IJDAR (2011) 14:87–100
DOI 10.1007/s10032-010-0125-5

ORIGINAL PAPER

A protocol to characterize the descriptive power and the complementarity of shape descriptors

Muriel Visani · Oriol Ramos Terrades · Salvatore Tabbone

Received: 21 November 2009 / Revised: 2 June 2010 / Accepted: 5 August 2010 / Published online: 23 September 2010

Abstract Most document analysis applications rely on the extraction of shape descriptors, which may be grouped into different categories, each category having its own advantages and drawbacks (O.R. Terrades et al. in Proceedings of ICDAR'07, pp. 227–231, 2007). In order to improve the richness of their description, many authors choose to combine multiple descriptors. Yet, most of the authors who propose a new descriptor content themselves with comparing its performance to the performance of a set of single state-of-the-art descriptors in a specific applicative context (*e.g.* symbol recognition, symbol spotting…). This results in a proliferation of the shape descriptors proposed in the literature. In this article, we propose an innovative protocol, the originality of which is to be as independent of the final application as possible and which relies on new quantitative and qualitative measures. We introduce two types of measures: while the measures of the first type are intended to characterize the descriptive power (in terms of uniqueness, distinctiveness and robustness towards noise) of a descriptor, the second type of measures characterizes the complementarity between multiple descriptors. Characterizing upstream the complementarity of shape descriptors is an alternative to the usual approach where the descriptors to be combined are selected by trial and error, considering the performance characteristics of the overall system. To illustrate the contribution of this protocol, we performed experimental studies using a set of descriptors and a set of symbols which are widely used by the community namely ART and SC descriptors and the GREC 2003 database.

Keywords Document analysis · Shape descriptors · Symbol description · Performance characterization · Complementarity analysis

1 Introduction

Over the last decades, there has been a growing interest about performance evaluation in the domain of graphics recognition. Many contests have been organized, concerning raster-to-vector conversion [2–4], arc segmentation [5] and symbol recognition [6,7]. Most symbol recognition methods rely on a two-step procedure: (1) symbol description (representation) by extracting a feature vector with one (or more) descriptor(s) and (2) supervised classification of the symbols to recognize, based on their feature vectors. Several shape descriptors have been proposed in the literature [8–10] and most of them have been applied to the task of symbol recognition.

This paper is focused on the first step of symbol description, which is a crucial step that may be used for many other tasks in the field of document analysis (symbol spotting…). Ramos et al. have introduced in [1] a taxonomy of the different shape descriptors frequently used for symbol representation. The new categorization they propose is made according to the properties of the different shape descriptors, pointing out their strengths and weaknesses. One of the main objectives of their work is to facilitate, for a given application, the choice of the best descriptor in that context.

Work supported by the Juan de la Cierva programme and the MITTRAL project (TIN2009-14633-C03-01).

M. Visani
L3I, University of La Rochelle, La Rochelle Cedex 1, France
e-mail: muriel.visani@univ-lr.fr

O. R. Terrades
Instituto Tecnológico de Informática,
Universidad Politécnica de Valencia, Valencia, Spain
e-mail: oriolrt@iti.upv.es

S. Tabbone (✉)
LORIA, University of Nancy 2, Vandoeuvre-les-Nancy, France
e-mail: salvatore.tabbone@loria.fr

However, when considering a problem of symbol recognition, selecting the descriptor which is best suited for a given type of symbol and/or noise can be a hard, or even an impossible, task. Instead, one may be interested in combining different descriptors from different categories in order to benefit from the advantages of all the descriptors to be combined. The combination may be performed at the level of the descriptor (early fusion) or at the level of the classifiers (late fusion). Early fusion is usually implicitly done with powerful classifiers like neural networks [11], boosting classifiers [12,13] and Support Vector Machines [14]. In those methods, descriptors are extracted and concatenated as a single feature vector. Later on, during the training process, each classifier combines the features from the different descriptors. However, for general applications where the number of classes is high and the symbols to recognize can be counted by thousands, these expert classifiers reach their limits as their performance may decrease drastically. In this case, late fusion schemes where the combination is performed at the level of the classifier are generally preferred [15]. Late fusion methods have been applied to shape descriptors for symbol recognition [16,17]. Even so, in these papers, the performance characteristics of the descriptors in terms of descriptive power were not evaluated (only the performance for recognition was studied). Additionally, the complementarity of the descriptors to be combined was not investigated upstream, even though it may be very useful when choosing the set of descriptors to be combined and the combination scheme which is best suited to this particular set of descriptors.

To the best of our knowledge, very few works have been proposed in the literature concerning the evaluation of the performance characteristics of symbol descriptors, most evaluations being focused on the final application. A methodology for characterizing the performance of shape descriptors for symbol recognition has been proposed in [18]. This paper additionally provides a general discussion concerning the main difficulties and problems one may be faced with when setting the data, evaluation metrics and evaluation protocol, to characterize the performance of a symbol recognition method. Delalandre et al. [19] propose a solution to generate ground-truth based on a system that builds synthetic graphical documents. In [20], two main performance characterization metrics have been proposed, but we will see that these measures have several drawbacks that need to be completed (see Sect. 3). Jouili et al. [21] propose a performance evaluation for symbol recognition based on graph matching measures. This evaluation is essentially quantitative, based on precision and recall rates.

In this paper, we propose an experimental protocol and both qualitative and quantitative measures for characterizing the descriptive power and the complementarity of different shape descriptors for symbol description. This methodology is as independent of the final application (symbol spotting, recognition) as possible. Contrary to the above-mentioned performance evaluation methodologies, we do not consider any classifier; at most we consider some dissimilarity or distance measure and the nearest neighbour rule, to characterize for instance the uniqueness and distinctiveness of a given descriptor. We introduce an innovative protocol and two types of measures: while the measures of the first type are intended to characterize the descriptive power (in terms of uniqueness, distinctiveness and robustness towards noise) of a descriptor, the second type of measures characterize the complementarity between multiple descriptors. Concerning the measures of the first type, we first recall the definitions of confusion matrices, recognition rate, precision, recall and Cumulative Match Characteristics (CMC) curves. Even though some of these measures are already used by many researchers in our community, our contribution here is that we link them to the notions of distinctiveness and uniqueness. Second, we introduce two measures that are original in the field of document analysis. These two measures are respectively the tolerance intervals, which characterize the robustness of descriptors towards noise, and a qualitative measure based on an analogy with Doddington's zoo, widely known in the field of biometrics, characterizing the symmetries in the confusions. Concerning the measures of the second type, we introduce original measures to characterize upstream the complementarity between multiple descriptors. These measures constitute an alternative to the usual approach where the descriptors to be combined are selected by trial and error, considering the performance characteristics of the overall system. It may also be helpful when choosing the best combination scheme for a given set of descriptors.

To illustrate our methodology, we present a case study using two well-known statistical descriptors: the Angular Radial Transform (ART) descriptor [22], based on region pixel values and the Shape Context (SC) [23] descriptor, based on contours. We use noisy versions of the GREC 2003 database (*cf.* Fig. 4), which is well known and widely used by researchers working in the field of document analysis. It has to be noted that with adequate dissimilarity or distance measures (*e.g.* edit distance) our methodology can also be applied to structural descriptors. Moreover, the proposed framework may be further used for characterizing the complexity of any symbol database.

