with the previous experiment, the results are also presented without a threshold, where the input invoice is always classified in the most similar class.

The results without the use of a threshold are summarized in Table 4.2, where the accuracy and the classification error are shown. We can observe that, for all measures and image resolutions, the best results are obtained for α values between 1.0 and 1.4.

The results with threshold are shown in Table 4.3, where four different measures have been used: accuracy $A = (TP + TN)/(TP + TN + FP + FN)$, precision $P = TP/(TP+FP)$, recall $R = TP/(TP+FN)$, and $F-measure = 2(P \times R)/(P+R)$. For each measure, the threshold that maximizes the $F-measure$ has been used. Observe that NMI obtains the best results and is also the least affected measure by the use of a threshold. It can be seen from Table 4.2 that, without the use of a threshold, the best results are obtained by the Tsallis-based measures for α values between 1.0 and 1.4. Otherwise, when a threshold is used (Table 4.3), the best results are obtained by Tsallis-based measures for $\alpha = 1.0$ (i.e., Shannon-based measures). Note too that the performance of most of the measures with image resolution of 50 pixels is similar to the one obtained with resolution of 100 pixels. Thus, we propose to use an image resolution of 50 pixels since its computational cost is four times lower than using a resolution of 100 pixels.

4.6 Conclusions

In this chapter, we have introduced three different mutual information generalizations for invoice classification. These measures have been inspired respectively by Kullback–Leibler distance, the difference between entropy and conditional entropy, and the Jensen–Shannon divergence, and their ratio with the Tsallis joint entropy. The experiments have been carried out on a testing database and a real-world database, both with and without the use of a threshold. When the threshold is used, Tsallis-based measures obtain the best results for α values between 1.0 and 1.4 whereas, when the threshold is not used, the best results are obtained for $\alpha = 1.0$, i.e, when Shannon-based measures are applied. In both cases, low resolution images have been used to show the good performance of the mutual information-based measures. Finally, the invoice registration using measures based on mutual information generalizations 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.

	I_α		NMI_α		ME_α		NME_α		JTI_α		$NJTI_\alpha$	
α *value*	A	E	A	E	A	E	A	E	A	E	A	E
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	3.00	**100.0**	0.00	**98.95**	3.00	**100.0**	0.00	**98.95**	3.00	**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

Table 4.1: For the color invoice database, the accuracy (A) and the classification error (E) for 100 pixels image height and different α values. Bold numbers indicate the best α values for each measure.

		50		100	
Measure	α	**A**	**E**	**A**	**E**
I_α	0.4	99.35	13.00	99,32	12,21
	0.6	99.52	12.47	99,58	14,07
	0.8	99.58	10.20	99,61	13,21
	1.0	99.67	8.75	99.65	12.38
	1.2	**99.83**	17.00	99,69	13,36
	1.4	99.83	21.33	**99,72**	15,50
NMI_α	0.4	99.58	15.40	99.47	12.79
	0.6	99.80	22.71	99.66	13.50
	0.8	99.83	18.17	99.80	19.71
	1.0	99.95	44.50	99.84	22.50
	1.2	**99.97**	88.00	99.83	21.83
	1.4	99.92	30.00	**99.86**	27.40
ME_α	0.4				
	0.6				
	0.8				
	1.0	**99.67**	8.75	**99.65**	12.38
	1.2	94.78	5.69	90.01	9.71
	1.4	72.23	10.44	57.53	11.18
NME_α	0.4				
	0.6				
	0.8				
	1.0	**99.95**	44.50	99.84	22.50
	1.2	99.89	17.75	**99.89**	24.75
	1.4	99.86	8.20	99.66	9.42
JTI_α	0.4	91.95	14.54	96.99	28.45
	0.6	98.63	8.77	99.12	18.81
	0.8	99.66	13.00	99.58	14.40
	1.0	99.67	8.75	99.65	12.38
	1.2	99.72	9.10	**99.66**	16.17
	1.4	**99.75**	9.89	99.63	17.54
$NJTI_\alpha$	0.4	92.70	12.78	94.54	27.09
	0.6	99.09	8.41	99.10	11.63
	0.8	99.92	50.33	99.69	15.73
	1.0	**99.95**	44.50	99.84	22.50
	1.2	99.94	42.00	99.83	27.00
	1.4	99.94	38.50	**99.86**	36.20

Table 4.2: For the real-world invoice database, the accuracy (A) and the classification error (E) for different image heights (50 and 100 pixels) and α values. Bold numbers indicate the best combination of measure and α value for each resolution.

		50				100			
Measure	α	**A**	**P**	**R**	**F**	**A**	**P**	**R**	**F**
I_α	0.4	95.91	95.91	100	97.91	96.10	96.10	100	98.01
	0.6	96.27	96.27	100	98.10	96.46	96.46	100	98.20
	0.8	96.40	96.40	100	98.17	96.65	96.65	100	98.30
	1.0	98.77	99.41	99.33	**99.37**	99.02	99.44	99.55	**99.49**
	1.2	96.70	96.70	100	98.32	97.17	99.37	97.69	98.53
	1.4	96.68	96.68	100	98.31	96.98	99.34	97.53	98.43
NMI_α	0.4	96.68	96.68	100	98.31	96.49	96.49	100	98.21
	0.6	97.28	99.51	97.67	98.58	97.68	99.46	98.15	98,80
	0.8	99.24	99.47	99.75	99.61	98.37	99.49	98.82	99.15
	1.0	99.84	99.92	99.92	**99.92**	99.51	99.66	99.83	**99.75**
	1.2	99.16	99.41	99.72	99.57	98.47	99.49	98.93	99.21
	1.4	98.66	99.41	99.21	99.31	98.23	99.29	98.88	99.08
ME_α	0.4								
	0.6								
	0.8								
	1.0	98.77	99.41	99.33	**99.37**	99.02	99.44	99.55	**99.49**
	1.2	88.50	88.50	100	93.90	83.19	83.19	100	90.82
	1.4	64.28	64.28	100	78.26	50.46	50.46	100	67.08
NME_α	0.4								
	0.6								
	0.8								
	1.0	99.84	99.92	99.92	**99.92**	99.51	99.66	99.83	**99.75**
	1.2	98.88	99.55	99.30	99.42	97.68	99.18	98.43	98.80
	1.4	98.47	99.55	98.88	99.21	97.11	99.17	97.83	98.50
JTI_α	0.4	86.87	86.87	100	92.97	92.04	92.04	100	95.86
	0.6	94.09	94.09	100	96.95	95.45	95.45	100	97.67
	0.8	96.38	96.38	100	98.15	96.46	96.46	100	98.20
	1.0	98.77	99.41	99.33	**99.37**	99.02	99.44	99.55	**99.49**
	1.2	97.22	99.37	97.75	98.55	96.57	96.57	100	98.25
	1.4	97.28	99.17	98.00	98.59	96.51	99.42	96.96	98.18
$NJTI_\alpha$	0.4	87.17	87.17	100	93.14	87.68	87.68	100	93.44
	0.6	95.69	98.79	96.72	97.75	95.67	95.67	100	97.79
	0.8	98.64	99.41	99.19	99.30	97.38	99.40	97.89	98.64
	1.0	99.84	99.92	99.92	**99.92**	99.51	99.66	99.83	**99.75**
	1.2	99.29	99.61	99.66	99.64	98.20	99.52	98.62	99.07
	1.4	99.29	99.61	99.66	99.64	98.09	99.55	98.48	99.01

Table 4.3: For the real-world invoice database, the accuracy (A), precision (P), recall (R), and F-measure (F) for different image heights (50 and 100 pixels) and α values. Bold numbers indicate the best combination of measure and α value for each resolution.

CHAPTER 5

Shot boundary detection and keyframe selection

5.1 Introduction

Video shot boundary detection, or the segmentation of a video sequence in its constituent shots, is a fundamental step in video data management. In this chapter, we propose two new information-theoretic measures based on both Tsallis mutual information and Jensen-Tsallis divergence to detect the shot boundaries of a video sequence. Their performance is analyzed for a set of video sequences using several entropic indices, color spaces, and regular binnings. These similarity measures are also used to select the most representative keyframe of a video shot. In this case, we assume that the maximal representativeness of a frame is achieved when its average similarity with the other frames of the shot is maximum.

Several experiments analyze the performance of the proposed mutual information-based measures using different color spaces, such as HSV or Lab, and a few number of histogram bins. These measures are devised to detect abrupt video cuts and are not designed to deal with gradual transitions.

The content of this chapter has been published in "*Tsallis entropy-based information measure for shot boundary detection and keyframe selection*", Marius Vila, Anton Bardera, Qing Xu, Miquel Feixas, Mateu Sbert. Signal, Image and Video Processing, vol. 7, no. 3, pages 507-520, 2013 [Vila 2013].

5.2 Previous work

In this section, we review some basic concepts on video shot boundary detection and keyframe selection, and on basic information-theoretic tools.

5.2.1 Shot boundary detection and keyframe selection

As we have introduced in Section 2.4 a video shot is defined as a sequence of frames captured by one camera in a single continuous action in time and space. In general, video shots have similar visual features, such as color, texture, or motion and their boundaries are commonly associated to abrupt changes of these features. Many measures and algorithms have been proposed to detect video discontinuities and to extract the most meaningful or representative frames of a video sequence [Money 2008, Peng 2010].

Various approaches for shot boundary detection are based on information-theoretic measures, such as entropy and mutual information. Butz and Thiran [Butz 2001] present a novel approach that uses mutual information in the gray-scale space to measure changes between subsequent frames in image sequences. In order to compensate camera panning and zooming, they use affine image registration that make the approach computationally expensive. Cernekova et al. [Cernekova 2006] also present a shot detection method based on mutual information and joint entropy between frames. This approach achieves very good results using an adaptive thresholding based on the local mutual information values in a temporal window. Mentzelopoulos and Psarrou [Mentzelopoulos 2004] propose an entropy difference to perform spatial frame segmentation. Huan et al. [Huan 2008] use mutual information and canny edge detector to distinguish object motion. Xu et al. introduced Jensen-Shannon [Xu 2010, Xu 2014], Jensen-Rényi [Xu 2012] and f-divergences [Luo 2014] to detect shot boundaries and keyframes.

Although most methods use the RGB color space or the luminance value for analyzing the video sequence, some works have studied different color spaces. For instance, Gargi et al. [Gargi 2000] investigate the efficacy of some methods for cut detection and the effect of color space representation on the performance of histogram-based shot detection. Zhang et al. [Zhang 2006] also use the HSV histogram differences of two consecutive frames as a feature for evaluating the color information.

5.2.2 Tsallis entropy and Jensen-based divergence

As we have presented in Section 2.2.7, the Tsallis entropy $H_\alpha(X)$ of a discrete random variable X is defined by

$$H_\alpha(X) = \frac{1 - \sum_{x \in \mathcal{X}} p(x)^\alpha}{\alpha - 1}, \tag{5.1}$$

where $\alpha \in \mathbb{R} - \{1\}$ is called entropic index. Let us remember that the Shannon entropy, defined using natural logarithms, is recovered when $\alpha \to 1$. The Tsallis mutual information is defined [Taneja 1988, Tsallis 1998] by

$$I_\alpha(X;Y) = \frac{1}{1-\alpha}\left(1 - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \frac{p(x,y)^\alpha}{p(x)^{\alpha-1} p(y)^{\alpha-1}}\right). \tag{5.2}$$

Shannon mutual information is also recovered when $\alpha \to 1$.

As we have seen in Section 2.2.3.3, the *Jensen-Shannon inequality* [Burbea 1982] is defined by

$$JS(\pi_1, \ldots, \pi_n; p_1, \ldots, p_n) = H\left(\sum_{i=1}^{n} \pi_i p_i\right) - \sum_{i=1}^{n} \pi_i H(p_i) \geq 0, \tag{5.3}$$

where $JS(\pi_1,\ldots,\pi_n;p_1,\ldots,p_n)$ is the *Jensen-Shannon divergence* (JS-divergence) of probability distributions $\{p_1,\ldots,p_n\}$ with prior probabilities or weights $\{\pi_1,\ldots,\pi_n\}$ fulfilling $\sum_{i=1}^{n}\pi_i = 1$. The JS-divergence measures how far the probabilities p_i are from their likely joint source $\sum_{i=1}^{n}\pi_i p_i$ and equals zero if and only if all the p_i are equal. It is important to note that the JS-divergence is identical to $I(X;Y)$ when $\{\pi_1,\ldots,\pi_n\}$ and $\{p_1,\ldots,p_n\}$ represent, respectively, the input distribution and the probability transition matrix of the channel $X \rightarrow Y$, where $n = |\mathscr{X}|$ and $m = |\mathscr{Y}|$ [Burbea 1982, Slonim 2000b].

Similar to Equation (5.3), the *Jensen-Tsallis inequality* is given by

$$JT_\alpha(\pi_1,\ldots,\pi_n;p_1,\ldots,p_n) = H_\alpha(\sum_{i=1}^{n}\pi_i p_i) - \sum_{i=1}^{n}\pi_i H_\alpha(p_i) \geq 0, \tag{5.4}$$

where $JT_\alpha(\pi_1,\ldots,\pi_n;p_1,\ldots,p_n)$ is the *Jensen-Tsallis divergence* (JT-divergence) of probability distributions $\{p_1,\ldots,p_n\}$ with weights $\{\pi_1,\ldots,\pi_n\}$. Since Tsallis entropy is a concave function for $\alpha > 0$, JT-divergence is positive for $\alpha > 0$ [Martins 2008].

5.3 Shot boundary detection

In this section, we present two new approaches based, respectively, on Tsallis mutual information and Jensen-Tsallis divergence to detect the abrupt shot boundaries of a video sequence. Then, we describe three new measures to extract the most representative keyframes.

5.3.1 Mutual information-based similarity between frames

As we have mentioned in Section 5.2, Cernekova et al. [Cernekova 2006] presented a shot detection method based on mutual information. The authors defined the similarity between two consecutive frames i and j as

$$I^{RGB}(i;j) = I^R(i;j) + I^G(i;j) + I^B(i;j), \tag{5.5}$$

where the superindices R, G, and B stand for the red, green, and blue color components, respectively, and $I^c(i;j)$ is the mutual information (Equation (2.8)) between frames i and j for a given color component c. The marginal probabilities $p(x)$ and $p(y)$ used in the computation of mutual information are given by the normalized histograms of the corresponding color component of frames i and j, respectively, and the joint probability $p(x,y)$ is given by the probability of finding the value x in the frame i and the value y in the frame j at the same pixel location.

Cernekova et al. [Cernekova 2006] also proposed a ratio between the mutual information $I^{RGB}(i;j)$ and its average value in the neighbourhood of pair (i,j) in order to capture relative variations of $I^{RGB}(i;j)$ with respect to the surrounding

frames. This ratio was defined as

$$IR^{RGB}(i;i+1) = \frac{I^{RGB}(i;i+1)}{\frac{1}{2r}\sum_{j=i-r,j\neq i}^{i+r} I^{RGB}(j;j+1)}, \tag{5.6}$$

where r is the radius of the window centered in the transition between frame i and $i+1$ and given by the frames $\{i-r,\ldots,i-1,i+1,\ldots,i+r\}$ (see Figure 5.1).

The measure $I^{RGB}(i;j)$ (see Equation (5.5)) is now extended using Tsallis entropy and different color spaces. Thus, the *informational frame similarity* is defined by

$$I_\alpha^{xyz}(i;j) = I_\alpha^x(i;j) + I_\alpha^y(i;j) + I_\alpha^z(i;j), \tag{5.7}$$

where the superindices x, y, and z stand for the color components in a determined color space and $I_\alpha^c(i;j)$ is the Tsallis mutual information (Equation (5.2)) between frames i and j for a given color component c. As it was noted by Portes de Albuquerque et al. [Portes de Albuquerque 2004], the main motivation for the use in image and video processing of non-extensive measures, such as Tsallis entropy, is the presence of correlations between pixels of the same object in the image that can be considered as long-range correlations. Note that the use of the Tsallis generalization of mutual information will allow us to analyze the performance of the similarity measure using different entropic indices and, thus, to select the entropic index that better discriminates a shot boundary. In this chapter, we use the following color spaces: *Lab* (abbreviation for the CIE 1976, also called CIELAB), *HSV* (abbreviation for hue, saturation, and value), and *RGB*. Lab color space is perceptually uniform and has been designed to approximate human vision. HSV color space separates lightness from chrominance information and it is not perceptually uniform [Gonzalez 2002, Thompson 2011]. Both Lab and HSV color spaces have less redundancy between the color components than RGB encoding.

