

A Normalized Soft Window-Based Similarity Measure To Extend The Rand Index

Romain Quéré

Lab. Mathématiques, Image et Applications
Université de La Rochelle
PRES Limousin-Poitou-Charentes
France
Email: romain.quere@univ-lr.fr

Carl Frélicot

Lab. Mathématiques, Image et Applications
Université de La Rochelle
PRES Limousin-Poitou-Charentes
France
Email: carl.frelicot@univ-lr.fr

Abstract—This article addresses the problem of the construction of concordance measures between two crisp, fuzzy and possibilistic partitions from their coincidence matrices. Two existing approaches are reviewed and their advantages and drawbacks are exhibited so that a new Rand index taking profits of both is proposed. Numerous experimental results show that it outperforms other extensions of the Rand index.

Index Terms—Soft partitions, similarity measure, Rand index

I. INTRODUCTION

A partition of a dataset $X = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ of n objects is a set of c non-empty subsets of X . These subsets, called *clusters* thereafter, generally regroup objects along common properties they share. Partitions can be represented by a $(c \times n)$ partition matrix U , whose general term u_{ik} represents the membership degree of each $\mathbf{x}_k$ ($k = 1, n$) to the i th cluster ($i = 1, c$). Each column of U is then the membership vector $\mathbf{u}_k$ of the object $\mathbf{x}_k$. For the sake of clearness and simplicity, a partition will be identified with its corresponding partition matrix in the following. Depending on whether or not clusters are mutually exclusive and on whether or not objects of X totally belong to each cluster, partitions can be of different nature. We distinguish the three main categories [1], called the sets of *possibilistic*, *fuzzy* and *crisp* partitions, respectively defined as :

- $\mathcal{M}_{pcn} = \{U \in \mathbb{R}^{cn} : u_{ik} \in [0, 1]\},$
- $\mathcal{M}_{fcn} = \{U \in \mathcal{M}_{pcn} : \sum_{i=1}^c u_{ik} = 1\},$
- $\mathcal{M}_{hcn} = \{U \in \mathcal{M}_{fcn} : u_{ik} \in \{0, 1\}\}.$

Moreover, we will refer to the set $\mathcal{M}_{pcn} - \mathcal{M}_{hcn}$ as the set of *soft partitions*, according to the definition in [2].

In many fields, figuring out how close are two partitions of the same dataset X is a problem of great interest. Generally speaking, it can indeed be useful to confront how different experts class different objects. In clustering, and more precisely in cluster validity, it also enables to evaluate and compare the quality of different algorithms, or the same algorithm with different parameters. This important problem has been tackled since decades by attempting to define concordance measures assessing the degree of agreement between partitions, so that a large number of measures exist in the literature. One can distinguish two kinds of measures, depending on the use made of them, see [3] : an *external index* will be used

when one wants to assess the quality of a partition compared to a reference one, while a *relative index* will be used to only evaluate how close two partitions are. Since the frontier between these two categories is thin, we will more generally refer to both of them as *comparison indices* in this paper.

Early comparison indices have been dedicated to crisp partitions, and a new interest has recently emerged for this topic thanks to the definition of the new class of soft partitions mentioned above. Many fuzzy or soft concordance measures have been recently proposed, and one can distinguish between them those relying on direct approaches from those extending historical crisp ones. Among the formers, one can cite measures based on the comparison of the whole set of α -cuts of the two partitions [4], or on a measure of uncertainty of both partitions [5]. The latters, such as the fuzzy transfert distance [6], extend existing crisp indices in the sense they reduce to crisp ones while comparing two crisp partitions. This category of soft indices also regroup a class of propositions called *coincidence-based indices*, extending well known indices such as the Rand, Jaccard and Fowlkes-Mallow ones [7]. Such soft indices solve most of the intrinsic problems of the comparison of partitions, and specifically :

- the *correspondance* or *cluster matching* problem, as discussed in [2],
- the problem of comparing partitions with different number of clusters.

Indeed, by comparing how objects are paired, i.e. belong or not belong to the same cluster in both partitions, no preliminary cluster matching is required, and the number of clusters has computationally no importance with these indices. The comparison is based on the construction, for each partition to be compared, of a coincidence matrix crossing the partition with itself. Many propositions have been made to compute and compare such matrices. Most of them are based on fuzzy set-theoretic tools, extending thus directly the set-theoretic approach used for the crisp indices, e.g. in [8], [9] and [10]. Another approach is possible, as proposed in [11]. It consists in using a similarity measure between the two coincidence matrices based on a distance between the membership vectors of the partitions. Because each of these constructions presents drawbacks and advantages, we propose in this paper a new

soft Rand index taking profits of both approaches.

This paper is organized as follows. A review of existing soft Rand indices is presented in section II and their strengths and weaknesses are confronted. In section III, the new Rand index is proposed. Its main characteristics are investigated in section IV through an intensive comparative experimental study. A conclusion is given in section V.

II. SOME EXISTING COINCIDENCE-BASED SOFT RAND INDICES

Let $U \in \mathcal{M}_{hcn}$ be a crisp partition of a dataset X . The *coincidence matrix* Ψ_U of U is a symmetric $(n \times n)$ matrix whose general term

$$\psi_{U,kl} = \sum_{i=1}^c u_{ik} \wedge u_{il} = \begin{cases} 1 & \text{if } \mathbf{u}_k = \mathbf{u}_l \\ 0 & \text{else} \end{cases} \quad (1)$$

describes the degree to which the objects $\mathbf{x}_k$ and $\mathbf{x}_l$ of X belong to the same cluster in U . Given a second partition V of the same dataset and its coincidence matrix Ψ_V , the crisp Rand Index measuring the concordance between U and V is defined as follows:

$$RI(U, V) = \frac{n_{11} + n_{00}}{n_{11} + n_{10} + n_{01} + n_{00}}. \quad (2)$$