The paper is organized as follows. In Sect. 2, some innovative measures for evaluating the descriptive power of different shape descriptors, their robustness towards noise and their complementarity are proposed. In Sect. 3, we propose an experimental protocol and perform experimental results using ART and SC descriptors on the GREC 2003 database. These results are analysed to highlight the interest of using the proposed protocol and measures. While Sect. 4 provides a discussion about the measures we propose, Sect. 5 concludes this paper and presents the future work.

2 Evaluating the descriptive power of the descriptors and the complementarity between descriptors

In many applications such as symbol spotting, symbol recognition, the richness of a descriptor is related to its ability to group the different occurrences of one given symbol (uniqueness of the representation) and to discriminate them from other symbols (distinctiveness). In this direction, Valveny et al. have proposed in [20] two measures characterizing respectively the uniqueness and the distinctiveness of a shape descriptor: homogeneity and separability. While homogeneity is based on the distances between the descriptions of different occurrences of one symbol, separability is based on the distances between descriptions of different symbols. In this work, a good descriptor is characterized by high values of both homogeneity and separability. These measures are generic and may be used in many applicative contexts where a distance matrix between all the elements of the database can be computed. However, they have three main drawbacks. First, it is difficult to fix the thresholds which are necessary to characterize high values of these measures, since the distributions of the distances between feature vectors vary a lot from one database to another. Second, in many applications, we have a model image (which may be considered as the original symbol) and noisy versions of this model that we need to confront to all the models in the database. These two categories of images (models and noisy symbols) have to be considered separately, which is not the case with homogeneity and separability measures. Indeed, they rely on a distance matrix computed between all the symbols in the database, whatever their type. Third, in general the confusions between symbols are not symmetric (*e.g.* symbol 1 may be confused with symbol 2 and not the opposite). And neither homogeneity nor separability characterizes the symmetry of the confusions. Therefore, homogeneity and separability, which provide a coarse characterization of the richness of the different descriptors, have to be completed by other measures which overcome the drawbacks listed earlier.

To conceive such measures (which will be defined in Sects. 2.3 to 2.8), we first need to define more precisely the concepts of uniqueness and distinctiveness (see Sect. 2.1) and further to propose a protocol which is independent of the final application (see Sect. 2.2).

2.1 Definitions

Let us focus on the very conventional case where, for each symbol i in the database (with $i = 1, \ldots, c$), we have in the descriptor's representation space a symbol model S_i and n_i noisy versions of this model $\hat{S_i}^j$ (with $j = 1, \ldots, n_i$). In this case we can consider that a descriptor provides a perfect representation of a given symbol i when:

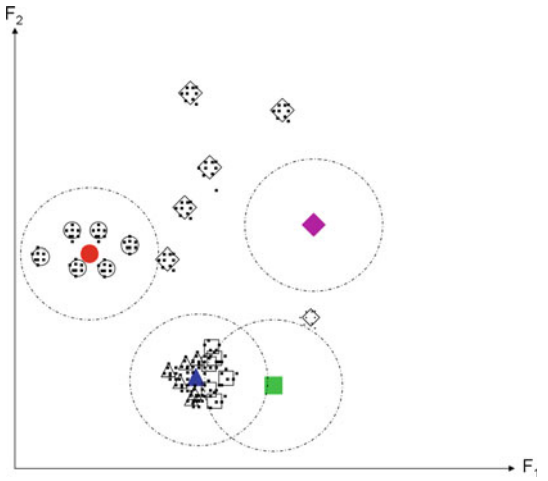

Fig. 1 Adaptation of Doddington's zoo (see Sect. 2.7) in our context. F_1 and F_2 are the two features of the descriptor's feature vector. In this example, we consider the Euclidean distance. The *filled shapes* are the models, each model having six noisy versions (*empty shapes* where noise is represented by *dots*). *Dashed circles* are situated at a distance θ from the models. The symbol *"circle"* is a sheep, the *"triangle"* is a lamb, the *"square"* is a wolf and the *"rhombus"* is a goat

– the representation of i is unique, *i.e.* feature vectors of noisy versions $\hat{S_i}^j$ of S_i are closer to S_i than to any other symbol model S_l (with $l \neq i$). Let us denote by d the distance (or dissimilarity measure) of interest. The representation of i may be considered as perfectly unique when:

$$\forall j = 1 \ldots n_i,\ \forall l = 1 \ldots c \ st\ l \neq i,\quad d(\hat{S_i}^j, S_l) > d(\hat{S_i}^j, S_i) \tag{1}$$

– the representation of i is distinctive, *i.e.* S_i is closer to its noisy versions $\hat{S_i}^j$ than to noisy versions $\hat{S_l}^m$ of other models S_l:

$$\forall l = 1 \ldots c \ st\ l \neq i,\ \forall m = 1 \ldots n_l,\quad d(\hat{S_l}^m, S_i) > \max_{j=1 \ldots n_i} d(\hat{S_i}^j, S_i) \tag{2}$$

Let us introduce the following notation: $NN(\hat{S}_i^j)$ being the nearest model of $\hat{S}_i^j$ in the descriptor's space (*i.e.* the result of the 1-nearest neighbour rule). The definitions of the uniqueness (see Eq. 1) and distinctiveness (see Eq. 2) of the symbol i may be respectively reformulated as follows:

$$\forall j = 1 \ldots n_i,\ NN(\hat{S}_i^j) = S_i \tag{3}$$

$$\forall l = 1 \ldots c \ st\ l \neq i,\ \forall m = 1 \ldots n_l,\ NN(\hat{S_l}^m) \neq S_i \tag{4}$$

For instance, the symbol "circle" in Fig. 1 is characterized by both perfect uniqueness and distinctiveness, while the symbol "triangle" is perfectly unique but not distinctive.

Equations 3 and 4 illustrate the notions of uniqueness and distinctiveness. However, they are not very useful in practice, because they only characterize perfect levels of uniqueness and distinctiveness. Relaxing them in order to allow the characterization of different levels of uniqueness and distinctiveness would require the introduction of additional parameters such as thresholds applied on the distance values. These parameters are difficult to settle in practice. In order to characterize the uniqueness and distinctiveness of a given descriptor, we therefore need to define a specific methodology.

2.2 Protocol

In order to characterize efficiently the descriptive power of different shape descriptors in terms of uniqueness and distinctiveness, we introduce the following protocol, which is independent of the final application (spotting, recognition…). First, the distances between all the noisy symbols and all the models are computed. Second, we apply the k nearest neighbour rule (kNN) in order to associate with each noisy symbol its k nearest models in the descriptor's representation space. Then, we can compute, for each descriptor, measures characterizing its descriptive power and robustness towards noise. Additionally, the complementarity between multiple descriptors may be measured.

In the remaining part of this section, we present two types of measures. The measures of the first type, introduced in Sects. 2.3 to 2.7, characterize indirectly the levels of uniqueness and/or distinctiveness (which are too parameter-dependent to be computed directly) of a single descriptor. We first recall in Sects. 2.3 to 2.5 the definitions of confusion matrices, recognition rate, precision, recall and Cumulative Match Characteristics (CMC) curves. Even though some of these measures are already used by many researchers in our community, our contribution here is that we link them to the notions of distinctiveness and uniqueness. Second, we introduce two measures that are original in the field of document analysis. These two measures are respectively the tolerance intervals (see Sect. 2.6), characterizing the robustness of descriptors towards noise, and a qualitative measure based on an analogy with Doddington's zoo, widely known in the field of biometrics, characterizing the symmetries in the confusions (see Sect. 2.7). Concerning the measures of the second type, we introduce in Sect. 2.8 original measures to characterize upstream the complementarity between multiple descriptors. These measures constitute an alternative to the usual approach where the descriptors to be combined are selected by trial and error, considering the performance characteristics of the overall system. It may also be helpful when choosing the best combination scheme for a given set of descriptors.