The ratio IR (see Equation (5.6)) is generalized by defining the *informational frame similarity ratio* as

$$IR_\alpha^{xyz}(i;i+1) = \frac{I_\alpha^{xyz}(i;i+1)}{IW_\alpha^{xyz}(i,r)}, \tag{5.8}$$

where the *average informational similarity in a window* is given by

$$IW_\alpha^{xyz}(i,r) = \frac{1}{2r}\sum_{j=i-r,j\neq i}^{i+r} I_\alpha^{xyz}(j;j+1), \tag{5.9}$$

where r is the radius of the window given by the frames $\{i-r,\ldots,i-1,i+1,\ldots,i+r\}$. In our experiments, we use $r=2$ (see Figure 5.1) as in Xu et al. [Xu 2010].

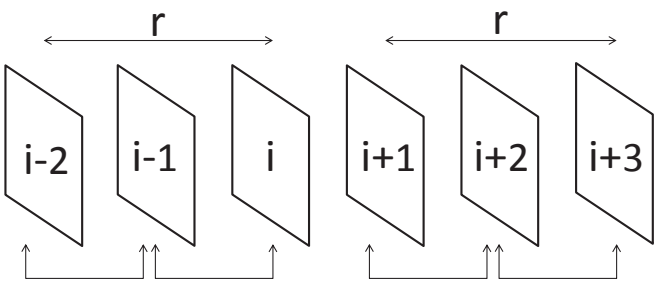


Figure 5.1: Set of frames with radius r centered in the transition between frame i and $i+1$.

5.3.2 Jensen-Tsallis-based similarity between frames

As we have commented in Section 5.2.1, Xu et al. [Xu 2010] applied the Jensen-Shannon divergence to quantify the frame dissimilarity. The dissimilarity between frames i and j was defined for an RGB color space as

$$\begin{aligned} JS^{RGB}(i,j) &= JS^{R}(\frac{1}{2},\frac{1}{2};p_i^R,p_j^R) \\ &+ JS^{G}(\frac{1}{2},\frac{1}{2};p_i^G,p_j^G) \\ &+ JS^{B}(\frac{1}{2},\frac{1}{2};p_i^B,p_j^B), \end{aligned} \tag{5.10}$$

where p_i^c and p_j^c are, respectively, the normalized histograms of frames i and j for a given color component c.

Similar to the extension of mutual information (Section 5.3.1), the measure JS^{RGB} is extended using Jensen-Tsallis divergence and several color spaces. From Equation (5.4), *Jensen-Tsallis divergence between two frames* i and j is defined as

$$\begin{aligned} JT_\alpha^{xyz}(i,j) &= JT_\alpha^{x}(\frac{1}{2},\frac{1}{2};p_i^x,p_j^x) \\ &+ JT_\alpha^{y}(\frac{1}{2},\frac{1}{2};p_i^y,p_j^y) \\ &+ JT_\alpha^{z}(\frac{1}{2},\frac{1}{2};p_i^z,p_j^z), \end{aligned} \tag{5.11}$$

where x, y, and z stand for the color components of a given color space, $JT_\alpha^c(\frac{1}{2},\frac{1}{2};p_i^c,p_j^c)$ is the Jensen-Tsallis divergence between the frames i and j for a given color component c (Equation (5.4)), and p_i^c and p_j^c are respectively given by the normalized histograms of frames i and j.

As we are interested in a similarity measure between frames, we compute now the complementary measure of $JT_\alpha^{xyz}(i,j)$ with respect to its maximum value. From Equations (5.1) and (5.4), it can be seen that the maximum value of JT_α^c between

two probability distributions depends on the parameter α and is given by

$$\begin{aligned} M_\alpha^c &= \max JT_\alpha^c(\frac{1}{2},\frac{1}{2};p_i^c,p_j^c) \\ &= \frac{1}{1-\alpha}\left(n^{1-\alpha}-1\right)-\frac{1}{1-\alpha}\left(\left(\frac{n}{2}\right)^{1-\alpha}-1\right), \end{aligned} \tag{5.12}$$

where $\alpha \neq 1$ and n is the number of histogram bins of the color component c. Using Jensen-Shannon divergence (Equation (5.3)), it can be seen that if $\alpha = 1$, then $M = 1$. Thus, $M_\alpha^c - JT_\alpha^c(\frac{1}{2},\frac{1}{2};p_i^c,p_j^c)$ can be seen as a similarity measure between two frames for a given color component c. This allows us to define the *Jensen-Tsallis frame similarity* between two frames i and j as

$$\begin{aligned} \mathrm{JT}_\alpha^{xyz}(i,j) &= M_\alpha^x - JT_\alpha^x(\frac{1}{2},\frac{1}{2};p_i^x,p_j^x) \\ &+ M_\alpha^y - JT_\alpha^y(\frac{1}{2},\frac{1}{2};p_i^y,p_j^y) \\ &+ M_\alpha^z - JT_\alpha^z(\frac{1}{2},\frac{1}{2};p_i^z,p_j^z), \end{aligned} \tag{5.13}$$

where x, y, and z stand for the color components of a given color space. Observe that this measure only deals with marginal probabilities, but not with joint probabilities, and, therefore, its computation is faster than the computation of Tsallis mutual information (Equation (5.7)).

Similar to the previous mutual information-based measures (Section 5.3.1), given a frame i, the *Jensen-Tsallis frame similarity ratio* between the frame similarity of pair $(i, i+1)$ and the average frame similarity in its neighbourhood is defined by

$$\mathrm{JTR}_\alpha^{xyz}(i,i+1) = \frac{\mathrm{JT}_\alpha^{xyz}(i,i+1)}{\mathrm{JTW}_\alpha^{xyz}(i,r)}, \tag{5.14}$$

where

$$\mathrm{JTW}_\alpha^{xyz}(i,r) = \frac{1}{2r}\sum_{j=i-r,j\neq i}^{i+r} \mathrm{JT}_\alpha^{xyz}(j,j+1) \tag{5.15}$$

is the *average Jensen-Tsallis similarity in a window* of radius r. In our experiments, we use $r = 2$ [Xu 2010]. Let us note that $\mathrm{JTR}_\alpha^{RGB}$ for $\alpha = 1$ is slightly different from the normalization of the Jensen-Shannon divergence proposed by Xu et al. [Xu 2010].

5.4 Keyframe selection

Given a video shot with m frames, three simple measures are proposed to extract the most representative keyframes. For a given frame i of a shot s, its average similarity with respect to the rest of frames of s can be computed using the Tsallis mutual

Filename	#F	#C
amc_jeep	1646	27
apo16005	691	4
FPTVKyrgyzstan	1439	22
FPTVPakistan	723	18
indi001	1686	15
indi002	844	6
indi007	3469	24
indi008	705	4
indi105	1109	9
indi106	3610	9
indi108	2673	13
loop_a_lot	1754	11
monkey_uncle	1688	31
newport	1765	16
newport_2	1537	12
newport_3	1538	11
newport_5	1713	15
newport_8	1749	10
parker_brothers	1595	34
rca_victor	1635	8
sharp_calculator	1743	13
tide	1736	15
trik_trak	1751	37
UGS07_007	3513	11
uist01_13	1766	15
uist97_11	2856	13
wth-02	385	2

Table 5.1: List of 27 videos (with filename, number of frames (#F), and number of shot boundaries (#C)) used in our experiments. Obtained from the video database The Open Video Project [Open Video Project].

information and the Jensen-Tsallis divergence.

The *average informational similarity* of frame i with respect to the other frames of shot s is defined by

$$AI_{\alpha}^{xyz}(i) = \frac{1}{m-1} \sum_{j=1, j\neq i}^{m} I_{\alpha}^{xyz}(i;j), \tag{5.16}$$

where m is the number of frames of shot s, and j represents a frame of shot s different of i. From this measure, the keyframe of a shot is given by the frame with the highest average similarity. Note that AI mainly takes into account the spatial distribution of intensity values and achieves high values when the distribution in the frame i highly correlates with the distribution in the other frames of the shot (see Figure 5.2).

Similarly, the *average Jensen-Tsallis similarity* of frame i with respect to the rest of frames of shot s is defined as

$$AJT_{\alpha}^{xyz}(i) = \frac{1}{m-1} \sum_{j=1, j\neq i}^{m} \mathrm{JT}_{\alpha}^{xyz}(i,j). \tag{5.17}$$

Observe that AJT only compares the histograms of the frames and that the spatial distribution of the intensities is not taken into account. In this case, high values will be obtained when the histogram of a frame is similar to the histograms of the rest of the frames (see Figure 5.2).

Finally, a more global strategy is proposed to quantify the similarity between the histogram of frame i and the mean histogram of shot s. Thus, the *global Jensen-Tsallis*

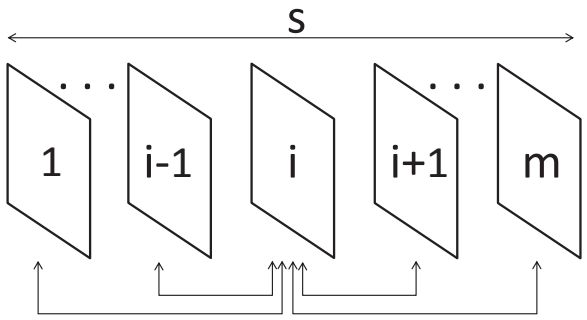


Figure 5.2: Computation scheme of the similarity between frame i and the rest of frames of shot s used by AI and AJT.

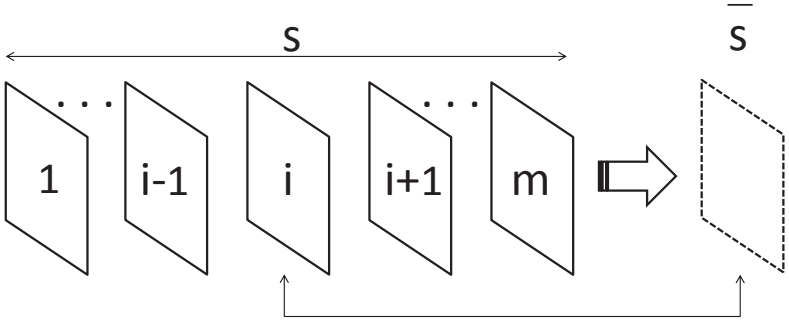


Figure 5.3: Computation scheme of the similarity between frame i and the virtual frame $\bar{s}$ used by GJT.

similarity (GJTS) for a frame i of shot s is defined as

$$GJT_{\alpha}^{xyz}(i) = \mathrm{JT}_{\alpha}^{xyz}(i,\bar{s}), \tag{5.18}$$

where $\bar{s}$ is interpreted as a virtual frame whose histogram is the average of the histograms of all frames of shot s. From this measure, the keyframe of shot s is given by the frame with maximum global similarity, that is, its histogram is the closest to the histogram that represents the whole shot (see Figure 5.3).

5.5 Results

In this section, we analyze the performance of the proposed Tsallis entropy-based measures, using different color spaces and histogram binnings, to deal with shot boundary detection and keyframe selection. These measures are compared with the mutual information-based measure proposed by Cernekova et al. [Cernekova 2006], which is a particular case of the measure IR_{α}^{xyz} (Equation (5.8)) when $\alpha = 1$ and the RGB color space with 256 bins is considered. The proposed measures are analyzed with two different databases. First, we use a training database, composed

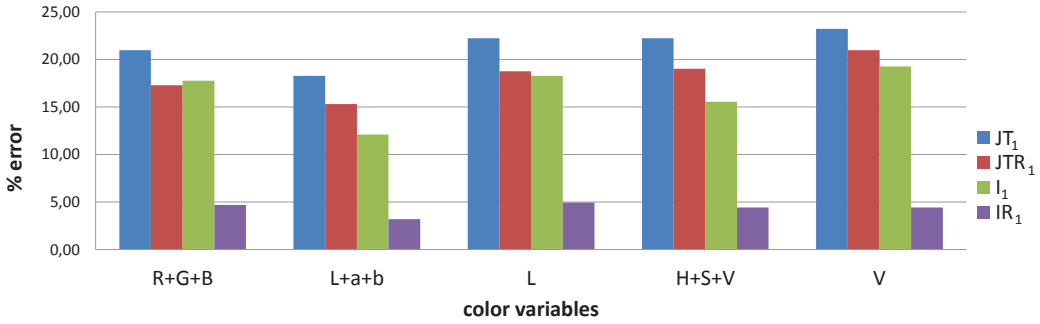


Figure 5.4: Error ratio percentage for the shot boundaries obtained by applying the measures JT_1, JTR_1, I_1, and IR_1, for different color variables.

of 27 videos from the Open Video Project [Open Video Project], to analyze in detail the performance of the proposed measures for several parameters, such as color space, regular binning, and entropic index. Second, we use a large testing database, provided by the TrecVid project [TRECVID], to evaluate the proposed measures with the optimal parameter configurations obtained with the training database. Our tests have been run on a PC with an Intel© Core™i5 430M 2.27GHz and 4 GB RAM.

5.5.1 Training database

In this section, we describe the experiments that have been performed with the training database to evaluate the behaviour of the proposed measures with different color spaces and histogram binnings. These experiments have been carried out over 27 videos with resolution of 320 × 240 pixels and a total duration of 30 minutes [Open Video Project]. Table 5.1 shows the list of videos used in our experiments, including both the number of frames and the number of shot boundaries. As the proposed measures have been devised to detect video cuts, we focus our attention on the detection of abrupt changes between shots and the fade sequences have been removed in the analyzed videos.

As we are interested in the discriminatory capacity of the proposed measures, we proceed in the following way. In this first experiment, for each video, the stopping criterion of our algorithms is given by the previously known number of shot boundaries. Thus, if the video has n shot boundaries, the frame similarity-based algorithm seeks the n lowest frame similarity values. In order to quantify the error ratio, the number of detected shot boundaries is compared with the manually defined one. The *error ratio* is defined by the ratio between the sum of the number of shot misdetections for all videos ($\#errors$) and the total number of shot boundaries also for all videos ($\#cuts$):

$$ER = \frac{\#errors}{\#cuts}. \tag{5.19}$$

First, we analyze the behaviour of I_α, IR_α, JT_α, and JTR_α with $\alpha = 1$ (i.e.,

Shannon entropy-based measures) using RGB, Lab, and HSV color spaces. Figure 5.4 shows the error ratio percentage for the shot boundaries obtained with the similarity measures I_1 and IR_1 based on mutual information I and JT_1 and JTR_1 based on Jensen-Shannon divergence JS, and using different color variables. For each color space, these measures are calculated with one color variable and three color variables. In the case of three variables, we assign one color component to each variable obtaining the combinations $R \oplus G \oplus B$, $L \oplus a \oplus b$, and $H \oplus S \oplus V$. In this experiment, we use 256 histogram bins to quantize each color component. For one variable, we use either the luminance variable (L in Lab) or the value variable (V in HSV), since these color components provide us the most basic information in an image and do not take into account the chromatic dimension. From Figure 5.4, we see that the mutual information-based measures achieve better results than the Jensen Shannon-based measures. In addition, we show how the use of ratios outperforms the accuracy of the results. We also see that the color space Lab, that is perceptually uniform, performs better than RGB and HSV. Note that, in general, the results obtained with only one variable (i.e., luminance or value) are worse than the ones obtained with three variables, where each color component is considered separately. Thus, we see that the chromaticity information plays an important role to detect the shot boundaries using the proposed measures.

To better illustrate the benefits of the normalization of mutual information and Jensen-Tsallis divergence, we have performed two experiments whose results are shown in Figures 5.5 and 5.6. Figure 5.5 shows the values of I_1, IW_1, and IR_1 between consecutive frames for the video *wth-02* using 256 bins, and $L \oplus a \oplus b$ color variables. This video is composed of only 3 shots where the first two ones correspond to standard video scenes while the frames of the last one are black images with random noise. In this video, the similarity measure I_1 between consecutive frames of the first two shots takes values around 5, while in the shot boundary between frames 110 and 111 the measure decreases to 1.25. In the third shot, the frames contain very little information and, thus, the shared information between two frames is also low (around 0.25). Thus, the mutual information I detects that these frames have low similarity and interprets them as shot boundaries. The informational frame similarity ratio IR overcomes this problem since it rates the similarity between two frames with the average similarity in its neighborhood (measure IW). In this case, the consecutive frames of the third shot have very low similarity but the average similarity in their neighborhood is also low and, thus, they are not detected as cuts. Figure 5.6 shows the similarity values JT_1, JTW_1, and JTR_1 between consecutive frames for the video *wth-02* using using 256 bins and $L \oplus a \oplus b$ color variables. As in the previous case, we can see that the measure JT obtains a small set of low values around the frame 287, when actually there is only one abrupt cut. This problem is also solved by the application of the normalized measure JTR. These two experiments illustrate the benefits of using the normalization of the information measures by their average on a window, as was previously proposed by Cernekova et al. [Cernekova 2006].