As defined, the Rand index takes values in $[0, 1]$ and equals 1 when $U \equiv V$. The number of pairs of objects $(\mathbf{x}_k, \mathbf{x}_l)$ belonging to the same cluster (respectively to different clusters) in both U and V is given by n_{11} (respectively n_{00}), so that the numerator in (1) represents the agreement of U and V . On the other side, disagreement between both partitions is represented by the numbers of pairs belonging to the same cluster in U (respectively V) and in different clusters in V (respectively U), namely n_{10} and n_{01} . Let us assume $\alpha, \beta \in \{0, 1\}$, the different numbers of pairs $n_{\alpha\beta}$ are given:

$$n_{\alpha\beta} = \sum_{k=2}^n \sum_{l=1}^{k-1} ((1 - \alpha) + (2\alpha - 1) \psi_{U,kl}) \wedge ((1 - \beta) + (2\beta - 1) \psi_{V,kl}). \quad (3)$$

This index, as many other crisp indices constructed on the same principle, *e.g.* in [7], can be equally computed from the contingency matrix crossing the two partitions to be compared, see [12]. This equivalence, that do not generally holds in the soft case, has led to two different classes of soft extensions, namely the *contingency-based*, see the recent work by Anderson et al [2], and the *coincidence-based* ones we are interested in. We recall hereafter two recent extensions of the Rand index belonging to this latter class to which we will compare our proposition.

A. The fuzzy set-theoretic approach

Since the crisp Rand index as defined by (2) is built on a logical and set-theoretic approach, some authors have proposed to extend it by use of fuzzy set-theoretic tools. Namely, the t-norm and the sum (or Σ -count) operators are used to extend the classical concepts of intersection (or conjunction) and cardinality, respectively. Briefly, a t-norm is a commutative,

associative and monotonic function $\top : [0, 1]^2 \rightarrow [0, 1]$ having 1 for neutral element. The most common t-norms are given in table I and an interest reader can refer to [13] for a complete review.

TABLE I
BASIC TRIANGULAR NORMS AND SOME PARAMETRIZED FAMILIES OF T-NORMS.

Standard	$\top_M(a, b) = \min(a, b)$
Product	$\top_P(a, b) = ab$
Łukasiewicz	$\top_L(a, b) = \max(a + b - 1, 0)$
Frank $\lambda \in \mathbb{R}_+^* \setminus \{1\}$	$\top_{F_\lambda}(a, b) = \ln \left(1 + \frac{(\lambda^a - 1)(\lambda^b - 1)}{\lambda - 1} \right)$
Hamacher $\lambda \in \mathbb{R}_+$	$\top_{H_\lambda}(a, b) = \frac{ab}{\lambda + (1 - \lambda)(a + b - ab)}$

Extending (1) to the soft domain results in defining the elements of the soft coincidence matrix $\Psi_U^\top$ of a soft partition matrix U by [3]:

$$\psi_{U,kl}^\top = \sum_{i=1}^c \top(u_{ik}, u_{il}). \quad (4)$$

However, it has been recently shown that such a direct extension leads to counterintuitive values for the diagonal elements $\psi_{U,kk}^\top$ of the soft matrix, see [9], [10]. Indeed, these elements, which represent the degree to which each object $\mathbf{x}_k$ belongs to the same cluster as itself, is generally not 1, and can moreover be much less than 1 because of the $\top$ aggregation operator. The authors thus recommend to normalize the matrix $\Psi_U^\top$ to bring its diagonal terms to ones. A generic solution can be found in [10], leading to a soft normalized coincidence matrix $\Phi_U^\top$ whose general term is defined by:

$$\phi_{U,kl}^\top = \frac{\psi_{U,kl}^\top}{\top \left(K_\top \left(\psi_{U,kk}^\top \right), K_\top \left(\phi_{U,ll}^\top \right) \right)} \quad (5)$$

where $K_\top : [0, 1] \rightarrow [0, 1]$ is a t-norm dependant function such that $\top(K_\top(a), K_\top(a)) = a$. Given another soft normalized coincidence matrix $\Phi_V^\top$ computed from another soft partition V , the numbers $n_{\alpha\beta}$ (3) are extended to the soft domain by:

$$n_{\alpha\beta}^\top = \sum_{k=2}^n \sum_{l=1}^{k-1} \top \left(((1 - \alpha) + (2\alpha - 1) \phi_{U,kl}^\top), ((1 - \beta) + (2\beta - 1) \phi_{V,kl}^\top) \right). \quad (6)$$

The derived *Normalized Soft Rand Index (NSRI)* is then computed directly from (2), and it is easy to show it reduces to its crisp counterpart when U and V are crisp.

B. The distance-based measure of equivalence approach

In [11], Hüllermeier and Rifqi propose another extension of the Rand index lying on a distance-based measure of equivalence between the membership vectors $\mathbf{u}_k$ of a soft partition U , so that the terms of the coincidence matrix Ψ_U^{HR} of U are defined as:

$$\psi_{U,kl}^{HR} = 1 - \frac{\|\mathbf{u}_k - \mathbf{u}_l\|}{\sqrt{2}}. \quad (7)$$

Given another soft partition V , the *Soft Equivalence Rand Index* ($SERI$) is then given by:

$$SERI(U, V) = \frac{1}{q} \sum_{k=2}^n \sum_{l=1}^{k-1} 1 - |\psi_{U,kl}^{HR} - \psi_{V,kl}^{HR}| \quad (8)$$

where $q = \frac{n(n-1)}{2}$ is the number of different pairs of objects in X . It is shown that $SERI$ reduces as well to the crisp Rand index when comparing two crisp partitions. Furthermore it presents the great advantage to take its maximum value of 1 when $U \equiv V$ whatever the space in which U and V are defined ; which is not the case for the previously presented $NSRI$. This lead to an interesting discussion about how the two coincidence matrices are aggregated to build the index, depending on the approach. Indeed, keep in mind the terms of the coincidence matrices describe the degree to which two elements belong to the same cluster in a given partition, and remember the Rand index is meant to assess how much the two partitions are in agreement. Then, while $\psi_{U,kl}$ is close to $\psi_{V,kl}$, one can intuitively conclude in an agreement between U and V concerning the $(\mathbf{x}_k, \mathbf{x}_l)$ pair of objects. As a direct consequence, the given pair should then have a high contribution to the value of the Rand index. However, this is not the case with the $NSRI$ presented in section II-A. As an example, let $a = 0.3$ and $b = 0.29$ be the degrees to which two objects $\mathbf{x}_k$ and $\mathbf{x}_l$ are paired in two partitions. With the standard t-norm $\top_M$, and according to (6) and (2), the contribution of a and b to the agreement $n_{11} + n_{00} = \min(0.3, 0.29) + \min(1 - 0.3, 1 - 0.29) = 0.99$, which is a value one can expect. Nevertheless, their overall impact on the Rand index is drastically pulled down by their contribution to the disagreement $n_{10} + n_{01} = \min(0.3, 1 - 0.29) + \min(1 - 0.3, 0.29) = 0.59$. At the opposite, the contribution of a and b to the $SERI$, given by $s^{HR}(a, b) = 1 - |a - b| = 0.99$ in (8), is total and avoid the previously exposed drawback. It is interesting to point out that this s^{HR} function, called *degree of concordance* by its authors, is actually a *similarity measure* since it satisfies the following axioms:

$$s(a, b) = s(b, a) \quad (\text{symmetry}) \quad (9)$$

$$s(a, a) \geq s(a, b), \quad \forall b \quad (\text{maximality}) \quad (10)$$

$$s(a, a) = 1 \quad (\text{unity}) \quad (11)$$

This remark lies at the heart of this paper. On one hand, by preserving the set-theoretic sense of the original Rand index, $NSRI$ presents the strong advantage to lie on a strong and admitted theory, and its well behaviour has been shown in [10]. However, as exhibited above, it also presents a serious drawback because of the way the terms of the coincidence matrices are aggregated. On the other hand, the $SERI$, by redefining the original Rand index formula by means of a similarity measure, correct this. In the next section, we propose a new Rand index which takes profit of both approach, based on a new family of similarity measures.

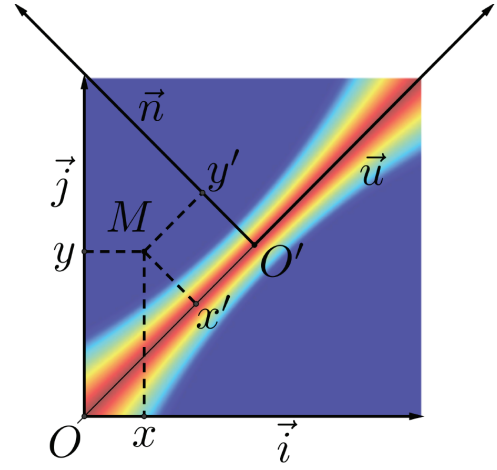


Fig. 1. Isosurface of s_p^w computed with w_T , $t = 0.1$, $r = 0.3$ and $o = \frac{1}{2}$ and geometrical explanation of its construction

III. THE NORMALIZED SOFT WINDOW-BASED SIMILARITY RAND INDEX

Let $O' = {}^t(o, o)$ be a point lying on the diagonal line $\Delta : x = y$ of the unit square and let $\vec{u}$ and $\vec{n}$ be two unit orthogonal vectors as shown in Fig. 1. Each point $M = {}^t(x, y) \in [0, 1]^2$, whose coordinates express in $(O, \vec{i}, \vec{j})$, can be transformed into $M' = {}^t(x', y')$, with respect to $(O', \vec{u}, \vec{n})$. Given a zero-mean symmetric window function w taking values in $[0, 1]$ such that $w(0) = 1$ and a symmetric profile function p taking values in $[0, 1]$ satisfying $p(x) > 0, \forall x$, the function $s_p^w : [0, 1]^2 \rightarrow [0, 1]$ defined by

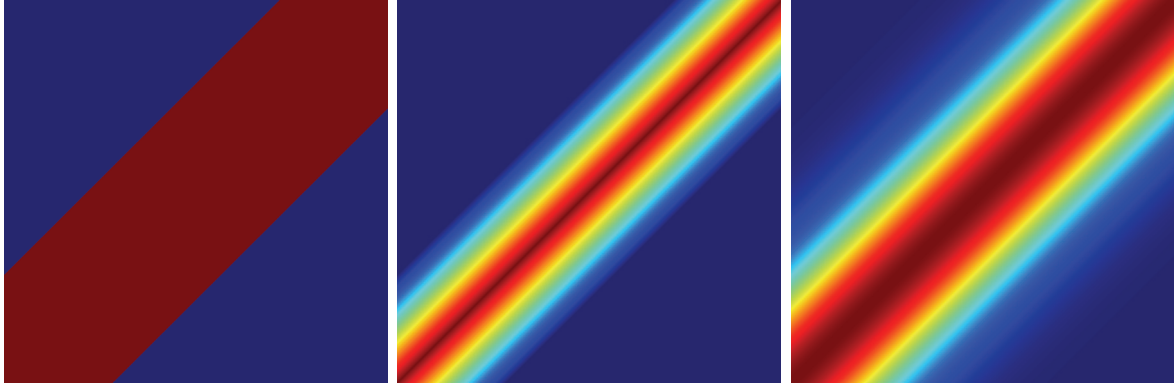
$$s_p^w(x, y) = w\left(\frac{y'}{p(x')}\right) \quad (12)$$

is a measure of similarity between x and y w.r.t. axioms (9), (10) and (11). The proof being trivial, it is left to the reader. In (12), w controls the way x and y are considered similar with regard to their difference. Formally, it is a weighting function acting on the distance between M and Δ such that the farther M is from Δ , the smaller $s_p^w(x, y)$ is. Thus, choosing a window function w dropping quickly from 1 leads to a drastic similarity function s_p^w , i.e. only close x and y values will be considered similar. Many symmetric window functions exist, some usual ones are reported in Table II. Fig. 2(b) presents the cuts of three s_p^w obtained with the three window functions w_U , w_T and w_G and whose isosurfaces are presented in Fig. 2(a). This example demonstrates how the behaviour of the similarity function s_p^w can be drastically modified according to w .