Table 1 A contingency matrix M. n_{il} is the number of noisy versions $\hat{S}_i^j$ of symbol i which nearest model is S_l in the representation space of the studied descriptor

$n_{i\cdot} - n_{\cdot l}$	$n_{\cdot 1}$	$n_{\cdot 2}$	…	$n_{\cdot c}$
$n_{1\cdot}$	n_{11}	n_{12}	…	n_{1c}
$n_{2\cdot}$	n_{21}	n_{22}	…	n_{2c}
⋮	⋮	⋮	⋱	⋮
$n_{c\cdot}$	n_{c1}	n_{c2}	…	n_{cc}

2.3 Confusion matrix

A confusion matrix M is a quantitative measure characterizing the descriptive power of a given descriptor. It is a contingency matrix computed from the array of distances between the descriptions of the noisy symbols and the models. This matrix contains c rows and c columns, where c is the number of models (see Table 1).The value in the cell n_{il} of the confusion matrix is the number of noisy versions $\hat{S}_i^j$ of symbol i which nearest model is S_l in the descriptor's representation space:

$$n_{il} = \sum_{j=1}^{n_i} \delta\left(NN(\hat{S}_i^j) = S_l\right) \tag{5}$$

with $NN(\hat{S}_i^j)$ being the model which is the nearest to $\hat{S}_i^j$ (*i.e.* the result of the $1NN$ rule following our protocol) and $\delta\left(NN(\hat{S}_i^j) = S_l\right) = 1$ if the nearest model of $\hat{S}_i^j$ is S_l, 0 otherwise. If we denote by $n = \sum_{i=1}^{c}\sum_{l=1}^{c} n_{il}$ the total number of noisy symbols, the descriptor may be considered as perfectly describing the database when the confusion matrix is diagonal (*i.e.* $\text{trace}(M) = n$).

Even if the confusion matrix is in general defined by using the 1-nearest neighbour rule (see Eq. 5), we can characterize the behaviour of the descriptor in an enlarged neighbourhood of the noisy symbol by considering confusion matrices $M(k)$ associated with higher ranks k, by defining:

$$n_{il}(k) = \sum_{j=1}^{n_i} \delta\left(kNN(\hat{S}_i^j) = S_l\right) \tag{6}$$

where $k\text{NN}(\hat{S}_i^j)$ is the kth nearest model to $\hat{S}_i^j$ (*i.e.* the result of the kNN rule following our protocol). We can note that the confusion matrix M shown in Table 1 is the same matrix as matrix $M(1)$, while matrix $M(k)$ with $k > 1$ may be obtained by replacing the values n_{il} in Table 1 with the $n_{il}(k)$ defined in Eq. 6. In the remaining part of this article, we consider by default confusion matrices at rank $k = 1$, unless we explicitly specify that we consider higher ranks $k > 1$ (for CMC curves for example, see Sect. 2.5).

2.4 Recognition rate, precision and recall

This section is dedicated to quantitative measures characterizing the richness of a given descriptor in terms of uniqueness and/or distinctiveness.

First of all, the recognition rate (RR) provides the percentage of noisy symbols such that their nearest model is the "good" one. It is computed from the confusion matrix (see Table 1 and Eq. 5) as follows:

$$RR = \frac{\text{trace}(M)}{n} \tag{7}$$

It has to be noted that we can also compute the recognition rates at any rank $k > 1$ by using the confusion matrix $M(k)$ (see Eq. 6):

$$RR(k) = \frac{\text{trace}(M(k))}{n} \tag{8}$$

Hereafter, we consider by default the recognition rate at rank $k = 1$, unless we explicitly specify that we consider higher ranks of k (for CMC curves for example, see Sect. 2.5).

While the recognition rate gives some information about the descriptive power of a descriptor on the whole database, one may be interested in the behaviour of the descriptor for a particular symbol. We can note that a symbol i which is badly described in the descriptor's representation space is associated with a large number of extra-diagonal elements n_{il} and/or n_{li} (with $l \neq i$) in the confusion matrix (see Eq. 5). A low distinctiveness for symbol i is characterized by high values of the column cells n_{li}. On the other hand, a low uniqueness for symbol i is characterized by high values of the row cells values n_{il} (see the definitions of distinctiveness and uniqueness given in Sect. 2.1). The level of distinctiveness and uniqueness for a given symbol i may therefore be respectively measured by using precision $P(i)$ and recall $R(i)$, where:

$$P(i) = \frac{n_{\text{ii}}}{\sum_{l=1}^{c} n_{\text{li}}} \tag{9}$$

$$R(i) = \frac{n_{\text{ii}}}{\sum_{l=1}^{c} n_{\text{il}}} \tag{10}$$

To characterize the distinctiveness and uniqueness for the whole dataset, one may consider only the scalar value corresponding to the average precision and/or recall among all the symbols $i = 1 \ldots c$:

$$P = \frac{1}{c} \sum_{i=1}^{c} \left(\frac{n_{\text{ii}}}{\sum_{l=1}^{c} n_{\text{li}}} \right) \tag{11}$$

$$R = \frac{1}{c} \sum_{i=1}^{c} \left(\frac{n_{\text{ii}}}{\sum_{l=1}^{c} n_{\text{il}}} \right) \tag{12}$$

We can note that, in the special case where the number n_i of noisy versions of symbol i is the same for all the c symbols i, the mean recall equals the recognition rate (given that $\sum_{l=1}^{c} n_{\text{il}} = n_i$, by construction).

2.5 Cumulative match characteristics curve

In order to characterize the behaviour of the descriptor in an enlarged neighbourhood of the symbols to describe (not only considering the nearest model), we may consider the recognition rates at ranks $k > 1$. The Cumulative Match Characteristic (CMC) curves are most widely used for evaluating the performance characteristics of semi-automated recognition systems where N candidates are proposed to a (often human) supervisor, the role of the supervisor being to select the good candidate. Such curves are useful to quickly visualize the cumulated recognition rates at different ranks k:

$$CMC(k) = \sum_{r=1}^{k} RR(r) \tag{13}$$

where $RR(k)$ is the recognition rate at rank $k \geq 1$ (see Eq. 8). For an example of a CMC curve see Fig. 7. If the CMC curve reaches a sufficiently high value at a rank k being tractable for the supervisor, then the semi-automated system is considered as effective.

2.6 Characterization of the Robustness towards noise

In this section, we present the Tolerance Interval, which has been defined in the context of face recognition in [25] in order to characterize the robustness of descriptors towards noise. Tolerance Intervals may be calculated in the case where the amount of noise is controlled by one parameter ω (*i.e.* in general when the noise is synthetically added to the images). For example, for Gaussian white noise the parameter is the standard deviation: $\omega = \sigma$. The recognition rate is computed as a function of the value of the noise parameter. A Tolerance Interval (TI) at $p\%$ may be defined as the range of values of parameter ω such that the recognition rate RR remains greater than $1 - p/100$. Examples of Tolerance Intervals are given in Table 3. For a fixed p, the larger is the Tolerance Interval, the more robust is the descriptor. Tolerance Intervals characterize in a compact way the robustness of a descriptor towards a given type of noise; they may be very helpful when choosing the descriptor which is best suited to a specific kind of noise.