In Figure 5.7, we analyze the capability to detect the video shot boundaries

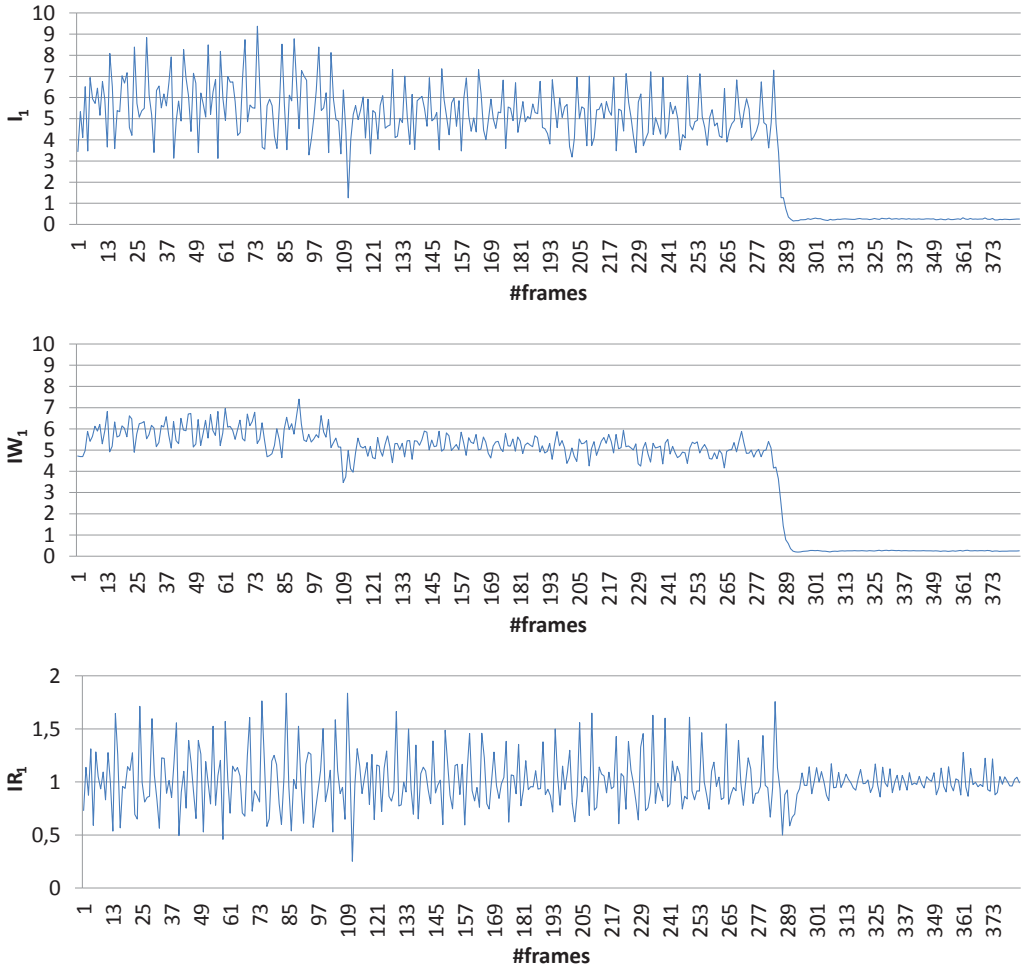


Figure 5.5: Similarity values I_1, IW_1, and IR_1 between consecutive frames for the video *wth-02* using 256 bins and $L \oplus a \oplus b$ color variables.

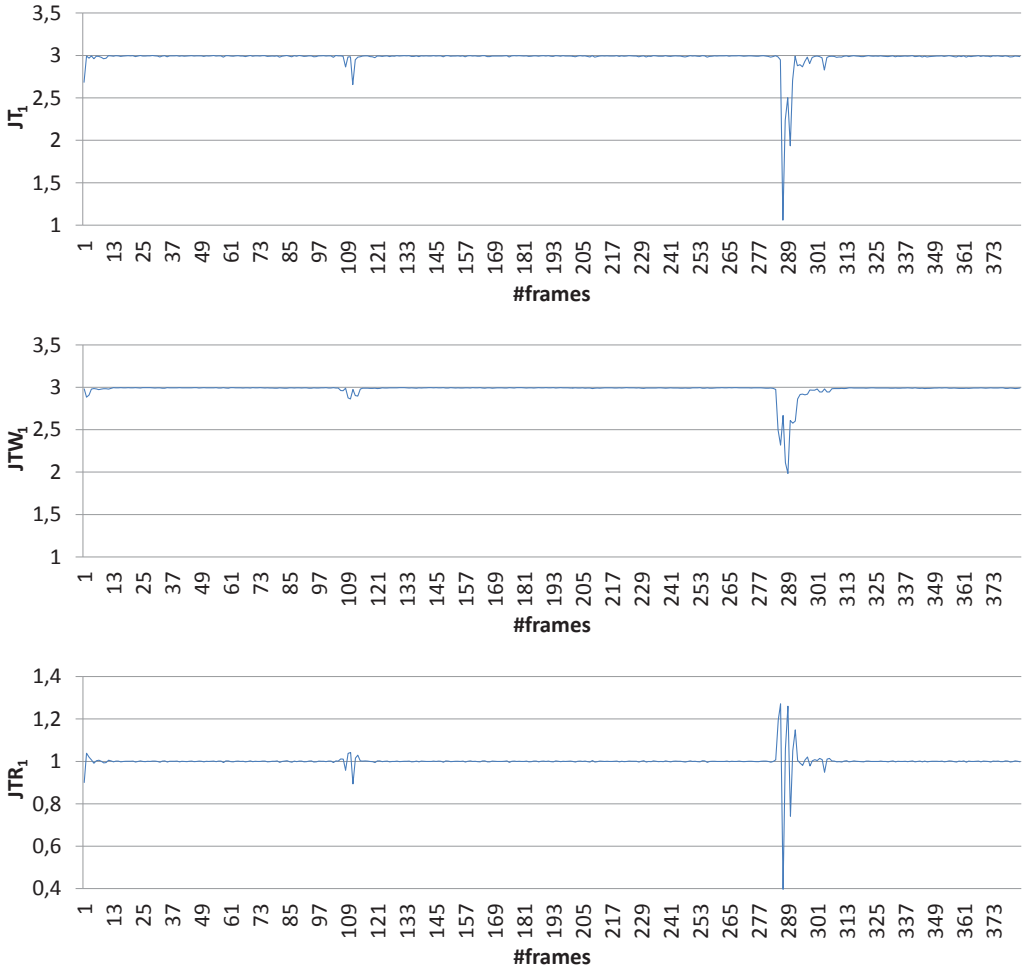
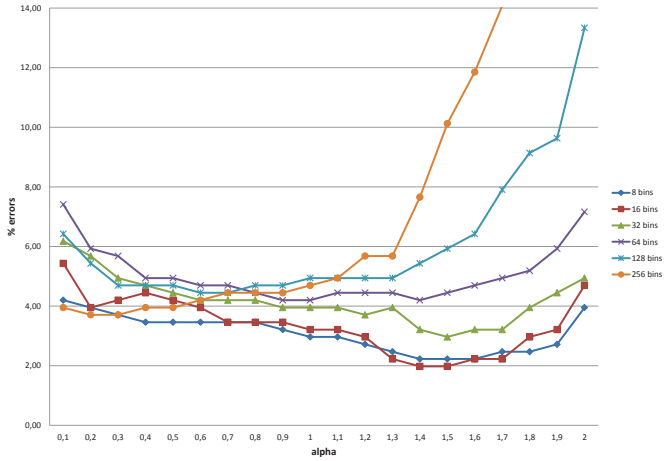
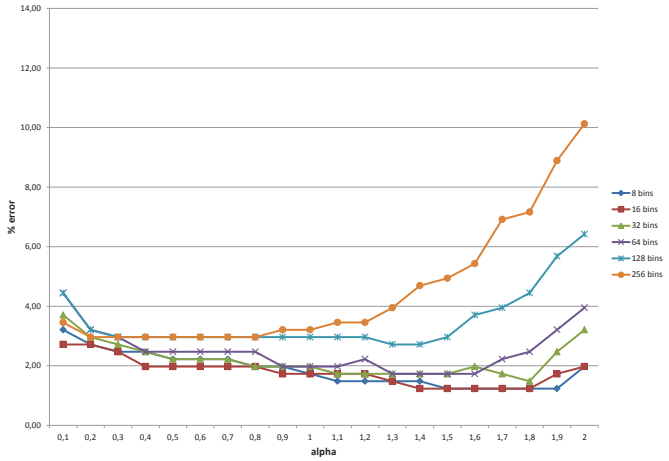


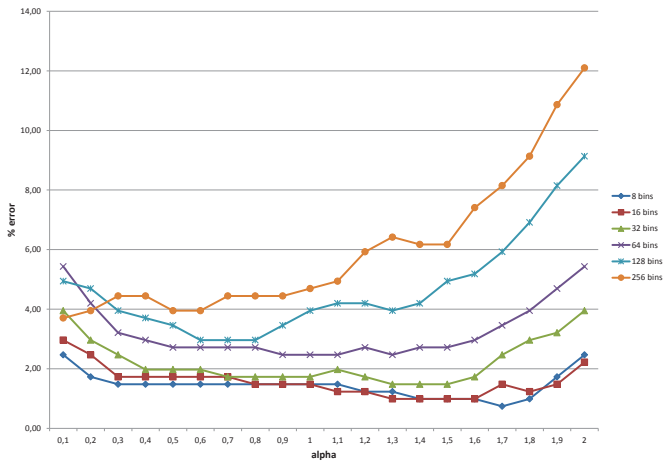
Figure 5.6: Similarity values JT_1, JTW_1, and JTR_1 between consecutive frames for the video *wth-02* using 256 bins and $L \oplus a \oplus b$ color variable.



(a) $R \oplus G \oplus B$



(b) $L \oplus a \oplus b$



(c) $H \oplus S \oplus V$

Figure 5.7: Error ratio percentage for the shot boundaries obtained with IR_α measure computed for different entropic indices in the range $[0.1, 2]$ and (a) $R \oplus G \oplus B$, (b) $L \oplus a \oplus b$ and (c) $H \oplus S \oplus V$ color variables using 8, 16, 32, 64, 128, and 256 histogram bins.

using the measure IR_α for different entropic indices, color spaces, and number of histogram bins. $R \oplus G \oplus B$, $L \oplus a \oplus b$, and $H \oplus S \oplus V$ color variables are analyzed for entropic indices ranging from 0.1 to 2 and considering a different number of histogram bins (8, 16, 32, 64, 128, and 256). It can be seen that, $L \oplus a \oplus b$ and $H \oplus S \oplus V$ color variables present a better performance than $R \oplus G \oplus B$ and, as in the first experiment presented in Figure 5.4, this fact confirms the hypothesis that the use of perceptually aimed color spaces improve the accuracy of the results. Observe also that the results are very sensitive to the number of histogram bins, obtaining the best results with only 8 bins (i.e., only considering 8 different values for each color component in the IR computation). Note also that the optimal entropic index depends on the number of bins. For instance, with the $L \oplus a \oplus b$ color variables and 256 bins, the optimal index is around 0.5, while, with the same color variables and 8 bins, the optimal one is around 1.7. In general, the lower the number of bins the higher the optimal entropic index. The best result is achieved with $H \oplus S \oplus V$ color variables, 8 bins, and an entropic index of 1.7, that is six times better than the one achieved with IR_1^{RGB} and 256 bins, as proposed by Cernekova et al. [Cernekova 2006]. Observe that the error is reduced from 4.69% to 0.74%.

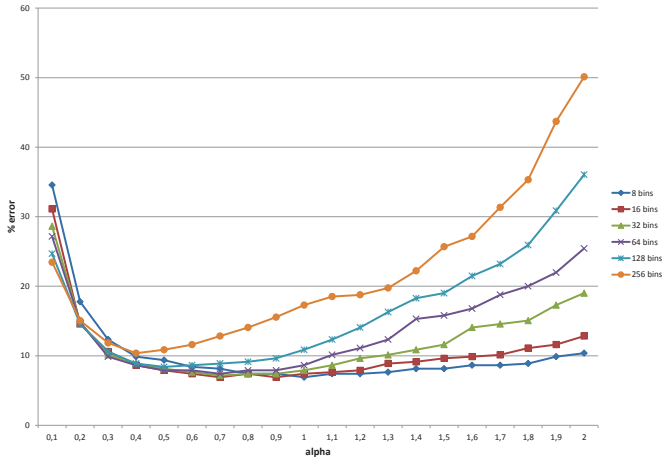
In Figure 5.8, we also analyze the capability to detect the video shot boundaries using the measure JTR for the same entropic indices, color spaces, and number of histogram bins than in the previous experiment. Note that $L \oplus a \oplus b$ and $H \oplus S \oplus V$ color variables also perform better than $R \oplus G \oplus B$ variables. In this experiment, in general the results improve with a low number of histogram bins, but the optimal result is achieved with 128 bins. With respect to the optimal entropic index, JTR has a behavior different than IR, since the best results are obtained in the range from 0.4 to 1.2. As in the first experiment (see Figure 5.4), IR is, in general, more accurate than JTR. However, it is also interesting to observe that $\mathrm{JTR}_{0.5}^{Lab}$ achieves slightly more precise results than IR_1^{RGB}. Note that JTR only deals with the frame histograms while IR is based on the intensities of each pair of matching pixels.

In a real scenario, the number of cuts is not a priori known. In this case, a simple strategy is to fix a threshold value so that a shot boundary is detected when the similarity measure between two frames is lower than the threshold value. This strategy is applied in the next experiment. To evaluate the results, the measures *precision* and *recall* are used. The precision is defined as

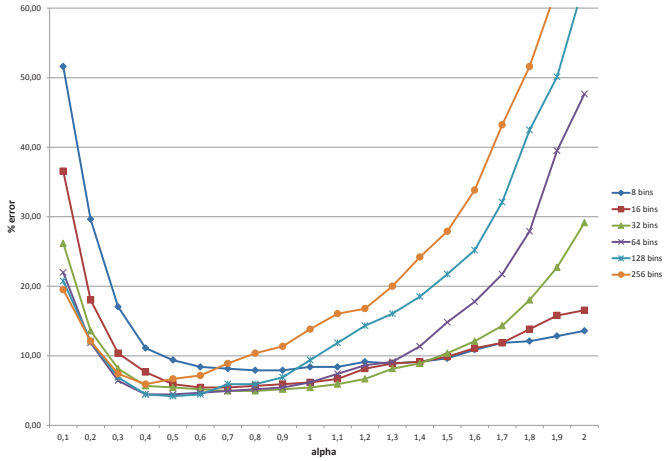
$$P = \frac{TP}{TP + FP}, \tag{5.20}$$

where TP is the number of true positives (i.e., the shot boundaries that the algorithm detects and that correspond to the real ones according to the ground truth) and FP is the number of false positives (i.e., the shot boundaries that the algorithm detects and that do not correspond to the real ones according to the ground truth). The recall is defined as

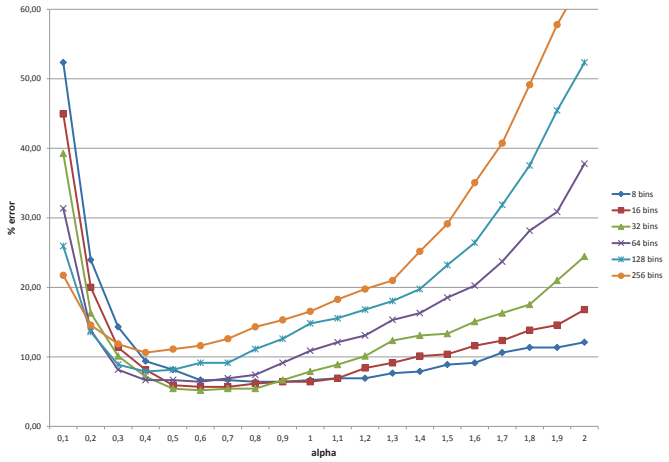
$$R = \frac{TP}{TP + FN}, \tag{5.21}$$



(a) $R \oplus G \oplus B$



(b) $L \oplus a \oplus b$



(c) $H \oplus S \oplus V$

Figure 5.8: Error ratio percentage for the shot boundaries obtained by applying JTR measure computed for different entropic indices in the range $[0.1, 2]$ and (a) $R \oplus G \oplus B$, (b) $L \oplus a \oplus b$, and (c) $H \oplus S \oplus V$ color variables using 8, 16, 32, 64, 128, and 256 histogram bins.