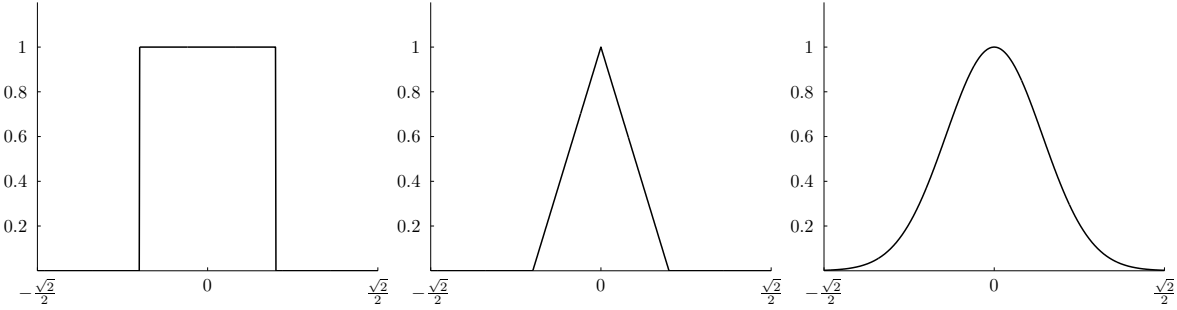
TABLE II
USUAL WINDOW FUNCTIONS

Uniform	$w_U(x) = 1_{(x \leq 1)}$
Triangular	$w_T(x) = (1 - x)1_{(x \leq 1)}$
Gaussian	$w_G(x) = \exp(-x^2)1_{(x \leq 1)}$
Epanechnikov	$w_E(x) = (1 - x^2)1_{(x \leq 1)}$

In the same manner, the profile function p influences the similarity between x and y with respect to the distance between



(a) $w = w_U, w_T$ and w_G (from left to right), $t = 0.2, r = 0, o$ is undefined



(b) Cut views of s_p^w plotted above along $x + y = 1$

Fig. 2. Isosurfaces and cut views of some s_p^w

M and O' . This allows to favour or not some specific ranges of values, according to the chosen *origin* parameter o . We propose to use the following p function:

$$p(x') = t + r x'^2 \quad (13)$$

where $t \in]0, \sqrt{2}]$ and $r \in \mathbb{R}_+$ are user defined parameters, called respectively the *thickness* and *curvature* parameters.

It is easy to show that with this particular profile function, O' has no influence whenever $r = 0$, so that no value for o is required.

As an example, Fig. 1 presents the isosurface of s_p^w with the following settings: $w = w_T$, a thickness $t = 0.1$, a curvature $r = 0.3$ and an origin $o = \frac{1}{2}$. One can observe that, for all M , the closer x and y are to $\frac{1}{2}$, the more selective $s_p^w(x, y)$ is, as expected. Fig. 3 also show examples of s_p^w functions using most of w functions of Table II with particular values for t, r and o , so that each row exhibits the role of each parameter. For instance, Fig. 3(a) shows the influence of t on the selectivity of s_p^w , according to its value (a small one, a middle one and the maximum one).

From this construction, we propose to define a new index measuring the concordance between two soft partitions U and V , called the *Normalized Soft Window-based similarity Rand Index* (*NSWRI*), by:

$$NSWRI(U, V) = \frac{1}{q} \sum_{k=2}^n \sum_{l=1}^{k-1} s_p^w(\phi_{U,kl}^\top, \phi_{V,kl}^\top) \quad (14)$$

where $\phi_{U,kl}^\top$ and $\phi_{V,kl}^\top$ are the terms of the normalized coincidence matrices Φ_U and Φ_V constructed according to (5).

Whatever the parameters used for the similarity function s_p^w , it is easy to show that this index reduces to the original crisp Rand index while comparing two crisp partitions. Moreover, like the *SERI*, it reaches its maximum value of 1 when $U \equiv V$, whatever their nature.

The set of parameters an user has to choose in order to compute $NSWRI(U, V)$ is $\{\top, w, t, r, o\}$. Since the range of values of each parameter is wide, we will only show some examples in the the following experimental results.

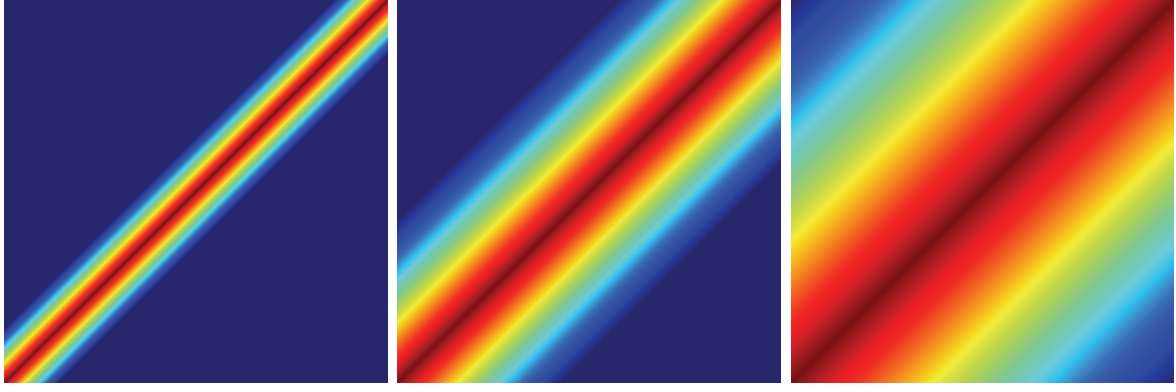
IV. EXPERIMENTAL RESULTS

A. Numerical experiments on fuzzy partitions.

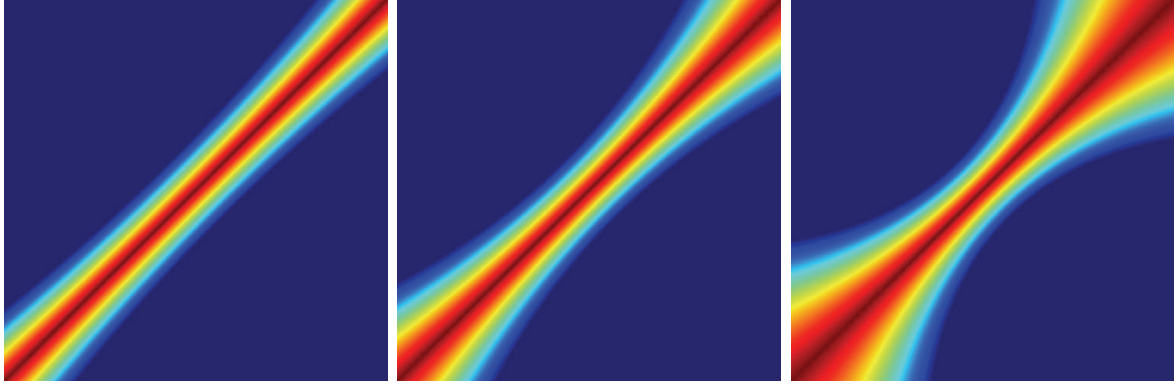
In this section, we propose to investigate the specificities and possibilities of this new index by mean of a simple set of synthetic partitions composed of one crisp partition