2.7 The zoo qualitative characterization

All the previous measures are intrinsically quantitative. In particular, we have shown how the precision and recall measures may be used at the symbol level to characterize the asymmetries in the confusions of a single descriptor for a given symbol (*e.g.* symbol i may be confused with others

and not the opposite). However, when trying to compare the descriptive power of multiple descriptors for a given symbol, it is difficult to consider jointly multiple precision and recall values. That is why we introduce a qualitative measure based on the definition of categories of symbols. Our categorization is inspired from Doddington et al.'s terminology [26]. This terminology was first introduced in the field of speaker recognition for biometrics. Figure 1 provides examples of the different categories. We give the original definitions by Doddington et al., followed by their adaptations in our context.

- "In our model, sheep dominate the population and systems perform nominally well for them". In our context, sheep are the symbols which are well represented by the descriptor, with both high uniqueness and distinctiveness. Therefore we can define a sheep i as a symbol associated with high values of both precision and recall. In Fig. 1 the symbol "circle" is a sheep;
- "Lambs, in our model, are those speakers who are particularly easy to imitate". In our context, lambs are symbols characterized by a low distinctiveness and therefore associated with a low precision. In Fig. 1 the symbol "triangle" is a lamb;
- "Wolves, in our model, are those speakers who are particularly successful at imitating other speakers". In our context, lambs are symbols characterized by a low uniqueness and thus associated with a low recall. In Fig. 1 the symbol "square" is a wolf;
- "Goats tend to adversely affect the performance of systems by accounting for a disproportionate share of the missed detections". A goat is a model such that its noisy versions are in general farther than a given threshold θ to all the models (dotted circles in Fig. 1, in the case of an Euclidean distance). In Fig. 1 the symbol "rhombus" is a goat.

While in the case of sheep the descriptive power of the descriptor may be considered as satisfactory, in the case of goats the descriptive power is low. But, for a given symbol, the behaviours of two different descriptors may differ. For instance, symbol i may be a sheep with descriptor $D1$ and a wolf with descriptor $D2$. At the level of the whole database, these behaviours may be complementary, *e.g.* in the case where symbol i is a sheep with descriptor $D1$ and not a sheep with descriptor $D2$, and vice-versa for symbol j. In that case, the description may be improved by combining the two descriptors, instead of considering a single descriptor. That is the reason why, in the following section, we introduce measures to characterize the complementarity of different descriptors.

2.8 Ensemble measures characterizing the complementarity of different descriptors

In this section, we introduce quantitative measures of the complementarity between different descriptors. We can note that, in most cases, confronting the confusion matrices of different descriptors does not help to characterize their complementarity. Indeed, while the confusion matrix provides information at the symbol level, the complementarity occurs at the level of the noisy image. For instance, when the confusion matrix says that, for a given symbol i, only 15 noisy versions over 30 are well described by descriptor $D1$ and by descriptor $D2$, these 15 well-described images may be the same (in this case the two descriptors are not complementary at all for this symbol) or totally distinct (in this case the two descriptors are perfectly complementary for this symbol), but from the confusion matrix we cannot guess which case we are dealing with. That is why we introduce the following complementarity measures that may be computed at different ranks $k \geq 1$, with $\hat{S}_i^j$ being a noisy version of the original model S_i and $kNN_D(\hat{S}_i^j)$ the kth nearest model of $\hat{S}_i^j$ in the representation domain associated with the descriptor D:

$$U(k) = \sum_{i=1}^{c} \sum_{j=1}^{n_i} \delta\left(\left\{kNN_{D1}(\hat{S}_i^j) = S_i\right\} \cup \left\{kNN_{D2}(\hat{S}_i^j) = S_i\right\}\right) \tag{14}$$

$$I(k) = \sum_{i=1}^{c} \sum_{j=1}^{n_i} \delta\left(\left\{kNN_{D1}(\hat{S}_i^j) = S_i\right\} \cap \left\{kNN_{D2}(\hat{S}_i^j) = S_i\right\}\right) \tag{15}$$

$$I_{D1}(k) = \sum_{i=1}^{c} \sum_{j=1}^{n_i} \delta\left(\left\{kNN_{D1}(\hat{S}_i^j) = S_i\right\} \cap \left\{kNN_{D2}(\hat{S}_i^j) \neq S_i\right\}\right) \tag{16}$$

$$I_{D2}(k) = \sum_{i=1}^{c} \sum_{j=1}^{n_i} \delta\left(\left\{kNN_{D2}(\hat{S}_i^j) = S_i\right\} \cap \left\{kNN_{D1}(\hat{S}_i^j) \neq S_i\right\}\right) \tag{17}$$

$$C(k) = \sum_{i=1}^{c} \sum_{j=1}^{n_i} \delta\left(\left\{kNN_{D1}(\hat{S}_i^j) \neq S_i\right\} \cap \left\{kNN_{D2}(\hat{S}_i^j) \neq S_i\right\}\right) \tag{18}$$

The measure $U(k)$ is the number of noisy images such that their kth nearest model is the "good" one for at least one of the two descriptors $D1$ or $D2$. The measure $I(k)$ is the number of noisy images such that their kth nearest model is the "good" one for both descriptors $D1$ and $D2$. We can note that, by construction, $I(k) \leq U(k)$. The measure I_{D1} is the total number of noisy images such that their kth nearest model is the "good" one for descriptor $D1$ but not for $D2$ (and vice-versa for I_{D2}). We can note that $U(k) = I(k) + I_{D1}(k) + I_{D2}(k)$. Finally, $C(k)$ is the total number of noisy images such that their kth nearest model is not the "good" one, neither for descriptor $D1$ nor for $D2$. We can note that

$C(k) + U(k) = n$, where n is the total number of noisy symbols in the database. Figure 8 provides a visual example of such measures with descriptors ART and SC. Let us note the following relationship: $\frac{I(k)+I_D(k)}{n} = RR_D(k)$, where $RR_D(k)$ is the recognition rate at rank k (see Eq. 8) associated with descriptor D. Of course the measures given in Eqs. (14–18) can be directly extended with more than two descriptors.

The numbers of images which are well represented by one descriptor but not by the other one (*i.e.* I_{D1} and I_{D2}) allow us to quantify the complementarity of the two descriptors. In particular, the value of $U(1)$ is the maximal number of symbols that may be well-described at rank 1 (*i.e.* the objective value) when conceiving a strategy for selecting, for each symbol, the best descriptor for this symbol. The more the objective value $\frac{U(1)}{n}$ exceeds the maximal recognition rate of the two descriptors, the more complementary are these two descriptors. Characterizing the complementarity between descriptors is very interesting, as considering a combination of complementary descriptors may improve the richness of the description of the symbol compared to considering a single descriptor.