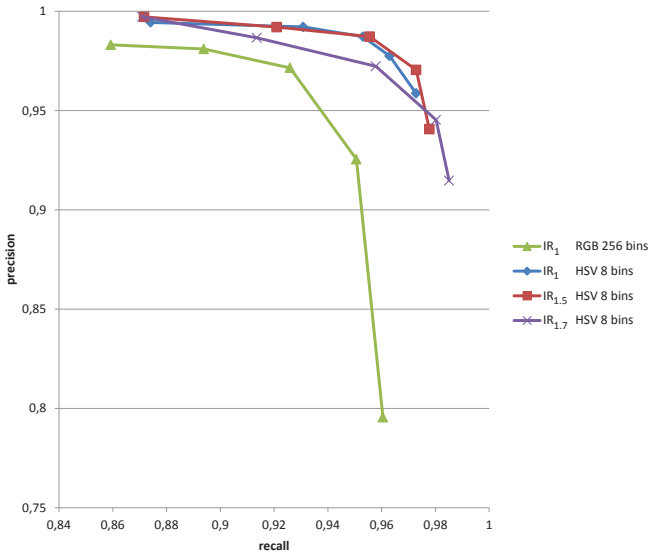
where FN is the number of false negatives (i.e., the shot boundaries that the algorithm does not detect and that correspond to the real ones according to the ground truth). The measures precision and recall take values in the range $[0,1]$, being 1 the best value. In Figure 5.9, we show the precision and recall values for the IR-based measures (Figure 5.9(a)) and the JTR-based measures (Figure 5.9(b)) with different threshold values. In Figure 5.9(a), precision and recall for IR_1^{RGB} with 256 bins are represented using threshold values in the range $[0.35, 0.45]$ with steps of 0.025, and the precision and recall for IR_1^{HSV}, $IR_{1.5}^{HSV}$, and $IR_{1.7}^{HSV}$ with 8 bins are represented using threshold values in the range $[0.15, 0.25]$ with steps of 0.025. Observe that, when the threshold value is increased, the precision decreases while the recall increases, and vice versa. As in the previous experiments, the results obtained with the $H \oplus S \oplus V$ color variables and 8 histogram bins clearly outperform the ones obtained with $R \oplus G \oplus B$ and 256 bins. Although the differences are small, the best results are now obtained with $\alpha = 1$ and $\alpha = 1.5$, while in the previous experiments the best performance was achieved with $\alpha = 1.7$.

Figure 5.9(b) shows the results obtained with JTR_1^{RGB} with 256 bins using threshold values in the range $[0.92, 0.96]$ with steps of 0.01, and the ones obtained with JTR_1^{Lab} with 32 bins and $\mathrm{JTR}_{0.5}^{Lab}$ with 128 bins using threshold values in the range $[0.95, 0.99]$ with steps of 0.01. As in the previous case, the measure applied with the RGB color space obtains worse results than with the Lab color space. In this experiment, the use of Tsallis entropy does not improve the results obtained with Shannon entropy (i.e., $\alpha = 1$). In particular, JTR_1^{Lab} with 32 bins obtain similar results to $\mathrm{JTR}_{0.5}^{Lab}$ with 128 bins.

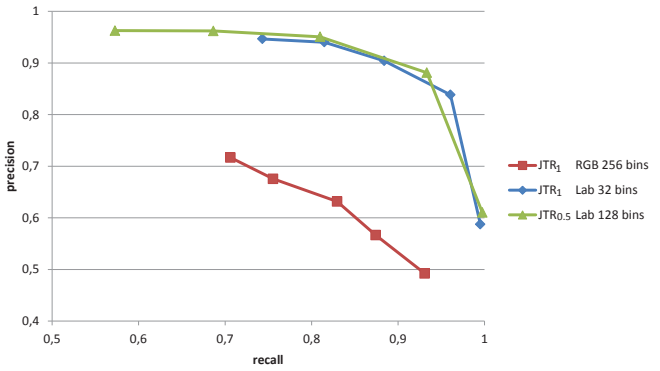
Finally, Figure 5.10 presents the results of selecting the keyframes for the video *UGS07_007* which has 12 shots. Results are shown for the three alternative measures, AI_α, AJT_α, and GJT_α. For AI_α we have used $\alpha = 1.7$, $H \oplus S \oplus V$, and 8 bins, and for AJT_α and GJT_α we have used $\alpha = 0.5$ and $L \oplus a \oplus b$, and 128 bins. Observe that these parameters have been chosen according to the ones that achieve the optimal results in shot boundary detection. We can observe small differences between the selected frames depending on the used measure. Due to the similarity of the results, we recommend to use the method based on GJT_α due to its lower computational cost. In this case, each frame has to be only compared with the mean histogram, while the measures AI_α and AJT_α require the comparison with all the other frames in the shot.

5.5.2 Testing database

In this section, we use a testing database to compare the proposed measures in a real environment. This large database is provided by the TrecVid project [TRECVID] and it contains 17 videos with resolution of 352×288 pixels and a total duration of 7 hours and 29 minutes (see Table 5.2, where the number of frames and the number of shot boundaries for each video are also shown). These videos were used to test several methods in the shot boundary detection task in TrecVid 2007. The results obtained by these methods are available in [TRECVID 2007]. The shot boundary



(a) *IR*-based measures



(b) JTR-based measures

Figure 5.9: Precision and recall values for (a) the *IR*-based measures and (b) the JTR-based measures with different threshold values.

Figure 5.10: The most representative keyframes for the video *UGS07_007* have been obtained using the $H \oplus S \oplus V$ color variables, $\alpha = 1.7$, 8 histogram bins, and the measures (first row) $AI_{1.7}^{HSV}$ with 8 bins, (second row) $AJT_{0.5}^{Lab}$ with 128 bins and, (third row) $GJT_{0.5}^{Lab}$ with 128 bins.

Filename	#Frames	#Cuts
BG_2408	35892	101
BG_11362	16416	104
BG_34901	34389	224
BG_35187	29025	135
BG_36182	29610	95
BG_36537	50004	259
BG_37359	28908	164
BG_37822	21960	119
BG_38150	52650	215

Filename	#Frames	#Cuts
BG_9401	50049	89
BG_14213	83115	106
BG_35050	36999	98
BG_36028	44991	87
BG_36506	15210	77
BG_36628	56564	192
BG_37417	23004	76
BG_37879	29019	95

Table 5.2: List of 17 videos (with filename, number of frames, and number of shot boundaries) used in our experiments. Obtained from the TrecVid project [TRECVID].

positions are given by the ground truth provided by the TrecVid project together with the video database. As with the training database, the gradual transitions have not been considered in our experiments.

Table 5.3 summarizes the results obtained with the proposed measures applied to the testing database. In this experiment, we have only analyzed the Shannon-based measures IR_1^{RGB} and JTR_1^{RGB} and the measures that have achieved the best results with the training database. A shot boundary is detected when a pair of frames has a similarity value lower than a given threshold. The obtained results are compared with the ground truth.

In addition to the previously defined measures precision and recall, we also use the harmonic mean of both measures, also called *F-measure*, as a single value that summarizes both precision and recall. The F-measure is defined as

$$F = \frac{2}{\frac{1}{P} + \frac{1}{R}}. \tag{5.22}$$

As it can be seen in Table 5.3, the best results are obtained with $IR_{1.7}^{HSV}$ and 8 bins using a threshold value of 0.2. Note also that the use of the Tsallis generalization notably improves the results (F-measure varies from 0.9384 in the second row to 0.9495 in the forth row), and also the reduction of histogram bins and the use of the HSV color space have a great impact in the results (F-measure varies from 0.8558 in the first row to 0.9384 in the second row). If the results are analyzed in a detailed way, we observe that approximately the half of the errors (50 of 104) comes from a single video, BG_36628, and are due to a scene with black frames corrupted by noise. As these frames contain very little information, the proposed measure does not perform properly because the normalization makes it very sensitive to the presence of noise. This problem could be easily solved by, for instance, do not taking into account frames with low information (i.e., low entropy value), similarly to the strategy proposed by Cernekova et al. [Cernekova 2006] to detect fades. The

Measure	Color variables	#Bins	Th	Precision	Recall	F
IR_1	$R \oplus G \oplus B$	256	0.4	0.7811	0.9463	0.8558
IR_1	$H \oplus S \oplus V$	8	0.2	0.9542	0.9231	0.9384
$IR_{1.5}$	$H \oplus S \oplus V$	8	0.2	0.9565	0.9352	0.9457
$IR_{1.7}$	$H \oplus S \oplus V$	8	0.2	0.9531	0.9459	0.9495
JTR_1	$R \oplus G \oplus B$	256	0.94	0.8664	0.8936	0.8798
JTR_1	$L \oplus a \oplus b$	32	0.98	0.7793	0.9696	0.8641
$\mathsf{JTR}_{0.5}$	$L \oplus a \oplus b$	128	0.98	0.8460	0.9584	0.8987

Table 5.3: Results obtained using the testing database and a threshold (Th) value as stopping criterion.

Measure	Color variables	#Bins	F-measure
IR_1	$R \oplus G \oplus B$	256	0.9186
IR_1	$H \oplus S \oplus V$	8	0.9495
$IR_{1.5}$	$H \oplus S \oplus V$	8	0.9562
$IR_{1.7}$	$H \oplus S \oplus V$	8	0.9548
JTR_1	$R \oplus G \oplus B$	256	0.8855
JTR_1	$L \oplus a \oplus b$	32	0.9052
$\mathsf{JTR}_{0.5}$	$L \oplus a \oplus b$	128	0.9047

Table 5.4: Results obtained using the testing database and the number of cuts as stopping criterion.

JTR-based measures obtain in general worse results than the IR-based ones, but note that $\mathsf{JTR}_{0.5}^{Lab}$ with 128 histogram bins obtain clearly better results than the measures IR_1^{RGB} and JTR_1^{RGB}).

Table 5.4 shows the results obtained with the same measures than the previous experiment when the number of cuts is a priori known and used as a stopping criterion. These results can be seen as the best possible achievable results when the best particular threshold is selected for each video. Note that in this case, precision and recall take the same value, since FP is equal to FN and, thus, the F-measure (i.e., the harmonic mean) also takes this value. The general behaviour of the measures is similar to the previous experiment, obtaining the best results when the perceptual color spaces and the generalized entropies are used. There are some minor differences, such as the better behaviour of $IR_{1.5}^{HSV}$ with respect to the one of $IR_{1.7}^{HSV}$. From this fact, it can be established that $IR_{1.5}^{HSV}$ has more capacity to obtain good results, while $IR_{1.7}^{HSV}$ is a more stable measure to be used with a single threshold value for all the videos.

Another important aspect to be considered is the computation time. Figure 5.11 plots the F-measure versus the computation time per frame in milliseconds spent on detecting the shot boundaries for the measures IR_1^{RGB} with 256 bins, IR_1^{HSV} with 8 bins, $IR_{1.7}^{HSV}$ with 8 bins, JTR_1^{RGB} with 256 bins, JTR_1^{Lab} with 32 bins,

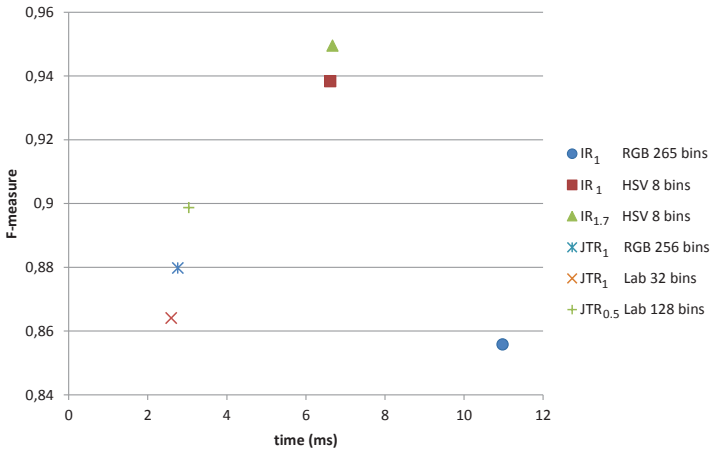


Figure 5.11: F-measure and computation time per frame in milliseconds, for the measures analyzed in Table 5.3.

and $\mathrm{JTR}_{0.5}^{Lab}$ with 128 bins. As we could expect, the computational time is very sensitive to the number of bins. On the other hand, the difference between the Shannon entropy-based measures and their extension based on Tsallis entropy is not significant (especially, for a low number of bins). It is also important to notice that the computational cost of the measures based on Jensen-Tsallis divergence are between two times and four times lower than the measures based on mutual information, since in the first case the computation of the joint histogram between frames is not required.

As a conclusion, we have seen that IR with $\alpha = 1.7$ and perceptually aimed color spaces achieves the best performance, significantly improving the performance of the mutual information-based measure proposed by Cernekova et al. We have also seen that JTR, with $\alpha = 0.5$, 128 bins, and $L \oplus a \oplus b$ color variables, obtains a good tradeoff between accuracy and computational cost.

5.6 Conclusions

In this chapter we have presented and analyzed the behaviour of two different information-theoretic approaches based on Tsallis mutual information and Jensen-Tsallis divergence to deal with video shot boundary detection and keyframe selection. Their discriminatory capacity has been analyzed for several color spaces (RGB, HSV, and Lab), regular binnings, and entropic indices. The performance of these approaches has been successfully compared with mutual information and Jensen-Shannon divergence presented by Cernekova et al. [Cernekova 2006] and Xu et al. [Xu 2010], respectively. Different experiments have shown that the optimal detection capacity is obtained by the Tsallis mutual information similarity measure using HSV and Lab color spaces. In general, the reduction of the number of

histogram bins also improves the obtained results. Finally, we have also proposed three different measures to select the most representative keyframe of each shot. One of these measures is based on Tsallis mutual information whereas the other two are based on Jensen–Tsallis divergence.

CHAPTER 6

Image informativeness

6.1 Introduction

The most basic information measure, the Shannon entropy, has been used to quantify the information content or uncertainty of a random variable. Given a color space (e.g., CIELab), the information associated with the lightness of the whole image can be computed from the entropy of the lightness histogram. The same procedure can be applied to any color component of a color space. The main drawback of histogram entropy, called image entropy, is the fact that it does not take into account the spatial distribution of pixels. This means that, for instance, after a simple swapping of pixels, the image entropy would give the same result. In other words, Shannon entropy is not an adequate measure to characterize, for instance, the structure of an image.

In this chapter, we present a set of information measures that capture several aspects of image information, from a local perspective, taking into account the vicinity of pixels, to an evolutionary perspective, based on the difficulty of extracting or discovering the image information. Thus, we focus our attention not on how the image information varies when distortion is applied but on the quantification of the information of a single image. This fact enables us to evaluate the image quality and also to provide a set of image features that could be used to deal with several image processing problems such as image classification and optimization of image acquisition parameters.

Thus, we analyze four information-theoretic measures, three of them (entropy rate, excess entropy, and erasure entropy) consider the image as a stationary stochastic process, while the fourth (partitional information) is based on an information channel between image regions and histogram bins. Experimental results, applied to natural and synthetic images, analyze the performance of these measures to characterize several informativeness aspects of an image. We also analyze their behavior under some image effects such as blurring, contrast change, and noise.

The content of this chapter has been published in "*Analysis of image informativeness measures*", Marius Vila, Anton Bardera, Miquel Feixas, Philippe Bekaert, Mateu Sbert. IEEE International Conference on Image Processing pages 1086-1090, October 2014 [Vila 2014].

6.2 Previous work

In this section we review some previous work related to the concepts used in this chapter.

Measurement of image quality is crucial for many image processing algorithms. Traditionally, image quality assessment algorithms predict visual quality by comparing a distorted image against a reference image, typically by modeling the Human Visual System (HVS), or by using arbitrary signal fidelity criteria. Sheikh and Bovik [Sheikh 2006] proposed an information fidelity criterion that quantifies the Shannon information that is shared between the reference and the distorted images relative to the information contained in the reference image itself. Wang and Li [Wang 2011] tested the hypothesis that when viewing natural images, the optimal perceptual weights for pooling should be proportional to local information content, which can be estimated in units of bit using advanced statistical models of natural images.

Rigau et al. [Rigau 2008a] presented a set of information-theoretic measures to study some informational aspects of a painting related to its palette and composition. Some of these measures, based on the entropy of the palette, the compressibility of the image, and an information channel to capture the composition of a painting, were used to discriminate different painting styles and to analyze the evolution of van Gogh's artwork [Rigau 2008b], revealing a significant correlation between the values of the measures and van Gogh's artistic periods. Rigau et al. [Rigau 2010] also investigated whether informational measures can support the claim of art critics on his evolution of palette and composition. They also studied how far van Gogh's last period was from his other periods, and tried to trace his artistic development. To this end, they employed informational measures together with a set of measures, such as entropy rate and excess entropy, that take into account spatial information.