$U_1 = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \in \mathcal{M}_{h34}$ and eight fuzzy partitions in $\mathcal{M}_{f34}$:

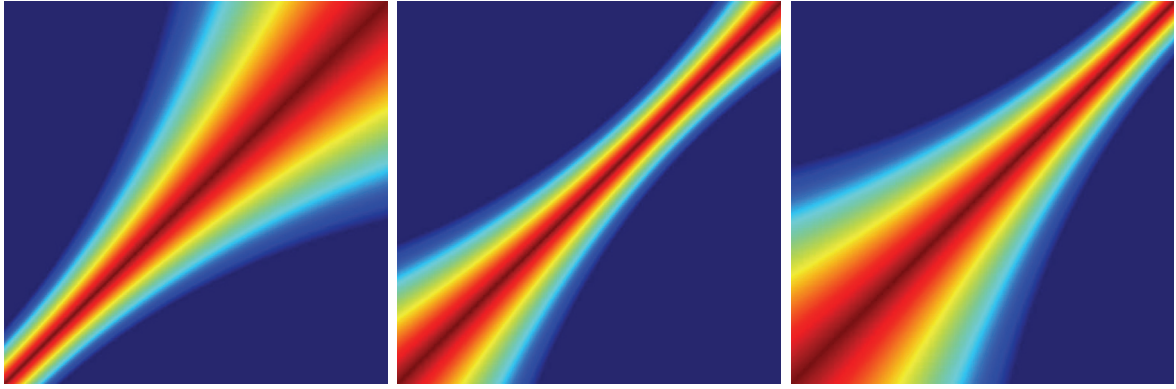
$$\begin{aligned} U_2 &= \begin{pmatrix} 0.05 & 0.10 & 0.85 & 0.95 \\ 0.00 & 0.90 & 0.15 & 0.00 \\ 0.95 & 0.00 & 0.00 & 0.05 \end{pmatrix}, U_3 = \begin{pmatrix} 0.10 & 0.10 & 0.80 & 0.90 \\ 0.00 & 0.80 & 0.20 & 0.00 \\ 0.90 & 0.10 & 0.00 & 0.10 \end{pmatrix}, \\ U_4 &= \begin{pmatrix} 0.20 & 0.20 & 0.70 & 0.80 \\ 0.00 & 0.70 & 0.20 & 0.10 \\ 0.80 & 0.10 & 0.10 & 0.10 \end{pmatrix}, U_5 = \begin{pmatrix} 0.20 & 0.30 & 0.60 & 0.70 \\ 0.10 & 0.60 & 0.30 & 0.10 \\ 0.70 & 0.10 & 0.10 & 0.20 \end{pmatrix}, \\ U_6 &= \begin{pmatrix} 0.30 & 0.30 & 0.50 & 0.60 \\ 0.10 & 0.50 & 0.30 & 0.20 \\ 0.60 & 0.20 & 0.20 & 0.20 \end{pmatrix}, U_7 = \begin{pmatrix} 0.30 & 0.35 & 0.40 & 0.50 \\ 0.20 & 0.40 & 0.35 & 0.25 \\ 0.50 & 0.25 & 0.25 & 0.25 \end{pmatrix}, \\ U_8 &= \begin{pmatrix} 0.35 & 0.35 & 0.35 & 0.35 \\ 0.30 & 0.35 & 0.35 & 0.30 \\ 0.35 & 0.30 & 0.30 & 0.35 \end{pmatrix}, U_9 = \begin{pmatrix} 0.40 & 0.45 & 0.30 & 0.35 \\ 0.30 & 0.35 & 0.45 & 0.25 \\ 0.30 & 0.20 & 0.25 & 0.40 \end{pmatrix}. \end{aligned}$$



(a) $w_T, t = 0.1, 0.3$ and $\frac{\sqrt{2}}{2}$ (from left to right), $r = 0, o$ is undefined



(b) $w_T, t = 0.1, r = 0.1, 0.3$ and 0.8 (from left to right), $o = 0.5$



(c) $w_T, t = 0.1, r = 0.3, o = 0, 0.7$, and 1 (from left to right)

Fig. 3. Isosurfaces of some s_p^w

These partitions are chosen such that:

- $U_2, \dots, U_7$ reduce to U_1 if hardened according to the maximal membership degree, while U_8 may not because of ties and U_9 does not
- $U_{\alpha+1}$ is fuzzier than U_α , $\forall \alpha \in \{1, \dots, 7\}$ according to the *Partition Entropy* [1] (see table III), which indicates the amount of fuzziness in a given partition:

$$PE(U) = -\frac{1}{n} \sum_{i=1}^c \sum_{k=1}^n u_{ik} \log(u_{ik}) \quad (15)$$

One can intuitively state that, according to their proximity

TABLE III
FUZZINESS OF PARTITIONS U_α , $\forall \alpha \in \{1, \dots, 8\}$

	U_1	U_2	U_3	U_4	U_5	U_6	U_7	U_8
$PE(U)$	0	0.29	0.45	0.69	0.85	0.98	1.06	1.10

with U_1 , these partitions can thus be sorted decreasingly as $\{U_2, U_3, U_4, U_5, U_6, U_7, U_9, U_8\}$. Note that U_9 is considered closer to U_1 than U_8 . Although this may seem counterintuitive, it can be explained simply. Indeed, U_1 is fully deterministic whereas U_8 presents, at the extreme opposite, complete uncer-