3 Experimental study

In this section, we perform an experimental study to illustrate the effectiveness of the measures we define in Sect. 2. The objective is to show how to use these measures for (1) comparing multiple descriptors in terms of descriptive power and noise robustness and (2) measuring the complementarity of multiple descriptors. For this purpose, we consider two well-known shape descriptors (that will be described in Sect. 3.1): ART and SC. The main objective here is not to characterize the performance of these two descriptors, but rather to illustrate the contribution brought to the community by our innovative protocol and measures. We selected ART and SC among the large variety of available shape descriptors for two main reasons. First, because of the paper size limit, we could not consider more than two descriptors. Second, ART and SC belong to different categories of shape descriptors (ART is 2D while SC is a 1D descriptor based on contours). This fact certainly makes them complementary to some extent, which is interesting to illustrate our complementarity measure.

This section is organized as follows. A brief description of the considered descriptors is given in Sect. 3.1. Next, the databases we use and their features are detailed in Sect. 3.2. Then, in Sect. 3.3 we compute and analyse the measures proposed in Sect. 2.

3.1 Statistical shape descriptors

Among the various shape descriptors that have been proposed in the literature [8–10], we selected in this paper two well-known statistical descriptors which are essentially different in their primitives: Angular Radial Transform (ART) [22] and Shape Context (SC) [23]. While ART is based on a 2D primitive (region inside the image) and provides a polar-based feature vector, SC is a 1D primitive (relying on the extraction of shape contours) resulting in a histogram-based feature vector.

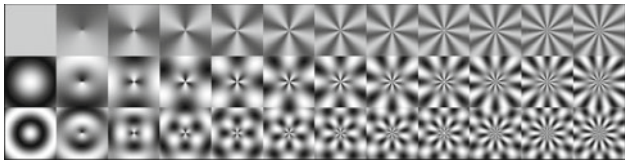

Fig. 2 Real parts of the basis functions of the descriptor ART, for n from 0 to $N = 2$ (*from top to bottom*) and m from 0 to $M = 11$ (*from left to right*). It has to be noted that their imaginary parts are similar except the quadrature phase difference

3.1.1 The ART descriptor

The ART descriptor [22] is the result of a complex 2D transform defined on a unit circle using polar coordinates. More precisely, ART coefficients are defined by the projection of the original image represented in polar coordinates on a basis of orthogonal complex functions $V_{n,m}(\rho, \theta) = A_m(\theta)R_n(\rho)$ (*cf.* Fig. 2). These basis functions are defined by multiplying a radial function R_n of parameter n by an angular function A_m of parameter m, the pair of parameters (n, m) defining the order of the coefficient $F_{n,m}$.

Invariance to similarity transforms is achieved by (1) using an exponential functional in the angular function (to get invariance towards rotations) and (2) centring and scaling the shape image before computing the coefficients (to achieve invariance towards scale and translation).

Finally, the distance between shapes is measured by the Manhattan (L_1) distance.

3.1.2 The SC descriptor

The SC descriptor [23] is based on relative spatial locations between some points extracted from the contours of the shape to analyse. Figure 3 illustrates its underlying principle. The shape to be described is represented by a discrete set of points extracted from the external and internal contours of the shape.

This descriptor can be considered invariant to scaling if the background is not too complex, since the radial distances are normalized by the average distance between all the pairs of points of the shape. In addition, it is invariant towards translation and can easily be made invariant towards rotation. And, given that the SC descriptor provides coarse information extracted from the whole shape, it is relatively robust towards occlusions.

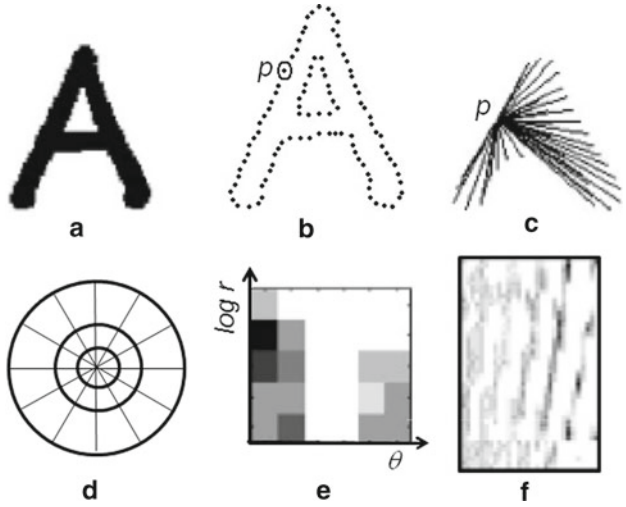

Fig. 3 Principle of the SC descriptor. **a** shape to be described, **b** contour points sampled from this shape, **c** set of vectors associated with the reference point p extracted from the contour, **d** classes (*bins*) used for the histograms, **e** histogram of the coordinates of the vectors shown in c (*i.e.* shape context of point p), **f** set of shape contexts of the shape shown in a

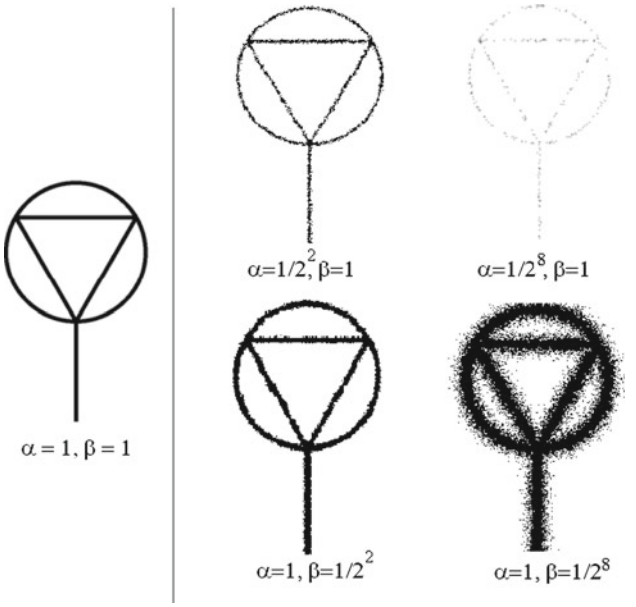

Fig. 4 Symbol # 86 with different levels of noise (from left to right, *unnoisy symbol*, first row: *noisy symbol* with noise α at levels $N = 2$ and $N = 8$, second row: noisy symbol with noise β at levels $N = 2$ and $N = 8$

Even though the χ^2 distance was initially used to compare shape contexts from different symbols [23], more recently numerous authors have chosen to consider shape contexts as feature vectors and to compare them by using the L_2 (Euclidean) distance [24].

3.2 Experimental protocol

In our experiments, we consider the GREC 2003 symbol database [6]. This database contains 150 models of symbols, which are used to generate noisy versions by applying the Kanungo algorithm [27]. The Kanungo noise is an additive noise applied to binary images; it is controlled by six parameters. Among these parameters, we chose to vary α and β, which simulate the presence of an ascending amount of noise in the image. When α decreases, the probability for a symbol pixel to be inverted and considered as a background pixel increases (which may be seen as some kind of "salt" noise), while when β decreases the probability to invert a background pixel as a symbol pixel increases ("pepper" noise). It has to be noted that these probabilities of inversion decrease according to the distance from the contour of the shape. Five databases, each one containing 30 random noisy versions of each of the 150 model symbols are generated for each $\alpha = \frac{1}{2^N}$, with N varying from 2 to 10 by a step of 2 (*cf.* Fig. 4). Five databases are constructed similarly by varying parameter β. At the end, we obtain a database containing the 150 model images (one image per symbol model) and 10 test databases, each one containing $30 \times 150 = 4,500$ noisy symbols.