Bardera et al. [Bardera 2009b] introduced a split-and-merge algorithm based on the definition of an information channel between a set of regions (input) of the image and the intensity histogram bins (output). From this channel, the maximization of the mutual information gain is used to optimize the image partitioning. Then, the merging process of the regions obtained in the previous phase is carried out by minimizing the loss of mutual information. Bardera et al. [Bardera 2009a] also presented an information-theoretic approach for thresholding-based segmentation that uses the excess entropy to measure the structural information of a 2D or 3D image and to locate the optimal thresholds. This approach is based on the conjecture that the optimal thresholding corresponds to the segmentation with maximum structure, i.e., maximum excess entropy.

6.3 Image information measures

In this chapter, an image is considered as a random variable B taking intensity bin values b from a finite set $\mathscr{B}$. Each value $b \in \mathscr{B}$ represents a bin of intensity values that can be either composed by a single intensity value or by a set of similar intensity values depending on the used quantization. The probability distribution of the random variable B is given by $p(b) = Pr[B = b]$. Thus, according to Equation (2.2), the *Shannon entropy* $H(B)$ of the random variable B is defined by

$$H(B) = -\sum_{b \in \mathscr{B}} p(b) \log p(b). \tag{6.1}$$

The term $-\log p(b)$ represents the information content associated with the intensity bin b. Thus, the entropy gives us the average amount of information of the intensity values in the image. From Equation (6.1), we can see that the entropy depends only on the probabilities of the intensity values, but not on their spatial distribution. Therefore, as we have mentioned in Section 6.1, two perceptually different images can give the same image entropy value.

In order to consider the spatial structure in the image information computation, two different approaches are presented. First, an image is modeled as a stationary stochastic process to quantify the image information from the vicinity of pixels and, second, an information channel between image regions and histogram bins is used to study the difficulty of extracting the information of an image.

6.3.1 Stationary stochastic process-based measures

This first approach models an image as a stationary stochastic process $\{B_i\}$, which is an indexed sequence of random variables characterized by the joint probability distribution $p(b_1, b_2, \ldots, b_n) = \Pr\{(B_1, B_2, \ldots, B_n) = (b_1, b_2, \ldots, b_n)\}$ with $(b_1, b_2, \ldots, b_n) \in \mathscr{B}^n$ for $n \geq 1$ [Cover 1991, Yeung 2008]. In our case, the sequence of states will be determined by consecutive positions in the image, considering, thus, the spatial information.

6.3.1.1 Entropy rate

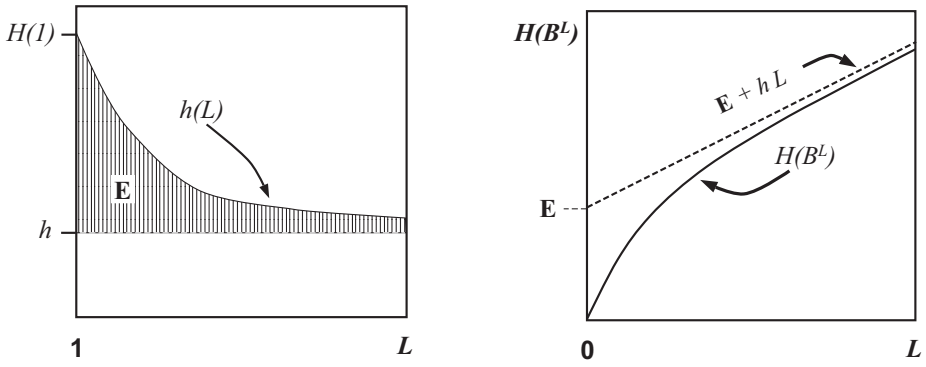
As we have seen in Section 2.2.4, following the notation used in the work of Feldman and Crutchfield [Feldman 2003], the *entropy rate* can be rewritten as

$$h = \lim_{L \to \infty} \frac{H(B^L)}{L} = \lim_{L \to \infty} h(L), \tag{6.2}$$

where

$$h(L) = H(B^L) - H(B^{L-1}) \tag{6.3}$$

$$= H(B_L | B_1, \ldots, B_{L-1}) \tag{6.4}$$



(a) (b)

Figure 6.1: Two different graphical representations of the excess entropy measure, corresponding to Equations (6.5) and (6.6), respectively. Images obtained from [Feldman 2003].

is the entropy of a symbol conditioned on a block of $L-1$ adjacent symbols. The entropy rate of a sequence quantifies the average amount of information per symbol b and the optimal achievement for any possible compression algorithm [Cover 1991]. The entropy rate is always equal or lower than the Shannon entropy and is only equal when there is no correlation between consecutive symbols. Entropy rate has been applied to image processing by Rigau et al. [Rigau 2010].

6.3.1.2 Excess entropy

A complementary measure to the entropy rate is the *excess entropy*, which is a measure of the *structure* of a system. Structure here is taken to be a statement which expresses the degree of correlation between the components of a system. The *excess entropy* is defined by

$$E \equiv \sum_{L=1}^{\infty} (h(L) - h) \tag{6.5}$$

and captures how $h(L)$ converges to its asymptotic value h. Figure 6.1(a) is a graphical representations of the excess entropy measure, which is represented by the shaded area, corresponding to the sum of differences between $h(L)$ and the limit h. If Equation (6.3) is inserted into Equation (6.5), the sum telescopes and an alternate expression for the excess entropy [Feldman 2003] is obtained:

$$E = \lim_{L \to \infty} [H(B^L) - h \cdot L]. \tag{6.6}$$

Hence, excess entropy is the y-intercept of the straight line to which $H(B^L)$ asymptotes as indicated in Figure 6.1(b).

It is important to note that, when we take into account only a few number

of symbols in the entropy computation, the system appears more random than it actually is. This excess randomness measures how much additional information must be gained about the configurations in order to reveal the actual uncertainty h [Feldman 2002]. Excess entropy is commonly used and well understood in one dimension, but some difficulties are found in its extension to higher dimensions. Excess entropy has been introduced to image processing by Bardera et al. [Bardera 2009a]. Excess entropy, which provides us with a measure of the regularities presenting in an image, can also be interpreted as the degree of predictability of a pixel given its neighbours. In order to compute the excess entropy, two main considerations have to be taken into account:

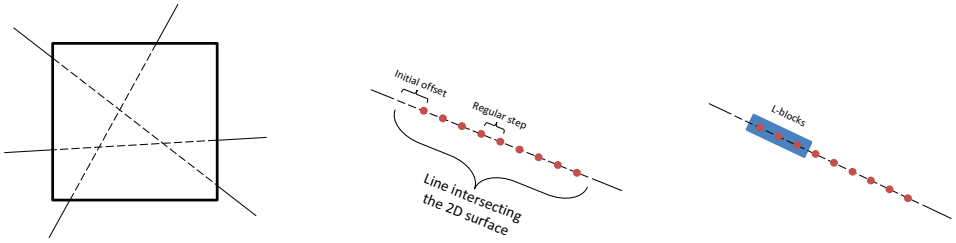
- The first is the definition of the neighbourhood concept for a pixel. While neighbourhood is unique and unambiguous in 1D, its extension to 2D introduces ambiguity, since a sequence of L-block neighbour pixels can be selected in different manners [Feldman 2003].

- The second is the computation of L-block entropies when $L \to \infty$. In practice, L-block entropies for high L are not computable, since the number of elements of the joint histogram (required to compute joint probabilities $p(b^L)$) is given by B^L, where B is the cardinality of the system. Note that in our case, B is the number of clusters or bins of the segmented image histogram, i.e., the number of colors of the image. Thus, a tradeoff between the accuracy of the measure, given by L, and the number of clusters $|\mathscr{B}|$ is required.

To overcome the neighbourhood problem, uniformly distributed random lines, also called global lines [Sbert 1993] are used. Global lines sample the 2D-surface stochastically in the sense of integral geometry, i.e., invariant to translations and rotations [Santaló 1976]. These lines are generated from the walls of a convex bounding box containing the surface [Castro 1998]. This can be done taking a random point on the surface of the convex bounding box and a cosinus distributed random direction as it is illustrated in Figure 6.2(a). The sequence of intensity values (L-block X^L) needed to estimate the joint probabilities is captured at evenly spaced positions over the global lines from an initial random offset, that ranges from 0 to the step size (see Figure 6.2(b)). Points chosen on each line provide us with the intensities to calculate the L-block entropies, required to compute the excess entropy (see Figure 6.2(c)). In this manner, the 2D-neighbourhood problem is reduced to 1D, where the concept of neighbourhood is well defined. In our implementation, N is taken as an input parameter of the algorithm, while L is determined from N such that the computation of the joint histogram is attainable.

6.3.1.3 Erasure entropy

As we have seen in Section 2.2.4, entropy rate of a collection of random variables can be interpreted as the uncertainty associated with a given symbol if all the preceding symbols are known. However, what if we condition on both the past and the future?



(a) (b) (c)

Figure 6.2: (a) Global lines are cast from the walls of the bounding box, (b) intensity values are captured at evenly spaced positions over the global lines from an initial random offset, and (c) neighbour intensity values are taken in L-blocks.

This idea is carried out by the erasure entropy [Verdú 2006] that can be seen as the uncertainty associated with a given symbol if all the preceding and succeeding symbols are known.

Similar to Equation (2.20), the erasure entropy of a collection of L random variables $B_1, \ldots, B_L$ is given by

$$H^-(B_1, \ldots, B_L) = \sum_{i=1}^{L} H(B_i | B_{/i}), \tag{6.7}$$

where $B_{/i} = \{B_j, j = 1, \ldots L, j \neq i\}$.

When entropy rate is applied to a data indexed by multidimensional sets, such as images, it requires an artificial definition of the preceding symbols (past), while erasure entropy does not suffer from that drawback.

For any collection of discrete random variables $\{B_1, \ldots B_L\}$, $H^-(B_1, \ldots B_L) \leq H(B_1, \ldots B_L)$ with equality if and only if $\{B_1, \ldots B_L\}$ are independent. Thus, a collection of random variables has zero erasure entropy if it has zero entropy, but the converse is not true. For example, if $B_1 = B_2$ then $H(B_1, B_2) = H(B_1)$ whereas $H^-(B_1, B_2) = 0$.

Analogously to the entropy rate (see Section 2.2.4), erasure entropy rate quantifies how the entropy of a sequence of L random variables increases with L. The erasure entropy rate h^- of a stochastic process $\{B_i\}$ is defined by

$$h^- = \lim_{L \to \infty} \frac{1}{L} H^-(B_1, B_2, \ldots, B_L) \tag{6.8}$$

when the limit exists. Thus, erasure entropy rate can be defined as the limit of the arithmetic mean of the conditional entropies of each symbol given all preceding and succeeding symbols. Erasure entropy rate represents the average information content per symbol in a stochastic process. For a stationary stochastic process, the erasure entropy rate exists and is equal to

$$h^{-} = \lim_{L\to\infty} h^{-}(0), \tag{6.9}$$

where $h^{-}(0) = H(B_0|B_{-L}^{-1}, B_1^L)$, B_{-L}^{-1} symbolizes the previous samples (past) and B_1^L the posterior samples (future).

6.3.2 Information channel-based measure

This second approach introduces spatial information into the information measure by considering an information channel $R \to B$ between the random variables R (input) and B (output), which represent, respectively, the set of regions $\mathscr{R}$ of an image and the set of intensity bins $\mathscr{B}$. This channel is defined by a conditional probability matrix $p(B|R)$ which expresses how the pixels corresponding to each region of the image are distributed into the histogram bins. Thus, each row r of the $p(B|R)$ matrix corresponds to the normalized histogram of the region r. The input distribution $p(R)$, which represents the probability of selecting each image region, is defined by $p(r) = \frac{n(r)}{N}$ (i.e. the relative area of region r).

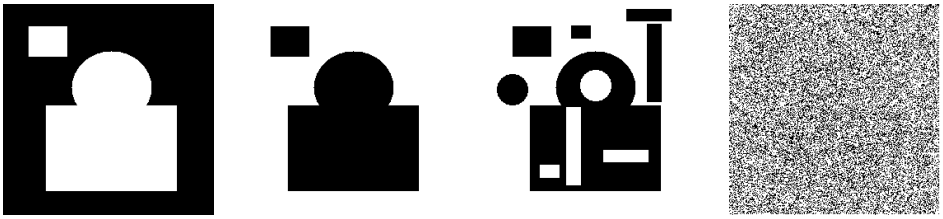
6.3.2.1 Partitional information

As it was proposed by Rigau et al. [Rigau 2004] and Bardera et al. [Bardera 2009b], the information bottleneck method presented in Section 2.2.6 can be applied to this information channel. Following a top-down partition procedure, a greedy algorithm based on a binary space partitioning (BSP) can be used to find a partition that maximizes the mutual information of the channel. The BSP partitioning algorithm can be represented by an evolving binary tree where each leaf corresponds to a terminal region of the image. At each partitioning step, the tree gains information from the original image such that each internal node k contains the information I_k gained with its corresponding splitting, which is given by

$$\begin{aligned} \delta I_{\tilde{r}} &= I(R,B) - I(\widetilde{R},B) \\ &= p(\tilde{r}) JS\left(\pi_1, \pi_2; p(B|r_1), p(B|r_2)\right), \end{aligned} \tag{6.10}$$

where $\pi_1 = \frac{p(r_1)}{p(\tilde{r})}$ and $\pi_2 = \frac{p(r_2)}{p(\tilde{r})}$. Note that Equation (6.10) is a particular case of Equation (2.31). As we have seen in Section 2.2.3.3 the JS-divergence $JS\left(\pi_i, \pi_j; p(B|r_1), p(B|r_2)\right)$ between two regions can be interpreted as a measure of *dissimilarity* between them respect to the intensity values. At a given moment, $I(R,B)$ can be obtained adding up the information available at the internal nodes of the tree weighted by $p(k)$, where $p(k) = \frac{n(k)}{N}$ is the relative area of the region associated with node k and $n(k)$ is the number of pixels of this region. Thus, the mutual information of the channel is given by

$$I(R,B) = \sum_{k=1}^{T} p(k) I_k, \tag{6.11}$$



(a)	(b)	(c)	(d)
H : 0.9427	H : 0.9427	H : 0.9427	H : 0.9427
h : 0.0932	h : 0.0932	h : 0.1830	h : 0.9215
E : 0.8551	E : 0.8551	E : 0.7657	E : 0.0003
h^- : 0.0253	h^- : 0.0253	h^- : 0.0591	h^- : 0.9212
PI : 0.9580	PI : 0.9580	PI : 0.9337	PI : 0.7669

Figure 6.3: Synthetic images and their entropy (H), entropy rate (h), excess entropy (E), erasure entropy (h^-), and partitional information (PI) values (a-d).

where T is the number of internal nodes. It is important to stress that the best partition can be decided locally. If this procedure is performed until the number of regions reach the number of pixels, the $I(R,B)$ curve leads to the entropy of the image $H(B)$. The integral of this curve can be seen as a measure of difficulty of describing the spatial distribution of the intensities. Thus, a new information measure, which we call *partitional information*, can be defined as

$$PI = \frac{\sum_{k=1}^{N} I(R_k, B)}{N \cdot H(B)}, \tag{6.12}$$

where R_k represents the set of regions of the image after k partitions. This measure takes values in $[0,1]$, leading to high values when the image has a simple structure and low values when the image is complex.

6.4 Results

In this section, we analyze the information content of a group of synthetic and real images using the image information measures presented in Section 6.3. All images used in our experiments have been converted to grayscale in order to obtain an unique intensity value between 0 and 256 for each pixel. Thus, probabilities of H have been computed using 256 intensity bins. To convert RGB values to grayscale values, we have used the CIE 1931 transformation $Y = 0.2621R + 0.7152G + 0.0722B$, where R, G, and B are, respectively, the values of red, green, and blue channels, and Y is the luminance obtained. To compute h, E, and h^-, the intensity values of an image are captured at evenly spaced positions over uniformly distributed random lines, also called global lines (see [Bardera 2009a] for more details). To compute h and E we have used Equations (6.2) and (6.6), respectively, taking the neighbour intensity values in L-blocks of size 3. h^- has been computed

using Equation (6.7) with L-blocks of size 1.