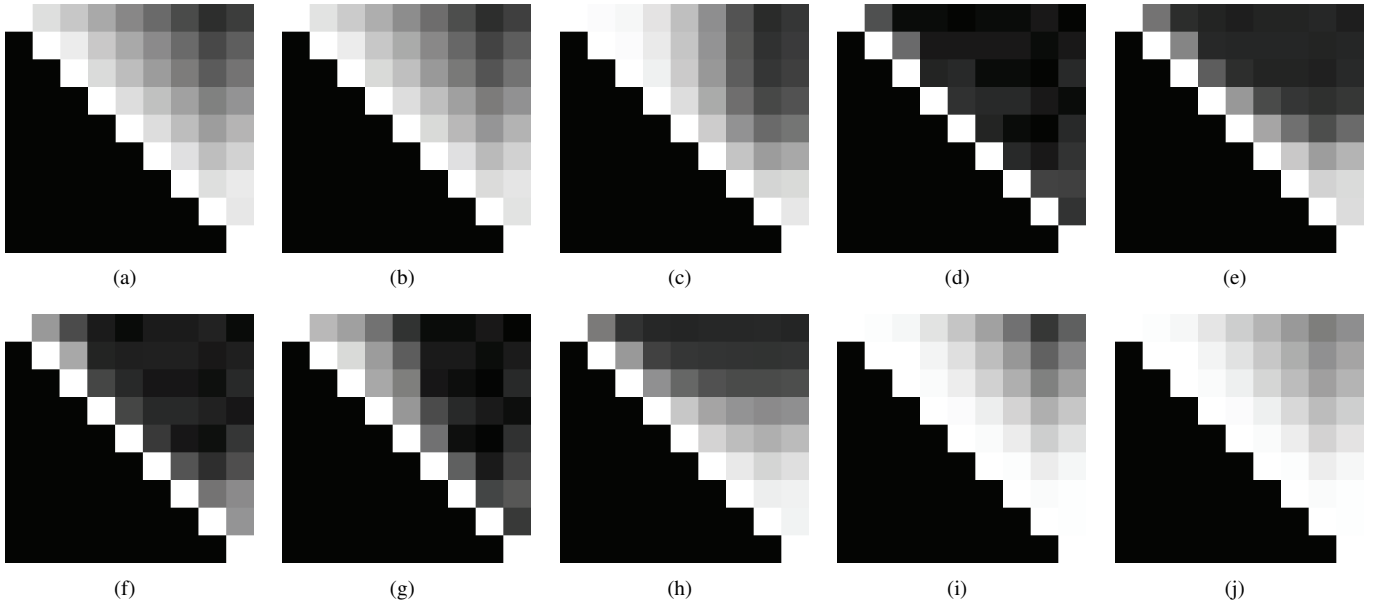


Fig. 4. Dissimilarity matrices crossing the nine partitions U_α with *SERI* (a) and *NSWRI* with different parameters (b)–(j) described in table IV

tainty. Thanks to a possible permutation of clusters, U_9 then shares more information in common with U_1 than U_8 does, as confirmed by the partition entropy $PE(U_9) = 1.07$.

So as to validate our proposal regarding to the intuitive proximities discussed above, nine dissimilarity matrices crossing all the pairs of partitions (U_i, U_j) with *NSWRI* are computed for different values of each of the five parameters $\{\top, w, t, r, o\}$. These nine dissimilarity matrices are presented in Fig. 4 with an additional matrix computed with *SERI* for comparison, presented in Fig. 4(a). The parameters used to compute the nine dissimilarity matrices (b)–(j) are reported in table IV. For each of them, partitions are ordered in rows and columns according to their index. The brighter a cell of a matrix is, the higher is the value of the index.

As one can directly see, the statement made earlier about the relative proximities between the nine partitions mostly holds, according to the first row of each dissimilarity matrix, especially for (b), (c), (e), (h), (i) and (j) where *NSWRI* is not too much drastic. While the statement does not hold, *i.e.* with (d), (f) and (g), it is interesting to note that *NSWRI* almost provides a strong separation between the partitions, so that a threshold could be easily tuned to decide whether two partitions are similar or not. Moreover, other statements can be made while comparing some couples or triples of dissimilarity matrices:

- (a) and (b) clearly show that *NSWRI* can be tuned to behave in the same way *SERI* does.
- (b) and (c) demonstrate the influence of the t-norm. Indeed, by choosing a drastic t-norm such as $\top_{H_{100}}$, our proposal favours the less fuzzy partitions.
- the influence of the thickness parameter t is crucial as it enables to tune how drastic the index is, as shown by (b) and (d).
- (e), (f) and (g) show the influence of the chosen origin

parameter o while the curvature r is not null: choosing o near 0 leads *NSWRI* to favour extremely fuzzy partitions, while choosing o near 1 will lead *NSWRI* to the opposite behaviour.

- (e) and (h) demonstrate the influence of the degree of curvature chosen for r on the previously described behaviour, *i.e.* a high value of r reinforces it.
- (b), (i) and (j) testify of the influence the window function w has on the selectivity and the range of response our proposal have. For instance, partitions U_α and $U_{\alpha+1}$ are considered closer by *NSWRI* using w_E than using w_T .

All these remarks show that our *NSWRI* is a really versatile tool for comparing partitions, that can be tuned to behave in many different ways.

TABLE IV
PARAMETERS USED TO COMPUTE *NSWRI* IN THE EXPERIMENTATIONS DESCRIBED BY FIG. 4

Dissimilarity matrix	$\top$	w	t	r	o
(b)	$\top_M$	w_T	$\frac{\sqrt{2}}{2}$	0	$\times$
(c)	$\top_{H_{100}}$	w_T	$\frac{\sqrt{2}}{2}$	0	$\times$
(d)	$\top_M$	w_T	0.1	0	$\times$
(e)	$\top_M$	w_T	0.1	0.3	0
(f)	$\top_M$	w_T	0.1	0.3	0.5
(g)	$\top_M$	w_T	0.1	0.3	1
(h)	$\top_M$	w_T	0.1	0.8	0
(i)	$\top_M$	w_E	$\frac{\sqrt{2}}{2}$	0	$\times$
(j)	$\top_M$	w_G	$\frac{\sqrt{2}}{2}$	0	$\times$

B. Numerical experiments on possibilistic partitions.

Consider the set composed of the crisp partition U_1 of the previous section and eight possibilistic $U_{\{2\dots 9\}}^p$ partitions defined from the previous fuzzy partitions $U_{\{2\dots 9\}} \in \mathcal{M}_{p34}$ such as $U_\alpha^p = \frac{U_\alpha}{2}$. As previously, the nine partitions are crossed

using *SERI* and *NSWRI* to compute two dissimilarity matrices presented in Fig. 5. The parameters used for the computation of *NSWRI* are the same as those used for Fig. 4(b), see table IV.