In this experiment, we compare each noisy image from the 10 noisy databases to the database containing the model images. For this purpose, we use for each descriptor its associated distance (the Manhattan distance for ART and the Euclidean distance for SC) and the protocol presented in Sect. 2.2.

3.3 Experimental results

The analysis of the experimental results is in four steps. During the first step (Sect. 3.3.1), a coarse evaluation of the performance characteristics of the descriptors ART and SC is provided. For this coarse evaluation, we consider the usual performance measures (recognition rate, precision and recall, see Sect. 2.4). The second step (Sect. 3.3.2) is a comparison of the robustness of the two descriptors towards noise. For this comparison, we compute Tolerance Intervals (see Sect. 2.6) and we consider the two different types of Kanungo noise (α "salt" noise and β "pepper" noise). Then, a subset of databases (among the most noisy) are selected for the third step of the analysis (Sect. 3.3.3). The third step is a detailed analysis relying on the confusion matrices (see Sect. 2.4), the CMC curves (see Sect. 2.5), and the qualitative measures we introduce in Sect. 2.7. The fourth step, given in Sect. 3.3.4, is a study of the complementarity of ART and SC, based on the complementarity measures defined in Sect. 2.8.

Table 2 Mean recall and precision associated with the two descriptors on the databases with the levels of noise $\alpha = \frac{1}{2^N}$ and $\beta = \frac{1}{2^N}$ (with $N = 2, 4, 6, 8, 10$)

The values between brackets are the corresponding standard deviations. We can note that the mean recall equals the recognition rate because the number $n_i = 30$ of noisy versions is constant over all the symbols i (see Sect. 2.4)

N	Descriptor	Noise α		Noise β	
		Mean precision (SD)	Mean recall (SD)	Mean precision (SD)	Mean recall (SD)
2	ART	100% (0)	100% (0)	100% (0)	100% (0)
	SC	99.06% (0.059)	98.98% (0.076)	99.23% (0.05)	99.09% (0.066)
4	ART	94.69% (0.201)	96.09% (0.193)	100% (0)	100% (0)
	SC	99.14% (0.055)	98.87% (0.081)	98.99% (0.057)	98.80% (0.08)
6	ART	93.05% (0.232)	94.71% (0.217)	99.74% (0.032)	99.58% (0.052)
	SC	99.13% (0.059)	98.8% (0.092)	97.55% (0.09)	96.33% (0.148)
8	ART	90.60% (0.24)	90.53% (0.251)	89.46% (0.279)	91.87% (0.269)
	SC	98.814% (0.071)	98.76% (0.09)	71.96% (0.396)	66.33% (0.402)
10	ART	63.6% (0.335)	52.91% (0.339)	19.09% (0.355)	28.64% (0.447)
	SC	95.19% (0.089)	94.69% (0.121)	3.30% (0.151)	4.4% (0.168)

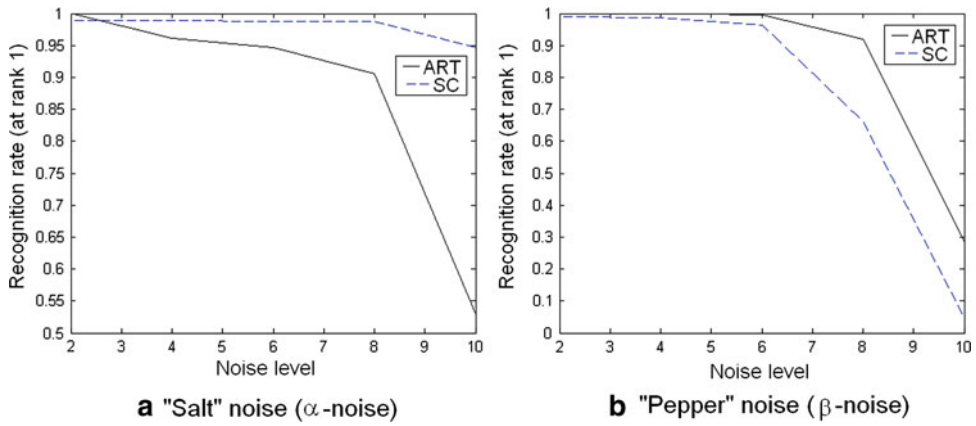

Fig. 5 Recognition rates RR (at rank 1) as a function of the level N of noise, with a noise of type α (*left*) and β (*right*)

3.3.1 *Overview of the performance characteristics of ART and SC*

The recognition rate (RR), depending on the type and amount of noise, is given in Table 2 and Fig. 5. We can note that, as explained in Sect. 2.4, the recognition rate equals the mean recall in our case where the number $n_i = 30$ of noisy versions is constant over all the symbols i. Table 2 also gives the mean precision (at rank $k = 1$ and the standard deviations of the recognition rate and mean precision. We can see that, for both descriptors and both types of noise, the quality of the description decreases when the amount of noise increases. We can also note that the decrease is more abrupt in the presence of β noise. For levels of noise $N > 2$, SC is superior to ART for "salt" noise (α-noise) while ART is superior to SC for "pepper" noise (β-noise). When $N \leq 2$ ART is always superior to SC, whatever kind of noise is applied.

We can easily understand why the performance of SC decreases drastically in the presence of "pepper" noise: SC is based on points sampled from the contour of the symbol (*cf.* Fig. 3). When pepper noise is added to the image, the shape contours are modified. In that case, the SC description is computed from inaccurate points and becomes imprecise. Conversely, "salt" noise has a thinning effect on the symbol. Therefore, when the α-noise increases, the amount of information available for computing ART is reduced, which makes ART description unstable (see the high standard deviation values in Table 2).

3.3.2 *Comparison of the Robustnesses of ART and SC towards noise*

From Table 2 and Fig. 5, we compute the Tolerance Intervals (TI) (described in Sect. 2.6) corresponding to descriptors ART and SC, towards α-noise and β-noise. The TIs at levels $p = 5\%$ and $p = 20\%$ are given in Table 3.

The results in Table 3 are consistent with the shapes of the curves in Fig. 5. In particular, in the presence of α-noise, the TI of ART is narrower than the TI of SC, which implies that SC is more robust than ART towards α-noise. For β-noise and descriptor SC, the fact that the TIs at $p = 5\%$ and $p = 20\%$ are both equal to $[1, 6]$ is due to a drastic drop

Table 3 Tolerance Intervals (TIs) of ART and SC towards Kanungo noise of type α and β

Type of noise	α-noise level N		β-noise level N	
	$p = 5\%$	$p = 20\%$	$p = 5\%$	$p = 20\%$
ART	[1, 4]	[1, 8]	[1, 6]	[1, 8]
SC	[1, 8]	[1, 10]	[1, 6]	[1, 6]

These TIs are computed by using the recognition rates given in Table 2.The noise levels N are such that $\alpha = \frac{1}{2^N}$, (respectively $\beta = \frac{1}{2^N}$)

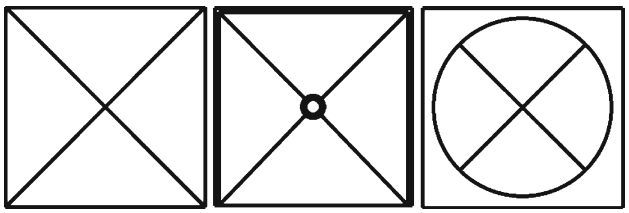

Fig. 6 From left to right: symbols 11, 87 and 125

in the recognition rates between the levels of noise $N = 6$ and $N = 8$. Table 3 shows the decrease in the quality of the description when the amount of noise is increased. In particular, none of the two descriptors is tolerant at $p = 5\%$ towards a noise of level $N = 10$ (neither for α-noise nor for β-noise). Consequently, the rest of this section is devoted to the detailed analysis of the results for the levels of noise $N = 6$ and $N = 8$ for α and β. Indeed, when applied on these databases, the behaviour of the descriptors is representative of the general case where the noise is not too strong and at least one of our two descriptors remains robust.