First, we use four synthetic images with pixel resolution of 256×256 to illustrate the behavior of the measures (see Fig. 6.3). All images have been created to have the same entropy value. The two first images 6.3(a-b) represent the same scene with the colors interchanged. In this case, h, E, h^{-}, and PI values are equal since these measures are not dependent on the colors themselves, but only on their probability and spatial distribution. In the third image 6.3(c), some shapes are moved with respect to the original image 6.3(b). The last image 6.3(d) has been generated by randomly swapping 200000 pixels of image 6.3(b). Thus, both images 6.3(c-d) alter the structure with respect to the original image 6.3(b) but keep the same probability for each color. Observe that h and h^{-} increase their value with the increase of uncertainty and variability in the image. This fact complicates the prediction of the value of a pixel from the spatial distribution and, therefore, the value of these measures increase. On the contrary, E and PI decrease with the decrease of the spatial structure. Since Fig. 6.3(d) has no spatial structure, the knowledge of the spatial distribution does not improve the capability of predicting a pixel value and, therefore, h and h^{-} are very close to the H value.

Second, we use a group of real images belonging to the Categorical Image Quality (CSIQ) Database [Larson 2010] developed at Oklahoma State University. It consists of 30 original images with pixel resolution of 512×512 an corresponding to five different subjects: animals, landscapes, people, plants, and urban environments. Each image is distorted using six types of distortions at five different levels, obtaining a total of 930 images. The distortion types used in CSIQ are JPEG and JPEG2000 compression, global contrast decrements, additive Gaussian pink noise, additive Gaussian white noise, and Gaussian blurring. Furthermore, with the aim of analyzing the behavior of all measures at different image resolutions, we have reduced the original images to pixel resolutions of 256×256, 128×128, and 64×64. Fig. 6.4 shows the absolute values of the measures of four images belonging to the CSIQ Database. Each row corresponds to a measure and is sorted from the lowest to the highest value, except PI, is sorted from the highest to the lowest value. Observe that the first image (column 1) obtains in all the cases the minimum value of the corresponding measure except with PI. This behaviour is mainly due to the low information content (entropy H) of this image. With respect to measure PI, the value is maximum due to the fact that this image can be more easily partitioned. A similar behaviour can be observed with the second image (column 2) with respect to third (column 3) and fourth (column 4) images. Observe also that the last two images, although having a conceptually different content, have similar and alternating values in the five measures considered. Our measures deal with intensity histogram and spatial distribution of pixel values, but do not consider the semantic content of the images.

Figs. 6.5(a-f) show several plots resulting from the application of additive Gaussian white noise, Gaussian blurring, global contrast decrements, additive Gaussian pink noise, JPEG, and JPEG2000 compression distortion, respectively. Each plot shows the mean values of each measure at five levels of distortion with respect

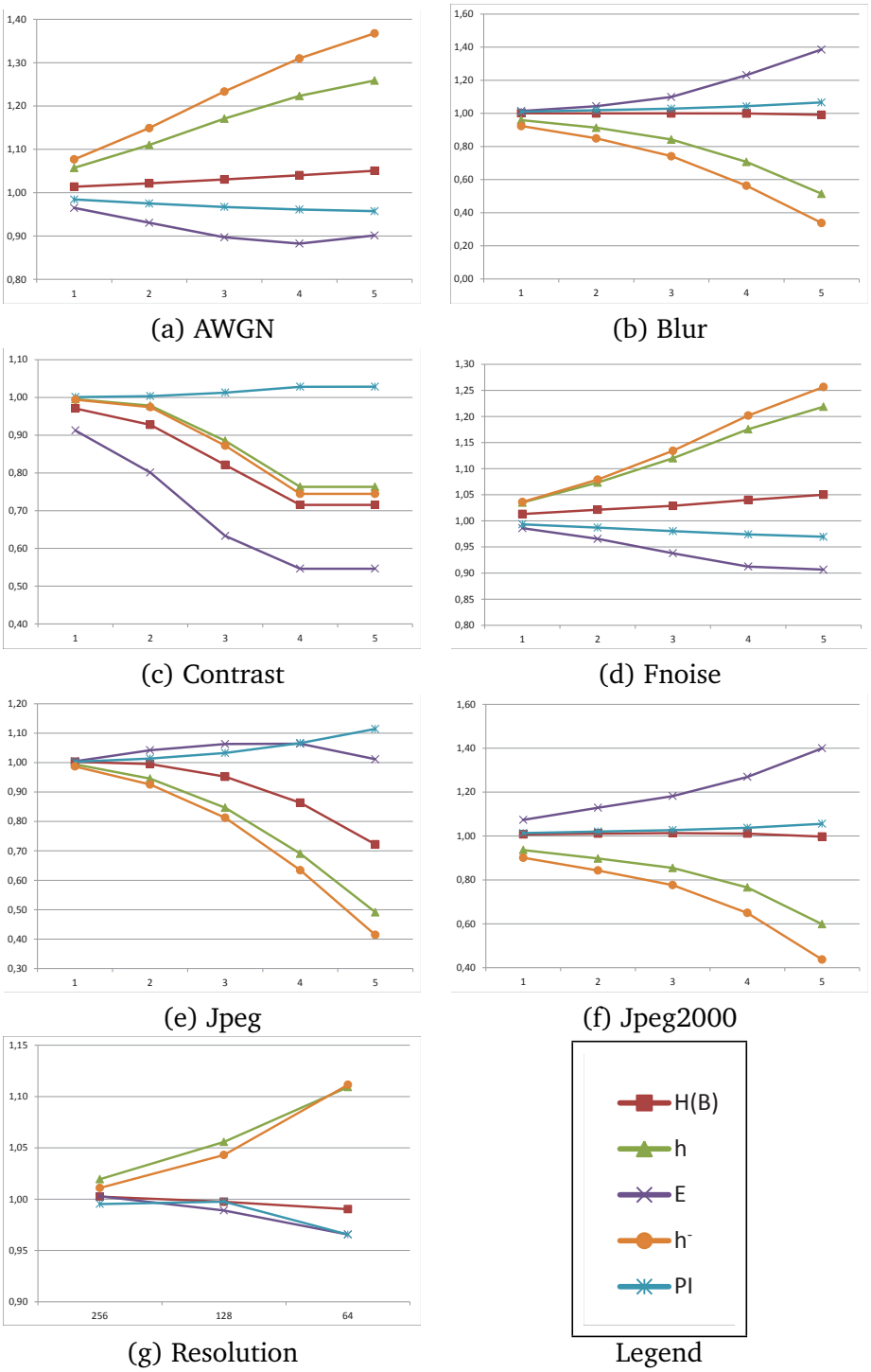
to the value in the original image. These six types of image effects can be divided into two main groups. On the one hand, at each new level of distortion, the additive Gaussian pink and white noise increase the variability of the image while the spatial structure decreases. On the other hand, the rest of image effects have a completely opposite behavior. In the first group, H, h, h^{-} increase with the distortion level, while E and PI decrease. In the second group, the behavior is complementary. The only exception corresponds to the E value with the global contrast reduction (see Fig. 6.5(c)). In this case, the decrease of E value is due to the fact that the reduction of contrast keeps the spatial structure but also produces a global loss of information. Observe that these results are consistent with the ones obtained by using the synthetic images.

Finally, Fig. 6.5(g) shows how the image resolution affects the value of the different measures with respect to the value in the original image. We can observe that H, PI, and E are nearly invariant to image resolution, while the other measures increase when the resolution is reduced. This reduction in the image resolution can be seen as a distortion that increases the variability of the image. Entropy rate and erasure entropy have obtained high values for images with high variability and low spatial structure, while excess entropy and partitional information show a complementary behavior. We can observe that excess entropy and partitional information are less sensitive to image resolution than the rest of measures.

6.5 Conclusions

In this chapter, we have presented two different approaches to quantify the information content of an image taking into account the spatial distribution of pixels. In the first approach, entropy rate, excess entropy, and erasure entropy have been used to quantify the image information from the vicinity of pixels and, in the second approach, an information channel between image regions and histogram bins has been applied to study the difficulty of extracting the information of an image. The measures have been applied to several synthetic and natural images, analyzing, in the latter case, the behavior of the measures under several types of image effects.

Figure 6.4: Values of the measures (Shannon entropy H, entropy rate h, excess entropy E, erasure entropy h^{-}, and partitional information PI) of four natural images. Each row corresponds to a measure and is sorted from the lowest to the highest value, except the partitional information that is sorted from the highest to the lowest value.



(a) AWGN (b) Blur

(c) Contrast (d) Fnoise

(e) Jpeg (f) Jpeg2000

(g) Resolution Legend

Figure 6.5: (a-f) Mean values of each measure for five levels of distortion with respect to the original image using Gaussian white noise, Gaussian blurring, global contrast decrements, additive Gaussian pink noise, JPEG, and JPEG2000 compression, respectively. (g) Mean values of each measure for three different image resolutions.

CHAPTER 7

Conclusions and future work

Contents

7.1 Contributions

The main goal of this thesis was to find good metrics based on information theory with the aim of developing robust similarity measures for multimedia data classification and retrieval. This objective has been achieved with the following contributions:

- We have analyzed the behavior of different similarity measures applied to invoice classification. Three types of measures, applied to document processing, have been presented and tested, and are based respectively on the intensity differences (sum of squared differences, sum of absolute differences, and correlation coefficient), the shared information (mutual information and normalized mutual information), and the normalized compression distance between two images, calculated from both image (PNG, JPEG, and JPEG2000) and file (GZIP and BZIP2) compressors. The experiments have been carried out on two testing databases and a real-world database. In both cases, low resolution images have been used to show the best performance of the mutual information-based measures, although an acceptable performance has also been obtained with the correlation coefficient and the normalized compression distance implemented using file and image compressors. We have demonstrated the suitability of several global similarity measures for invoice image classification.

- We have introduced three different mutual information generalizations for invoice classification. These measures have been inspired respectively by Kullback–Leibler distance, the difference between entropy and conditional entropy, and the Jensen–Shannon divergence, and their ratio with the Tsallis joint entropy. The experiments have been carried out on a testing database and

a real-world database, both with and without the use of a threshold, showing different behaviour depending on the measure and the entropic index. When the threshold is used, Tsallis-based measures obtain the best results for α values between 1.0 and 1.4 whereas, when the threshold is not used, the best results are obtained for $\alpha = 1.0$, i.e, when Shannon-based measures are applied. In both cases, low resolution images have been used to show the good performance of the mutual information-based measures. Finally, the document registration using measures based on mutual information generalizations 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.

- We have presented and analyzed the behaviour of two different information-theoretic approaches based on Tsallis mutual information and Jensen-Tsallis divergence to deal with video shot boundary detection and keyframe selection. Their discriminatory capacity has been analyzed for several color spaces (RGB, HSV, and Lab), regular binnings, and entropic indices. The performance of these approaches has been successfully compared with mutual information and Jensen-Shannon divergence presented by Cernekova et al. [Cernekova 2006] and Xu et al. [Xu 2010], respectively. Different experiments have shown that the best detection capacity is obtained by the Tsallis mutual information similarity measure using HSV and Lab color spaces. In general, the reduction of the number of histogram bins also improves the obtained results. We have also proposed three different measures to select the most representative keyframe of each shot. One of these measures is based on Tsallis mutual information and the other two are based on Jensen–Tsallis divergence. Due to the similarity of the results, we recommend to use the method based on global Jensen-Tsallis similarity due to its lower computational cost, as each frame has to be only compared with the mean histogram, while the rest of measures require the comparison with all the other frames in the shot.

- We have presented two different approaches to quantify the information content of an image taking into account the spatial distribution of pixels. In the first approach, entropy rate, excess entropy, and erasure entropy have been used to quantify the image information from the vicinity of pixels and, in the second approach, an information channel between image regions and histogram bins has been applied to study the difficulty of extracting the information of an image. The measures have been applied to several synthetic and natural images, analyzing, in the latter case, the behavior of the measures under several types of image effects. Results show that entropy rate and erasure entropy obtain high values for images with high variability and low spatial structure, while excess entropy and partitional information show a complementary behavior.

7.2 Future work

The ideas presented in this thesis can be expanded in different directions.

- In document classification,
  - we will study the performance of the similarity measures presented in Chapters 3 and 4 for other typologies of documents, such as scientific papers or journal pages.
  - we will analyze the behaviour of several Rényi-based generalizations of mutual information to calculate the similarity between scanned documents.
  - we will investigate different learning systems, such as support vector machine and neural networks, for document classification.
  - we will investigate whether the proposed measures for the video shot detection can be adapted to the classification of document images.
- In video processing,
  - we will more deeply investigate the performance of the proposed measures in Chapter 5 for video shot detection. In particular, we will investigate the performance of Tsallis mutual information and Jensen-Tsallis divergence for several image resolutions.
  - we will investigate how and why different measures have a different behavior on shot detection, for videos with different characterizations.
  - we will analyze the use of Tsallis entropy-based measures to deal with gradual transitions in the context of shot boundary detection.
  - we will investigate whether the proposed measures for the classification of invoice images can be adapted to video shot detection.
  - with respect to the color spaces, we will analyze the relationship between the correlation of the color components and their discriminatory capacity.
  - we will investigate new strategies for keyframe selection taking into account the variability within a shot and the similarities between keyframes of different shots and to compare the automatic keyframe selection with a manual selection.
  - all parameters are determined heuristically. Some of them could be more robustly obtained, by creating links between them and other parameters used.
- In image informativeness,
  - we will investigate the performance of the measures proposed in Chapter 6 to automatically adjust different camera effects such as focus, contrast or sharpness.

  - we will apply the presented measures as descriptors for document and image classification.

7.3 Publications

Publications that support the contents of this thesis:

- "*Tsallis Mutual Information for Document Classification*", Marius Vila, Anton Bardera, Miquel Feixas, Mateu Sbert. Entropy, vol. 13, no. 9, pages 1694-1707, 2011.
- "*Tsallis entropy-based information measure for shot boundary detection and keyframe selection*", Marius Vila, Anton Bardera, Qing Xu, Miquel Feixas, Mateu Sbert. Signal, Image and Video Processing, vol. 7, no. 3, pages 507-520, 2013.
- "*Analysis of image informativeness measures*", Marius Vila, Anton Bardera, Miquel Feixas, Philippe Bekaert, Mateu Sbert. IEEE International Conference on Image Processing pages 1086-1090, October 2014.
- "*Image-based Similarity Measures for Invoice Classification*", Marius Vila, Anton Bardera, Miquel Feixas, Mateu Sbert. Submitted.

Bibliography

[Alippi 2005] Cesare Alippi, F. Pessina and Manuel Roveri. *An adaptive system for automatic invoice-documents classification*. In International Conference on Image Processing, pages 526–529, 2005. (Cited on pages 19 and 22.)

[Appiani 2001] Enrico Appiani, Francesca Cesarini, Anna Maria Colla, Michelangelo Diligenti, Marco Gori, Simone Marinai and Giovanni Soda. *Automatic document classification and indexing in high-volume applications*. International Journal on Document Analysis and Recognition, vol. 4, no. 2, pages 69–83, 2001. (Cited on pages 18, 22 and 28.)

[Arai 1997] Hiroyuki Arai and Kazumi Odaka. *Form processing based on background region analysis*. In International Conference on Document Analysis and Recognition, pages 164–169, 1997. (Cited on page 22.)

[Bagdanov 2001] Andrew D. Bagdanov and Marcel Worring. *Fine-grained document genre classification using first order random graphs*. In Sixth International Conference on Document Analysis and Recognition, pages 79–83, 2001. (Cited on page 18.)

[Bagdanov 2003] Andrew D. Bagdanov and Marcel Worring. *First order Gaussian graphs for efficient structure classification*. Pattern Recognition, vol. 36, pages 1311–1324, 2003. (Cited on page 18.)

[Baldi 2003] Stefano Baldi, Simone Marinai and Giovanni Soda. *Using treegrammars for training set expansion in page classification*. In Seventh International Conference on Document Analysis and Recognition, pages 829–833, 2003. (Cited on page 18.)

[Bardera 2004] Anton Bardera, Miquel Feixas and Imma Boada. *Normalized similarity measures for medical image registration*. In Proceedings of Medical Imaging SPIE 2004, volume 5370, pages 108–118, February 2004. (Cited on pages 45 and 47.)

[Bardera 2009a] Anton Bardera, Imma Boada, Miquel Feixas and Mateu Sbert. *Image segmentation using excess entropy*. Journal of Signal Processing Systems, vol. 54, no. 1-3, pages 205–214, 2009. (Cited on pages 80, 83 and 86.)

[Bardera 2009b] Anton Bardera, Jaume Rigau, Imma Boada, Miquel Feixas and Mateu Sbert. *Image segmentation using information bottleneck method*. Image Processing, IEEE Transactions on, vol. 18, no. 7, pages 1601–1612, July 2009. (Cited on pages 80 and 85.)

[Bardera 2010] Anton Bardera, Miquel Feixas, Imma Boada and Mateu Sbert. *Image registration by compression*. Information Sciences, vol. 180, no. 7, pages 1121–1133, 2010. (Cited on page 27.)