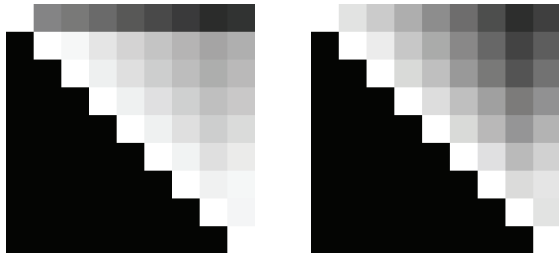


Fig. 5. Dissimilarity matrices crossing nine possibilistic partitions U_{α}^p , obtained with *SERI* (left) and *NSWRI* (right) with $\{\top_M, w_T, t = \frac{\sqrt{2}}{2}, r = 0\}$

One can observe that the *SERI* makes a strong distinction between the crisp and the possibilistic partitions. At the opposite, results given here by *NSWRI* strictly present no differences with the ones obtained with the previous set of fuzzy partitions U_{α} , as described in Fig. 4(b). This behaviour is the consequence of the normalization applied to the coincidence matrix of each partition. Moreover, this property holds for any set of parameters $\{\top, w, t, r, o\}$ while $\top \in \{\top_M, \top_P\}$ and barely holds with little differences with other t-norms, so that results obtained in previous section holds for possibilistic partitions. An interesting discussion is that no existing statement has already been made about the way possibilistic partitions must be compared to crisp or fuzzy ones, so that both *SERI* and *NSWRI* behaviour can be defended and used, depending on the application.

C. Comparing two fuzzy partitions provided by FCM.

We consider four well known datasets from the UCI Machine Learning Repository [14] presenting various characteristics in terms of: number n of samples, number p of features, true number c^* of classes/clusters and degree of overlap between clusters, as summarized in table V. For each dataset, the *Fuzzy C-Means algorithm* (FCM) [15] is run to produce a reference fuzzy partition R and a collection of 14 fuzzy c -partitions Q_c , $c \in 2, \dots, 14$. Each partition Q_c is compared to R with different indices, namely *NSRI*, *SERI* and our proposal *NSWRI* with different parameters, provided the index requires some parametrization. For each index I , the computed value of $I(R, Q_c)$ is stored for each value of c . Fig. 6 presents some examples of resulting curves. Each column is for a dataset, the first, second, third and fourth rows are respectively the results obtained by computing *NSWRI* with $\{\top_P, w_T, 0.1, 0.1, 0\}$, $\{\top_P, w_T, 0.1, 0.8, 0\}$, $\{\top_P, w_T, 0.1, 0.1, 1\}$ and $\{\top_P, w_T, 0.1, 0.8, 1\}$. *NSRI* is computed with $\top_P$ as well.

As expected, each index expresses its maximum value when c reaches the real number of classes c^* , and they share the tendency to drop from this maximum toward an asymptotic value, different in each case and for each index and depending

TABLE V
REAL DATASETS AND THEIR CHARACTERISTICS.

dataset	n	p	c^*	clusters property
Ionosphere	351	34	2	very strong overlap
Pima	768	9	2	strong overlap
Iris	150	4	3	slight overlap
Thyroid	215	5	3	slight overlap

on the overlap of the dataset. *NSWRI* does not always reach a value as large as the other two indices, but it presents a better dynamic and is thus more discriminant. One can observe that this behaviour is influenced by the choice of the origin parameter o . Indeed, the separation is emphasized in the whole set of examples when $o = 1$, backing the statements made in previous experiments, see section IV-A. Nevertheless, it seems that choosing a large value for r weakens the discriminant power of *NSWRI*, according to the results presented in each second and fourth row of Fig. 6.

V. CONCLUSION

In this article, we propose a new extension of the Rand index, meant to take benefits of two existing ones that belong to the class of coincidence-based comparison indices. These approaches lie respectively on fuzzy set-theoretic and similarity principles. It is shown that the former, although it is based on a direct extension of the crisp Rand index to the soft domain, presents some drawbacks because of the way the coincidence matrices are aggregated. These drawbacks are avoided by the latter, thanks to the use of a similarity measure. A new parametrized similarity measure is proposed to correct the drawbacks. It allows us to define a new versatile Rand index. Thanks to its flexibility, this index outperforms the existing ones as shown by numerous experimental results on both synthetic and real data. Moreover, by allowing to compare possibilistic partitions in a new way, the proposed concordance measure could become a privileged tool for many practitioners.

ACKNOWLEDGMENT

This work is partially supported by the French Ministry of Research through a PhD grant.

REFERENCES

- [1] J. Bezdek, *Pattern Recognition with fuzzy objective function algorithm*. Plenum Press, 1981.
- [2] D. Anderson, J. Bezdek, M. Popescu, and J. Keller, "Comparing fuzzy, probabilistic, and possibilistic partitions," *IEEE Transactions on Fuzzy Systems*, vol. 18, no. 5, pp. 906–918, 2010.
- [3] C. Borgelt, "Prototype-based classification and clustering," 2005, habilitation Thesis, Fakultät für Informatik der Otto von Guericke, Universität Magdeburg.
- [4] S. Bodjanova, "Comparison of fuzzy partitions based on their α -cuts," *Fuzzy Sets and Systems*, vol. 105, no. 1, pp. 99–112, 1999.
- [5] A. G. Di Nuovo and V. Catania, "On external measures for validation of fuzzy partitions," in *Lecture Notes in Computer Science* 4529, 2007, pp. 491–501.
- [6] I. Charon, L. Denoeud, A. Guenoche, and O. Hudry, "Maximum transfer distance between partitions," *Journal of Classification*, vol. 23, no. 1, pp. 103–121, June 2006.