3.3.3 Detailed analysis of the performance characteristics of ART and SC

To provide a more detailed analysis of the descriptors' behaviours, we show in Table 4 some extracts of the confusion matrices (see Sect. 2.3) for symbols 11, 87 and 125 (see Fig. 6). We consider these particular symbols because, in the presence of β-noise (at levels 6 and 8), they are subject to confusions. Let us now focus on the database with noise β at level $N = 6$ (Table 4a and c). Among the 4,500 noisy symbols of the whole database, ART badly describes only 19 noisy symbols. All of these 19 poor descriptions come from confusions between symbols 87 and 125. On the other hand, the SC descriptor badly describes 165 noisy symbols over 4,500, among which 29 poor descriptions are due to confusions between symbols 11 and 87.

To go deeper into the analysis of the database with β-noise of level $N = 6$, let us consider additionally the precision and recall measures (see Sect. 2.4) associated with the symbols 11, 87 and 125 and the qualitative definitions introduced in Sect. 2.7. From this analysis, we conclude that

Table 4 Extracts of the confusion matrices of the descriptor ART on databases with noise β and levels (a) $N = 6$ and (b) $N = 8$ and of the descriptor SC on databases with noise β and levels (c) $N = 6$ and (d) $N = 8$

GT	Nearest model		
	11	87	125
(a) ART with noise β at level $N = 6$			
11	30	0	0
87	0	11	19
125	0	0	30
(b) ART with noise β at level $N = 8$			
11	30	0	0
87	0	0	30
125	0	0	3
(c) SC with noise β at level $N = 6$			
11	1	29	0
87	0	30	0
125	0	0	30
(d) SC with noise β at level $N = 8$			
11	0	22	8
87	0	12	18
125	0	2	28

- symbol 11 (respectively symbol 125) is a sheep for descriptor ART (resp. SC), as the quality of the description of this symbol is satisfying (indeed $P = 1$ and $R = 1$ for symbol 11 with ART and $P = 0.81$ and $R = 1$ for symbol 125 with SC);
- symbol 125 (respectively symbol 87) is a lamb for descriptor ART (resp. SC), as other symbols may be confused with it. Indeed, $P = 0.61$ for symbol 125 with ART and $P = 0.51$ for symbol 87 with SC;
- symbol 87 (respectively symbol 11) is a wolf for descriptor ART (resp. SC), as this symbol may be confused with other symbols. Indeed, $R = 0.37$ for symbol 87 with ART and $R = 0.03$ for symbol 11 with SC).

A preliminary statistical study has shown that the databases are very homogeneous for both descriptors [20]. This means that the number of goats is very reduced in this context. Therefore, we did not look for goats, which would have required the settlement of an additional parameter θ (see Sect. 2.7). The fact that symbol 11 is a wolf for symbol 87 in the presence of pepper noise for descriptor SC is easily understandable. Indeed, the pepper noise located at the centre of the cross in symbol 11 may be confused with the small circle at the centre of symbol 87 (see Fig. 6). With ART, 11 noisy occurrences of symbol 87 have the "good" model 87 as nearest neighbour, while the remaining 19 occurrences have the model 125 as nearest neighbour. This phenomenon

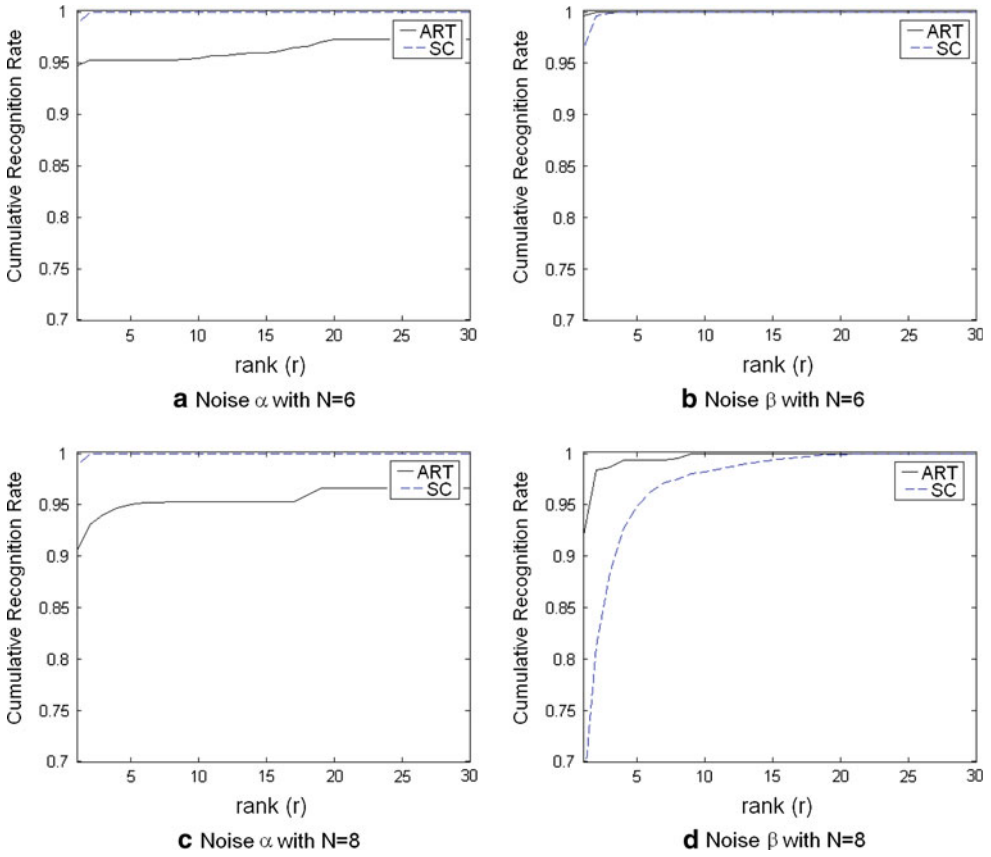

Fig. 7 The CMC curves associated with the databases (*first column*): $\alpha = \frac{1}{2^6}$ and $\alpha = \frac{1}{2^8}$ and (*second column*): $\beta = \frac{1}{2^6}$ and $\beta = \frac{1}{2^8}$

makes us guess that there is an overlap between the images of symbols 87 and 125 in the ART description space.

In addition, Table 4 shows that these dissymmetries in the confusion matrix increase when more pepper noise is added to the symbols (between levels $N = 6$ and $N = 8$). For instance, when the level of β-noise reaches $N = 8$, symbol 87 becomes a wolf for symbol 125 in the SC space, as the images of symbols 87 become closer to the model 125 than to the model 87.