[Baumann 1997] Stephan Baumann, Majdi Ben Hadj Ali, Andreas Dengel, Thorsten Jäger, Michael Malburg, Achim Weigel and Claudia Wenzel. *Message extraction from printed documents - a complete solution*. In Fourth International Conference on Document Analysis and Recognition, volume 2, pages 1055–1059, aug 1997. (Cited on page 18.)

[Behera 2005] Ardhendu Behera, Denis Lalanne and Rolf Ingold. *Combining color and layout features for the identification of low-resolution documents*. International Journal of Signal Processing, vol. 2, no. 1, pages 7–14, 2005. (Cited on page 22.)

[Browne 1999] Paul Browne, Alan F. Smeaton, Noel Murphy, Noel O'Connor, Sen Marlow and Catherine Berrut. *Evaluating and combining digital video shot boundary detection algorithms*. In Irish Machine Vision and Image Processing Conference, 1999. (Cited on page 20.)

[Bunke 2000] Horst Bunke. *Recent developments in graph matching*. In Proceedings of 15th International Conference on Pattern Recognition, 2000, volume 2, pages 117–124, 2000. (Cited on page 19.)

[Burbea 1982] Jacob Burbea and C. Radhakrishna Rao. *On the convexity of some divergence measures based on entropy functions*. IEEE Transactions on Information Theory, vol. 28, no. 3, pages 489–495, May 1982. (Cited on pages 12, 42, 58 and 59.)

[Butz 2001] Torsten Butz and Jean-Philippe Thiran. *Shot boundary detection with mutual information*. In International Conference on Image Processing, pages 422–425, 2001. (Cited on pages 20 and 58.)

[Byun 2000] Yungcheol Byun and Yillbyung Lee. *Form classification using DP matching*. In ACM Symposium on Applied Computing, volume 1, pages 1–4, 2000. (Cited on page 19.)

[Capek 2001] Martin Capek, Lukas Mroz and Rainer Wegenkittl. *Robust and fast medical registration of 3D-multi-modality data sets*. In Proceedings of the Internatinal Federation for Medical and Biological Engineering, volume 1, pages 515–518, June 2001. (Cited on page 49.)

[Castro 1998] Francesc Castro and Mateu Sbert. *Application of quasi-Monte Carlo sampling to the multipath method for radiosity*. In Proceedings of 3rd International Conference on Monte Carlo and Quasi-Monte Carlo methods in Scientific Computing, Claremont (CA), USA, June 1998. (Cited on page 83.)

[Cernekova 2006] Zuzana Cernekova, Ioannis Pitas and Christophoros Nikou. *Information theory-based shot cut/fade detection and video summarization*. IEEE Transactions on Circuits and Systems for Video Technology, vol. 16, no. 1, pages 82–91, 2006. (Cited on pages 20, 58, 59, 64, 66, 70, 74, 76 and 92.)

[Cesarini 1998] Francesca Cesarini, Marco Gori, Simone Marinai and Giovanni Soda. *INFORMys: a flexible invoice-like form-reader system*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 7, pages 730–745, 1998. (Cited on page 22.)

[Chen 2007] Nawei Chen and Dorothea Blostein. *A survey of document image classification: problem statement, classifier architecture and performance evaluation*. International Journal on Document Analysis and Recognition, vol. 10, no. 1, pages 1–16, 2007. (Cited on pages 18 and 21.)

[Cilibrasi 2005] Rudi Cilibrasi and Paul Vitányi. *Clustering by compression*. IEEE Transactions on Information Theory, vol. 51, no. 4, pages 1523–1545, April 2005. (Cited on pages 26 and 27.)

[Ciocca 2006] Gianluigi Ciocca and Raimondo Schettini. *An innovative algorithm for key frame extraction in video summarization.* J. Real-Time Image Processing, vol. 1, no. 1, pages 69–88, 2006. (Cited on page 20.)

[Collignon 1995] André Collignon, Dirk Vandermeulen, Paul Suetens and Guy Marchal. *Automated multi-modality image registration based on information theory*. Computational Imaging and Vision, vol. 3, pages 263–274, 1995. (Cited on page 44.)

[Cotsaces 2006] Costas Cotsaces, Nikos Nikolaidis and Ioannis Pitas. *Video shot detection and condensed representation. A review*. IEEE Signal Processing Magazine, vol. 23, no. 2, pages 28–37, April 2006. (Cited on page 20.)

[Cover 1991] Thomas M. Cover and Joy A. Thomas. Elements of information theory. Wiley-Interscience, 1991. (Cited on pages 6, 8, 9, 11, 12, 13, 14, 24, 42, 81 and 82.)

[Crutchfield 2003] James P. Crutchfield and David P. Feldman. *Regularities unseen, randomness observed: the entropy convergence hierarchy*. Chaos, vol. 15, pages 25–54, 2003. (Cited on page 13.)

[Csiszár 2004] Imre Csiszár and Paul C Shields. *Information Theory and Statistics: A Tutorial*. Foundations and Trends in Communications and Information Theory, vol. 1, no. 4, pages 417–528, 2004. (Cited on page 8.)

[Dhawale 2008] Chitra A. Dhawale and Sanjeev Jain. *Motion compensated video shot detection using multiple feature experts*. ICGST International Journal on

Graphics, Vision and Image Processing, vol. 08, pages 1–11, 2008. (Cited on page 20.)

[Diligenti 2003] Michelangelo Diligenti, Paolo Frasconi and Marco Gori. *Hidden tree Markov models for document image classification*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, pages 519–523, 2003. (Cited on pages 18 and 19.)

[Duda 2001] Richard O. Duda, Peter E. Hart and David G. Stork. Pattern classification. Wiley-Interscience, 2 edition, November 2001. (Cited on page 19.)

[Duygulu 2002] Pinar Duygulu and Volkan Atalay. *A hierarchical representation of form documents for identification and retrieval*. International Journal on Document Analysis and Recognition, vol. 5, no. 1, pages 17–27, 2002. (Cited on pages 18 and 22.)

[Eglin 2003] Véronique Eglin and Stéphane Bres. *Document page similarity based on layout visual saliency: application to query by example and document classification*. In Seventh International Conference on Document Analysis and Recognition, pages 1208–1212, aug. 2003. (Cited on page 18.)

[Esposito 2000] Floriana Esposito, Donato Malerba and Francesca A. Lisi. *Machine learning for intelligent processing of printed documents*. Journal of Intelligent Information Systems, vol. 14, pages 175–198, 2000. (Cited on pages 18 and 19.)

[Fan 2010] Huijie Fan, Linlin Zhu and Yandong Tang. *Skew detection in document images based on rectangular active contour*. International Journal on Document Analysis and Recognition, vol. 13, no. 4, pages 261–269, 2010. (Cited on page 18.)

[Feldman 1998] David P. Feldman and James P. Crutchfield. *Discovering non-critical organization: statistical mechanical, information theoretic, and computational views of patterns in simple one-dimensional spin systems*. Journal of Statistical Physics, page submitted, 1998. (Cited on page 13.)

[Feldman 2002] David Feldman. *A brief introduction to information theory, excess entropy and computational mechanics*. Department of Physics, University of California, July, 2002. (Cited on pages 13 and 83.)

[Feldman 2003] David P. Feldman and James P. Crutchfield. *Structural information in two-dimensional patterns: entropy convergence and excess entropy*. Physical Review E, vol. 67, no. 051104, 2003. (Cited on pages 81, 82 and 83.)

[Fränti 2000] Pasi Fränti, Alexey Mednonogov, Ville Kyrki and Heikki Kälviäinen. *Content-based matching of line-drawing images using the Hough transform*.

International Journal on Document Analysis and Recognition, vol. 3, pages 117–124, 2000. (Cited on page 18.)

[Furuichi 2006] Shigeru Furuichi. *Information theoretical properties of Tsallis entropies*. Journal of Mathematical Physics, vol. 47, no. 2, 2006. (Cited on pages 46 and 47.)

[Gargi 2000] Ullas Gargi, Rangachar Kasturi and Strayer H. Strayer. *Performance characterization of video-shot-change detection methods*. IEEE Transactions on Circuits and Systems for Video Technology, vol. 10, no. 1, pages 1 –13, feb 2000. (Cited on pages 20 and 58.)

[Gatos 1997] Basilios Gatos, Nikos Papamarkos and Christodoulos Chamzas. *Skew detection and text line position determination in digitized documents*. Pattern Recognition, vol. 30, no. 9, pages 1505–1519, 1997. (Cited on pages 28 and 30.)

[Gonzalez 2002] Rafael C. Gonzalez and Richard E. Woods. Digital image processing. Prentice Hall, Upper Saddle River (NJ), USA, 2002. (Cited on page 60.)

[Grana 2007] Costantino Grana and Rita Cucchiara. *Linear transition detection as a unified shot detection approach*. IEEE Transactions on Circuits and Systems for Video Technology, vol. 17, no. 4, pages 483 –489, april 2007. (Cited on page 20.)

[Günsel 1998] Bilge Günsel and A. Murat Tekalp. *Content-based video abstraction*. In International Conference on Image Processing, pages 128 –132 vol.3, oct 1998. (Cited on page 20.)

[Gupta 2007] Mithun Das Gupta and Prateek Sarkar. *A shared parts model for document image recognition*. In International Conference on Document Analysis and Recognition, pages 1163–1172, 2007. (Cited on pages 18 and 22.)

[Hajnal 2001] Joseph V. Hajnal, David J. Hawkes and Derek L. G. Hill. Medical image registration. CRC Press Inc., 2001. (Cited on pages 19, 23, 24 and 41.)

[Hamza 2006] Abdessamad Ben Hamza. *Nonextensive information-theoretic measure for image edge detection*. Journal of Electronic Imaging, vol. 15, no. 1, pages 13011.1—13011.8, 2006. (Cited on page 47.)

[Hamza 2008] Hatem Hamza, Yolande Belaïd, Abdel Belaïd and Bidyut Baran Chaudhuri. *An end-to-end administrative document analysis system*. In Document Analysis Systems, pages 175–182, 2008. (Cited on pages 21 and 22.)

[Hanjalic 2002] Alan Hanjalic. *Shot-boundary detection: unraveled and resolved?* IEEE Transactions on Circuits and Systems for Video Technology, vol. 12, no. 2, pages 90–105, 2002. (Cited on page 20.)

[Haralick 1994] Robert M. Haralick. *Document image understanding: geometric and logical layout*. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 385–390, jun 1994. (Cited on page 18.)

[Harvda 1967] Jan Harvda and František Charvát. *Quantification method of classification processes. Concept of structural α-entropy*. Kybernetika, vol. 3, pages 30–35, 1967. (Cited on page 15.)

[Héroux 1998] Pierre Héroux, Sébastien Diana, Arnaud Ribert and Éric Trupin. *Classification method study for automatic form class identification*. In Fourteenth International Conference on Pattern Recognition, volume 1, pages 926–928, aug 1998. (Cited on page 19.)

[Hesseler 2006] Wolfgang Hesseler and Stefan Eickeler. *MPEG-2 compressed-domain algorithms for video analysis*. EURASIP Journal on Applied Signal Processing, vol. 2006, pages 186–186, January 2006. (Cited on page 20.)

[Hill 2001] Derek L. G. Hill, Philipp G. Batchelor, Mark Holden and David J. Hawkes. *Medical image registration*. Physics in Medicine and Biology, vol. 46, pages R1–R45, 2001. (Cited on page 23.)

[Ho 2001] Tin Kam Ho. Multiple classifier combination: lessons and next steps, volume 47 of *Series in Machine Perception and Artificial Intelligence*, chapter 7, pages 171–198. World Scientific, 2001. (Cited on page 19.)

[Hu 1999a] Jianying Hu, Ramanujan S. Kashi and Gordon T. Wilfong. *Document classification using layout analysis*. In DEXA Workshop, pages 556–560, 1999. (Cited on page 19.)

[Hu 1999b] Jianying Hu, Ramanujan S. Kashi and Gordon T. Wilfong. *Document image layout comparison and classification*. In International Conference on Document Analysis and Recognition, pages 285–288, 1999. (Cited on pages 17, 18, 22 and 41.)

[Huan 2008] Zhao Huan, Li Xiuhuan and Yu Lilei. *Shot boundary detection based on mutual information and canny edge detector*. In Proceedings of the 2008 International Conference on Computer Science and Software Engineering - Volume 02, pages 1124–1128, 2008. (Cited on page 58.)

[Jain 2000] Anil K. Jain, Robert P. W. Duin and Jianchang Mao. *Statistical pattern recognition: a review*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pages 4–37, jan 2000. (Cited on page 19.)

[Khader 2010] Mohammed Khader, Abdessamad Ben Hamza and Prabir Bhattacharya. *Multimodality image alignment using information-theoretic approach*. In Interntational Conference on Image Analysis and Recognition, volume 6112 of *Lecture Notes in Computer Science*, pages 30–39. Springer, 2010. (Cited on page 45.)

[Kochi 1999] Tsukasa Kochi and Takashi Saitoh. *User-defined template for identifying document type and extracting information from documents*. In Fifth International Conference on Document Analysis and Recognition, pages 127–130, sep 1999. (Cited on page 19.)

[Kullback 1951] Solomon Kullback and Richard A. Leibler. *On information and sufficiency*. The Annals of Mathematical Statistics, vol. 22, no. 1, pages 79–86, 1951. (Cited on page 8.)

[Lam 1993] Stephen W. Lam. *An adaptive approach to document classification and understanding*. In International Association for Pattern Recognition Workshop on Document Analysis Systems, pages 231–251, 1993. (Cited on page 19.)

[Larson 2010] Eric C. Larson and Damon M. Chandler. *Most apparent distortion: full-reference image quality assessment and the role of strategy*. Journal of Electronic Imaging, vol. 19, no. 1, page 011006, 2010. (Cited on page 87.)

[Lavallee 1995] Stéphane Lavallee. *Registration for computer-integrated-surgery: methodology, state of the art*. Computer Integrated Surgery: Technology and Clinical Applications, pages 77–97, 1995. MIT Press, Cambridge, Massachusettes. (Cited on page 42.)

[Lee 2006] Manhee Lee, Hun-Woo Yoo and Dong-Sik Jang. *Video scene change detection using neural network: improved ART2*. Expert Systems with Applications, vol. 31, no. 1, pages 13–25, 2006. (Cited on page 20.)

[Lehmann 1999] Thomas M. Lehmann, Claudia Gonner and Klaus Spitzer. *Registration for computed-integrated-surgery: methodolgy, state of the art*. IEEE Transactions on Medical Imaging, vol. 18, no. 11, pages 1049–1074, November 1999. (Cited on page 43.)

[Lelescu 2003] Dan Lelescu and Dan Schonfeld. *Statistical sequential analysis for real-time video scene change detection on compressed multimedia bitstream*. IEEE Transactions on Multimedia, vol. 5, no. 1, pages 106 – 117, march 2003. (Cited on page 20.)

[Li 2004] Ming Li, Xi Chen, Xin Li, Bin Ma and Paul Vitányi. *The similarity metric*. IEEE Transactions on Information Theory, vol. 50, no. 12, pages 3250–3264, December 2004. (Cited on page 26.)

[Li 2008] Ming Li and Paul Vitányi. An introduction to Kolmogorov complexity and its applications. Springer-Verlag, 3rd edition, 2008. (Cited on page 26.)

[Liang 2002] Jian Liang, David S. Doermann, Matthew Y. Ma and Jinhong Katherine Guo. *Page classification through logical labelling*. In 16th International Conference on Pattern Recognition, volume 3, pages 477–480, 2002. (Cited on pages 18 and 19.)

[Lienhart 1999] Rainer Lienhart. *Comparison of automatic shot boundary detection algorithms*. In Proceedings of SPIE, pages 290–301, 1999. (Cited on page 20.)

[Lienhart 2001] Rainer Lienhart. *Reliable transition detection in videos: a survey and practitioner's guide*. International Journal of Image and Graphics, vol. 1, pages 469–486, 2001. (Cited on page 20.)

[Lopresti 2000] Daniel P. Lopresti. *String techniques for detecting duplicates in document databases*. International Journal on Document Analysis and Recognition, vol. 2, no. 4, pages 186–199, 2000. (Cited on pages 17, 18 and 41.)

[Luo 2014] Xiaoxiao Luo, Qing Xu, Mateu Sbert and Klaus Schoeffmann. *F-divergences driven video key frame extraction*. In IEEE International Conference on Multimedia and Expo, pages 1–6, July 2014. (Cited on page 58.)

[Maderlechner 1997] Gerd Maderlechner, Peter Suda and Thomas Brückner. *Classification of documents by form and content*. Pattern Recognition Letters, vol. 18, pages 1225–1231, November 1997. (Cited on page 18.)