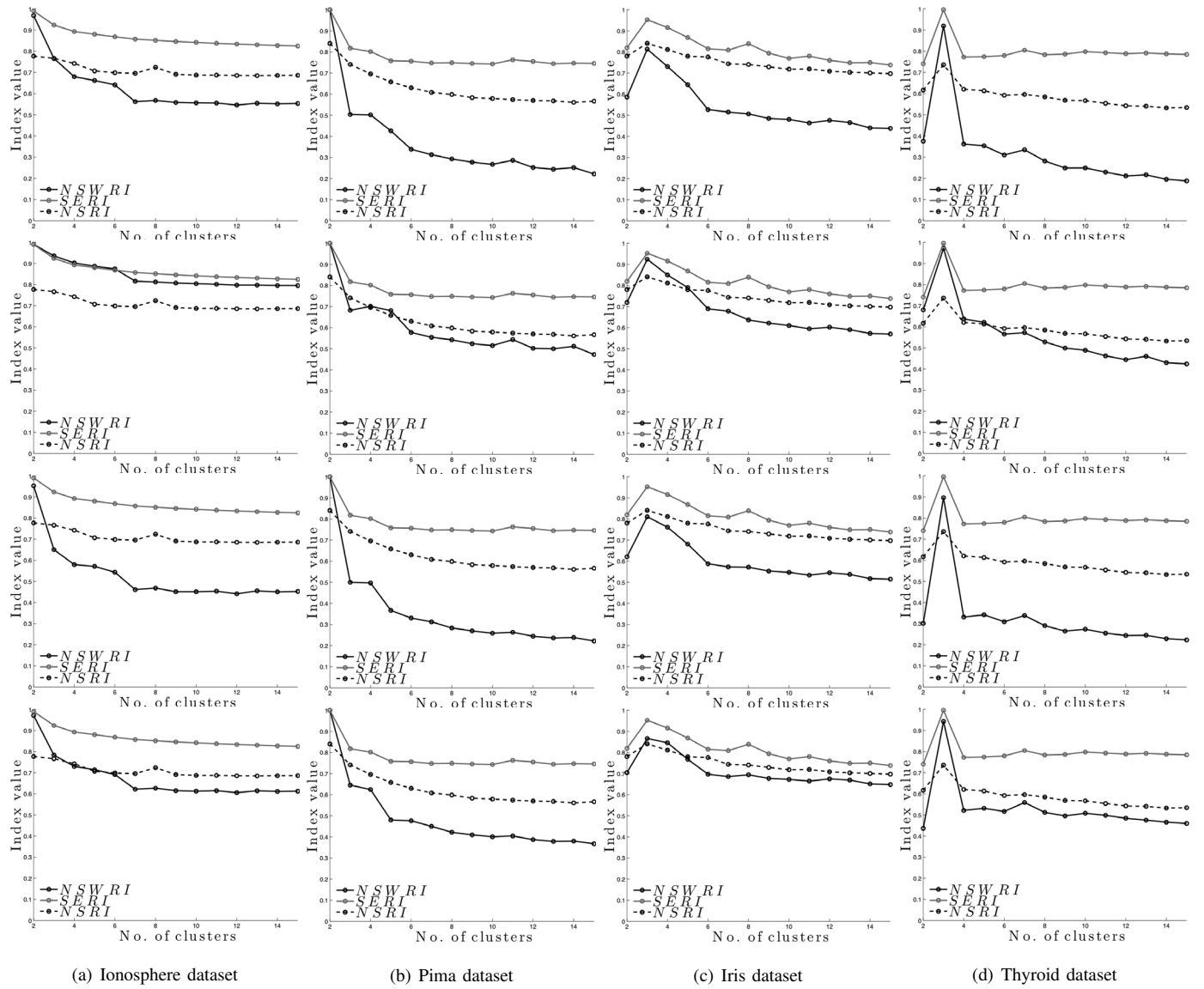


Fig. 6. $SERI(R, Q_c)$, $NSRI(R, Q_c)$, $NFRI(R, Q_c)$ and $NSWRI(R, Q_c)$ for four datasets, with $\{\top_P, w_T, 0.1, 0.1, 0\}$, $\{\top_P, w_T, 0.1, 0.8, 0\}$, $\{\top_P, w_T, 0.1, 0.1, 1\}$ and $\{\top_P, w_T, 0.1, 0.8, 1\}$

- [7] A. Albatineh, M. Niewiadomska-Bugaj, and D. Mihalko, "On similarity indices and correction for chance agreement," *Journal of Classification*, vol. 23, pp. 301–313, 2006.
- [8] C. Borgelt, "Resampling for fuzzy clustering," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 15, no. 5, pp. 595–614, 2007.
- [9] R. Brouwer, "Extending the rand, adjusted rand and jaccard indices to fuzzy partitions," *Journal of Intelligent Information Systems*, vol. 32, no. 3, pp. 213–235, 2009.
- [10] R. Quéré, H. L. Capitaine, N. Fraisseix, and C. Frélicot, "On normalizing fuzzy coincidence matrices to compare fuzzy and/or possibilistic partitions with the rand index," in *10th. IEEE International Conference on Data Mining*, 2010, pp. 977–982.
- [11] E. Hüllermeier and M. Rifqi, "A fuzzy variant of the rand index for comparing clustering structures," in *13th International Fuzzy Systems Association World Congress*, 2009, pp. 1294–1298.
- [12] M. Kendall and A. Stuart, *The Advanced Theory of Statistics*. Griffin, 1981, vol. 2.
- [13] E. Klement, R. Mesiar, and E. Pap, *Triangular Norms*, T. in Logic, Ed. Kluwer Academic, 2000.
- [14] A. Asuncion and D. Newman, "UCI machine learning repository," 2007. [Online]. Available: <http://www.ics.uci.edu/~mllearn/MLRepository.html>
- [15] J. Bezdek, J. Keller, R. Krishnapuram, and N. Pal, *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*. Kluwer Academic, 1999.