[Maes 1997] Frederik Maes, André Collignon, Dirk Vandermeulen, Guy Marchal and Paul Suetens. *Multimodality image registration by maximization of mutual information*. IEEE Transactions on Medical Imaging, vol. 16, no. 2, pages 187–198, 1997. (Cited on page 25.)

[Mao 2003] Song Mao, Azriel Rosenfeld and Tapas Kanungo. *Document structure analysis algorithms: a literature survey*. In Document Recognition and Retrieval, pages 197–207, 2003. (Cited on page 18.)

[Martins 2008] André F. T. Martins, Mário A. T. Figueiredo, Pedro M. Q. Aguiar, Noah A. Smith and Eric P. Xing. *Nonextensive entropic kernels*. In Proceedings of the 25th International Conference on Machine Learning, ICML '08, pages 640–647, 2008. (Cited on page 59.)

[Meng 1995] Jianhao Meng, Yujen Juan and Shih-Fu Chang. *Scene change detection in an MPEG-compressed video sequence*. In IS&T/SPIE Symposium Proceedings, volume 2419, pages 14–25. SPIE, 1995. (Cited on page 20.)

[Mentzelopoulos 2004] Markos Mentzelopoulos and Alexandra Psarrou. *Key-frame extraction algorithm using entropy difference*. In Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval, MIR '04, pages 39–45, 2004. (Cited on page 58.)

[Meshesha 2008] Million Meshesha and C. V. Jawahar. *Matching word images for content-based retrieval from printed document images*. International Journal on Document Analysis and Recognition, vol. 11, no. 1, pages 29–38, 2008. (Cited on page 18.)

[Mohamed 2009] Waleed Mohamed and Abdessamad Ben Hamza. *Nonextensive entropic image registration*. In International Conference on Image Analysis and Recognition, volume 5627 of *Lecture Notes in Computer Science*, pages 116–125. Springer, 2009. (Cited on page 45.)

[Money 2008] Arthur G. Money and Harry W. Agius. *Video summarisation: a conceptual framework and survey of the state of the art*. J. Visual Communication and Image Representation, vol. 19, no. 2, pages 121–143, 2008. (Cited on pages 19 and 57.)

[Nagasaka 1992] Akio Nagasaka and Yuzuru Tanaka. *Automatic video indexing and full-video search for object appearances*. In Proceedings of the IFIP TC2/WG 2.6 Second Working Conference on Visual Database Systems II, pages 113–127, Amsterdam, The Netherlands, 1992. North-Holland Publishing Co. (Cited on page 20.)

[Nagy 2000] George Nagy. *Twenty years of document image analysis in PAMI*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pages 38–62, 2000. (Cited on page 18.)

[Open Video Project] Open Video Project. *The open video project: a shared digitial video collection*. http://www.open-video.org/index.php. (Cited on pages 63 and 65.)

[Peng 2003] Hanchuan Peng, Fuhui Long and Zheru Chi. *Document image recognition based on template matching of component block projections*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 9, pages 1188–1192, 2003. (Cited on pages 17, 22 and 41.)

[Peng 2010] Jiang Peng and Qin Xiao-Lin. *Keyframe-based video summary using visual attention clues*. IEEE MultiMedia, vol. 17, no. 2, pages 64–73, 2010. (Cited on pages 19, 20 and 57.)

[Portes de Albuquerque 2004] Marcelo Portes de Albuquerque, Israel A. Esquef, Aline da Rocha Gesualdi Mello and Márcio Portes de Albuquerque. *Image thresholding using Tsallis entropy*. Pattern Recognition Letters, vol. 25, pages 1059–1065, 2004. (Cited on page 60.)

[Press 1992] William H. Press, Saul A. Teukolsky, William T. Vetterling and Brain P. Flannery. Numerical recipes in C. Cambridge University Press, 1992. (Cited on pages 44 and 49.)

[Rangoni 2011] Yves Rangoni, Abdel Belaïd and Szilárd Vajda. *Labelling logical structures of document images using a dynamic perceptive neural network*. International Journal on Document Analysis and Recognition, pages 1–11, 2011. (Cited on pages 18 and 19.)

[Rényi 1961] Alfréd Rényi. *On measures of entropy and information*. In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, volume 1, pages 547–561, Berkeley (CA), USA, 1961. University of California Press. (Cited on page 15.)

[Rigau 2004] Jaume Rigau, Miquel Feixas and Mateu Sbert. *An information theoretic framework for image segmentation*. In Proceedings of the IEEE International Conference on Image Processing, volume 2, pages 1193–1196, Singapore, Republic of Singapore, October 2004. (Cited on page 85.)

[Rigau 2008a] Jaume Rigau, Miquel Feixas and Mateu Sbert. *Informational aesthetics measures*. IEEE Computer Graphics and Applications, vol. 28, no. 2, pages 24–34, March 2008. (Cited on page 80.)

[Rigau 2008b] Jaume Rigau, Miquel Feixas and Mateu Sbert. *Informational dialogue with van Gogh's paintings*. In Computational Aesthetics in Graphics, Visualization, and Imaging. The Eurographics Association, 2008. (Cited on page 80.)

[Rigau 2010] Jaume Rigau, Miquel Feixas, Mateu Sbert and Christian Wallraven. *Toward auvers period: evolution of van Gogh's style*. In Computational Aesthetics, pages 99–106, 2010. (Cited on pages 80 and 82.)

[Sako 2003] Hiroshi Sako, Minenobu Seki, Naohiro Furukawa, Hisashi Ikeda and Atsuhiro Imaizumi. *Form reading based on form-type identification and form-data recognition*. In Seventh International Conference on Document Analysis and Recognition, pages 926–930. IEEE Computer Society, 2003. (Cited on page 18.)

[Santaló 1976] Lluís A. Santaló. Integral geometry and geometric probability. Addison-Wesley, Reading (MA), USA, 1976. (Cited on page 83.)

[Sbert 1993] Mateu Sbert. *An integral geometry based method for fast form-factor computation*. Computer Graphics Forum, vol. 12, no. 3, pages 409–420, 1993. (Cited on page 83.)

[Shannon 1948] Claude E. Shannon. *A mathematical theory of communication*. The Bell System Technical Journal, vol. 27, pages 379–423, 623–656, July, October 1948. (Cited on pages 5, 6 and 44.)

[Sheikh 2006] Hamid R. Sheikh and Alan C. Bovik. *Image information and visual quality*. IEEE Transactions on Image Processing, vol. 15, no. 2, pages 430–444, Feb 2006. (Cited on page 80.)

[Shin 2001] Christian Shin, David S. Doermann and Azriel Rosenfeld. *Classification of document pages using structure-based features*. International Journal on Document Analysis and Recognition, vol. 3, no. 4, pages 232–247, 2001. (Cited on pages 17, 18, 19, 22 and 41.)

[Shin 2006] Christian Shin and David S. Doermann. *Document image retrieval based on layout structural similarity*. In Proceedings of the International Conference on Image Processing, Computer Vision, & Pattern Recognition, Las Vegas, Nevada, USA, Volume 2, pages 606–612, 2006. (Cited on page 22.)

[Silva 2006] Ana Costa Silva, Alípio Mário Jorge and Luís Torgo. *Design of an end-to-end method to extract information from tables*. International Journal on Document Analysis and Recognition, vol. 8, no. 2-3, pages 144–171, 2006. (Cited on page 22.)

[Slonim 2000a] Noam Slonim and Naftali Tishby. *Agglomerative information bottleneck*. In Proceedings of Neural Information Processing Systems, pages 617–623. MIT Press, 2000. (Cited on page 14.)

[Slonim 2000b] Noam Slonim and Naftali Tishby. *Document clustering using word clusters via the information bottleneck method*. In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 208–215. ACM Press, 2000. Held in Athens, Greece. (Cited on pages 12 and 59.)

[Spitz 1999] A. Lawrence Spitz and Arman Maghbouleh. *Text categorization using character shape codes*. In SPIE Symposium on Electronic Imaging Science and Technology, pages 174–181, 1999. (Cited on page 18.)

[Studholme 1997] Colin Studholme. *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, 1997. (Cited on pages 23 and 25.)

[Taneja 1988] Inder J. Taneja. *Bivariate measures of type α and their applications*. Tamkang Journal of Mathematics, vol. 19, no. 3, pages 63–74, 1988. (Cited on pages 16, 46 and 58.)

[Tang 1997] Yuan Yan Tang and Jiming Liu. *Information acquisition and storage of forms in document processing*. In International Conference on Document Analysis and Recognition, pages 170–174, 1997. (Cited on page 22.)

[Tardini 2005] Giovanni Tardini, Costantino Grana, Rossano Marchi and Rita Cucchiara. *Shot detection and motion analysis for automatic MPEG-7 annotation of sports videos*. In 13th International Conference on Image Analysis and Processing, pages 653–660, 2005. (Cited on page 20.)

[Thompson 2011] William B. Thompson, Roland W. Fleming, Sarah H. Creem-Regehr and Jeanine K. Stefanucci. Visual perception from a computer graphics perspective. CRC Press, 2011. (Cited on page 60.)

[Tishby 1999] Naftali Tishby, Fernando C. Pereira and William Bialek. *The information bottleneck method*. In Hajek B and Sreenivas RS, editeurs, Proceedings of the 37th Annual Allerton Conference on Communication Control and Computing, volume pages, pages 368–377. University of Illinois, 1999. (Cited on page 14.)

[TRECVID] TRECVID. *TREC Video Retrieval Evaluation: TRECVID*. http://trecvid.nist.gov/. (Cited on pages 65, 72 and 74.)

[TRECVID 2007] TRECVID. *TRECVID 2007 shot boundary determination results*, 2007. http://www-nlpir.nist.gov/projects/tv2007/active/\results/shot.boundaries/runTable.web. (Cited on page 72.)

[Trier 1996] Øivind Due Trier, Anil K. Jain and Torfinn Taxt. *Feature extraction methods for character recognition - A survey*. Pattern Recognition, vol. 29, no. 4, pages 641–662, 1996. (Cited on pages 18 and 22.)

[Tsallis 1988] Constanino Tsallis. *Possible generalization of Boltzmann-Gibbs statistics*. Journal of Statistical Physics, vol. 52, pages 479–487, 1988. (Cited on page 15.)

[Tsallis 1998] Constanino Tsallis. *Generalized entropy-based criterion for consistent testing*. Physical Review E, vol. 58, no. 2, pages 479–487, 1998. (Cited on pages 16, 46 and 58.)

[Tsallis 2002] Constanino Tsallis. *Entropic nonextensivity: a possible measure of complexity*. Chaos, Solitons, & Fractals, vol. 13, no. 3, pages 371–391, 2002. (Cited on page 16.)

[Tsao 2003] Jeffrey Tsao. *Interpolation artifacts in multimodality image registration based on maximization of mutual information*. IEEE Transactions on Medical Imaging, vol. 22, pages 854–864, November 2003. (Cited on page 45.)

[Tseng 1997] Lin Yu Tseng and Rung Ching Chen. *The recognition of form documents based on three types of line segments*. In International Conference on Document Analysis and Recognition, pages 71–75, 1997. (Cited on pages 17, 18 and 41.)

[Unser 1999] Michael Unser. *Splines. A perfect fit for signal and image processing*. IEEE Signal Processing Magazine, vol. 16, pages 22–38, 1999. (Cited on page 44.)

[Urhan 2006] OË~guzhan Urhan, Kemal M. Güllü and Sarp Ertürk. *Modified phase-correlation based robust hard-cut detection with application to archive film*. IEEE Transactions on Circuits and Systems for Video Technology, vol. 16, no. 6, pages 753–770, June 2006. (Cited on page 20.)

[Verdú 2006] Sergio Verdú and Tsachy Weissman. *Erasure entropy*. In IEEE International Symposium on Information Theory, 2006. (Cited on page 84.)

[Verdú 1998] Sergio Verdú. *Fifty years of Shannon theory*. IEEE Transactions on Information Theory, vol. 44, no. 6, pages 2057–2078, 1998. (Cited on page 6.)

[Vila 2011] Marius Vila, Anton Bardera, Miquel Feixas and Mateu Sbert. *Tsallis mutual information for document classification*. Entropy, vol. 13, no. 9, pages 1694–1707, 2011. (Cited on pages 3 and 41.)

[Vila 2013] Marius Vila, Anton Bardera, Qing Xu, Miquel Feixas and Mateu Sbert. *Tsallis entropy-based information measures for shot boundary detection and keyframe selection*. Signal, Image and Video Processing, vol. 7, no. 3, pages 507–520, 2013. (Cited on pages 3 and 57.)

[Vila 2014] Marius Vila, Anton Bardera, Miquel Feixas, Philippe Bekaert and Mateu Sbert. *Analysis of image informativeness measures*. In IEEE International Conference on Image Processing, pages 1086–1090, October 2014. (Cited on pages 3 and 79.)

[Viola 1995] Paul A. Viola. *Alignment by maximization of mutual information*. PhD thesis, MIT Artificial Intelligence Laboratory (TR 1548), Massachusetts (MA), USA, 1995. (Cited on pages 23 and 25.)

[Wachowiak 2003] Matt P. Wachowiak, Renata Smolikova, Georgia D. Tourassi and Adel Said Elmaghraby. *Similarity metrics based on non-additive entropies for 2D-3D multimodal biomedical image registration*. In Proceedings of SPIE Medical Imaging, pages 1090–1100, 2003. Held in San Diego, CA. (Cited on page 45.)

[Wang 2011] Zhou Wang and Qiang Li. *Information content weighting for perceptual image quality assessment*. IEEE Transactions on Image Processing, vol. 20, no. 5, pages 1185–1198, May 2011. (Cited on page 80.)

[Wenzel 2001] Claudia Wenzel and Heiko Maus. *Leveraging corporate context within knowledge-based document analysis and understanding*. International Journal on Document Analysis and Recognition, vol. 3, no. 4, pages 248–260, 2001. (Cited on page 19.)

[Wolf 1996] Wayne Wolf. *Key frame selection by motion analysis*. In IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 2, pages 1228 –1231 vol. 2, may 1996. (Cited on page 20.)

[Xu 2010] Qing Xu, Pengcheng Wang, Bin Long, Mateu Sbert, Miquel Feixas and Riccardo Scopigno. *Selection and 3D visualization of video key frames*. In IEEE International Conference on Systems Man and Cybernetics, pages 52–59, 2010. (Cited on pages 20, 58, 60, 61, 62, 76 and 92.)

[Xu 2012] Qing Xu, Xiu Li, Zhen Yang, Jie Wang, Mateu Sbert and Jianfu Li. *Key frame selection based on Jensen-Rényi divergence*. In Proceedings of the 21st International Conference on Pattern Recognition, Tsukuba, Japan, pages 1892–1895, 2012. (Cited on page 58.)

[Xu 2014] Qing Xu, Yu Liu, Xiu Li, Zhen Yang, Jie Wang, Mateu Sbert and Riccardo Scopigno. *Browsing and exploration of video sequences: A new scheme for key frame extraction and 3D visualization using entropy based Jensen divergence*. Information Sciences, vol. 278, no. 0, pages 736 – 756, 2014. (Cited on page 58.)

[Yeung 2008] Raymond W. Yeung. Information theory and network coding. Information Technology: Transmission, Processing and Storage. Springer, 2008. (Cited on pages 6, 8, 9, 12, 24 and 81.)

[Yoo 2006] Hun-Woo Yoo, Han-Jin Ryoo and Dong-Sik Jang. *Gradual shot boundary detection using localized edge blocks*. Multimedia Tools and Applications, vol. 28, pages 283–300, 2006. (Cited on page 20.)

[Yuan 2007] Jinhui Yuan, Huiyi Wang, Lan Xiao, Wujie Zheng, Jianmin Li, Fuzong Lin and Bo Zhang. *A formal study of shot boundary detection*. IEEE Transactions on Circuits and Systems for Video Technology, vol. 17, no. 2, pages 168 –186, feb. 2007. (Cited on page 20.)

[Zabih 1995] Ramin Zabih, Justin Miller and Kevin Mai. *A feature-based algorithm for detecting and classifying scene breaks*. In Proceedings of the third ACM international conference on Multimedia, pages 189–200, 1995. (Cited on page 20.)

[Zhang 2006] Weigang Zhang, Jianqiu Lin, Xiaopeng Chen, Qingming Huang and Yang Liu. *Video shot detection using hidden markov models with complementary features*. In First International Conference on Innovative Computing, Information and Control, volume 3, pages 593 –596, September 2006. (Cited on page 58.)

[Zhuang 1998] Yueting Zhuang, Yong Rui, Thomas S. Huang and Sharad Mehrotra. *Adaptive key frame extraction using unsupervised clustering*. In International Conference on Image Processing, volume 1, pages 866–870, October 1998. (Cited on page 20.)