From this point of view, characterizing the results of the descriptions not only at the first rank (*i.e.* using the nearest model), but at higher ranks (*i.e.* using the k nearest models) is very important. Indeed, in the case where there is an overlap between two symbols (*e.g.* in the case of symbols 87 and 125 for ART), the "good" model may be among the two nearest models but not necessarily the nearest one. And we can consider that, if the "good" model is among the two nearest models, the quality of the description is better than if the "good" model is farther. In other words, we can consider that a wolf model which is near its lamb (see Fig. 1) is better described than a goat. In order to take into account higher ranks $k > 1$, we consider the CMC curves (see Sect. 2.5) shown in Fig. 7. Figure 4a and c shows that most of the confusions of the descriptor SC at rank 1 with noise α (see Table 2) are solved at ranks 2 or 3. This means that the descriptive power of SC in the presence of α-noise is relatively good, because even the noisy symbols which are badly classified by using the $1NN$ rule are well classified when considering the $3NN$ rule. On the contrary, we can see from Fig. 4d that the descriptive power of SC is bad with β-noise of level $N = 8$, as the CMC curve of SC does not reach 100% recognition rate before rank $k > 20$.

From Table 4a and c, we can also note that, with noise β at level $N = 6$, while ART does not manage to discriminate between symbol 87 and symbol 125, SC perfectly discriminates between these two symbols. And vice-versa for symbols 11 and 87 in the same database. Therefore, we can consider that ART and SC are complementary for symbols 11, 87 and 125 in the particular database with noise β at level $N = 6$. Which means that combining ART and SC to obtain a single descriptor may improve (on average) the quality of our description, compared to a single descriptor. Now let us expand our study of the complementarity of ART and SC to the whole dataset.

3.3.4 Analysis of the complementarity of ART and SC

To end this experimental study, we measure the complementarity of ART and SC by the measures defined in

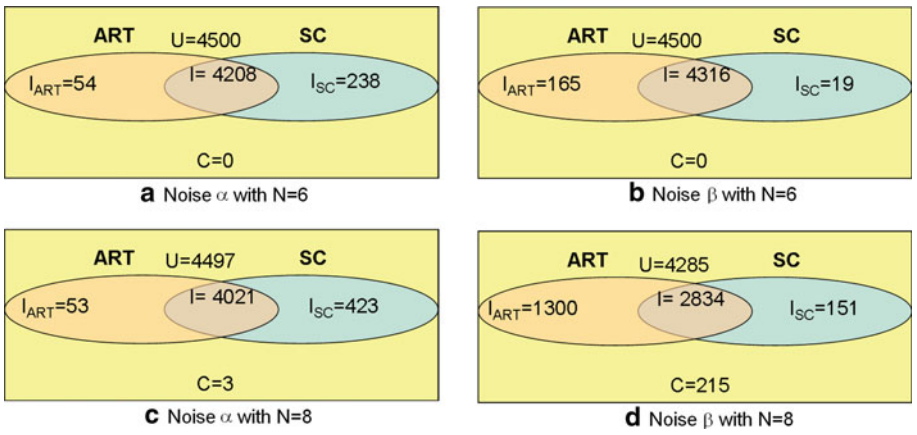

Fig. 8 Measures $U(k)$, $I(k)$, $I_{\mathrm{ART}}(k)$ and $I_{\mathrm{SC}}(k)$ (with $k = 1$, denoted as U, I, I_{ART} and I_{SC} for the sake of readability) with different levels of noise

Eqs. (14–18), where $D1$ = ART and $D2$ = SC. The results are given in Fig. 8.

From that figure we can see that both descriptors are really complementary on the four databases. For the two databases with a noise level $N = 6$, we can see that the two descriptors are perfectly complementary as $C = 0$ with the two types of noise (even if ART is superior to SC in the presence of β-noise while SC is superior to ART in the presence of α-noise). It means for instance that, for a task of recognition, any query symbol from these noisy databases may be correctly classified automatically by using the ART and SC descriptors. Provided an ideal adaptive descriptor selection method that would select, for each query symbol and each type of noise, the best performing descriptor for this particular symbol. With a noise level of $N = 8$, we can observe high values of both I_{ART} and I_{SC}, which means that combining ART and SC may enhance the average quality of description of the symbols, whatever type of noise is present in the images, compared to a single descriptor.

3.3.5 Conclusion of the experimental section

The results given in this section have shown that SC is superior to ART in the presence of "salt" noise while ART is superior to SC in the presence of "pepper" noise. Both descriptors are robust towards a small amount of noise but their performance decreases drastically when the amount of noise increases (especially with "pepper" noise). However, the SC descriptor remains more robust than ART with salt noise. We can also note that these two descriptors are complementary. Therefore, combining them may enhance the quality of description of a symbol compared to that of a single descriptor.

4 Discussion

Well-known and widely used evaluation measures such as the recognition rate (RR) and the Precision-Recall values (see Sect. 2.4) are very useful to measure whether a given descriptor is adapted to a particular context. They provide performance characteristics on a set of evaluation databases which are supposed to be representative of the images the system will find in a real environment. However, when no descriptor is superior to the others on all the databases (for different types of noise for example), then the issue of the usefulness of the evaluation measures proposed in Sects. 2.5 and 2.6 arises. Thereby, the Tolerance Interval quickly gives an idea of the robustness of descriptors towards noise, while the CMC curves characterize the quality of the description in the neighbourhood of the noisy symbols.

Nevertheless, the descriptive power of a given descriptor may vary from one symbol to another. Indeed, in any database and for any descriptor there are symbols which are well-described and others which are not (provided a database of sufficient complexity). In this context, using the qualitative measures introduced in Sect. 2.7 becomes useful, since it allows us to detect the overlapping symbols. The information given by these measures is equivalent to the information given by the CMC curves, but detailed for each symbol in the database, while the CMC remains general.

The complementarity measures (see Sect. 2.8) provide information about the benefit we can expect from the combination of multiple descriptors. Hence, the best configuration for a pair of descriptors is to be perfectly complementary, which is the case for ART and SC on the databases of noise level $N = 6$, for both α and β noise. Measuring upstream the complementarity of shape descriptors is an interesting alternative to the most widely used approach consisting in selecting the descriptors to be combined by trial and error, considering the performance characteristics of the overall system.

It has to be noted that the complementarity measures can also be used to characterize the complexity of a given database. Indeed, if we consider a well-chosen set of mutually complementary descriptors and that this set of descriptors gives poor results on a given database (*i.e.* the value of C is high), we can consider that this database is highly complex.

On the contrary, when (considering the same set of descriptors) the value of C is small (*i.e.* only a small number of samples are badly represented by all the descriptors), the database can be considered as less complex. When, in addition, the value of I almost equals the value of U (*i.e.* all the descriptors describe correctly almost all the samples), the complexity of the database may be considered as low.

To conclude this discussion, we can note that all the measures introduced in Sect. 2 may also be applied to several structural descriptors. Indeed, we have seen that what we only need in our protocol is a confusion matrix including distances or dissimilarity measures between noisy symbol descriptors and models. In this case, we can easily extend this framework to structural methods based on graph representation and graph similarity measures which quantify the effort needed to match one graph with a another [28–30].

5 Conclusion

In this paper, we introduced an experimental protocol and measures for characterizing the performance of descriptors in the context of symbol description. The measures we introduced are of two types. While the first type of measures is devoted to the descriptive power of each descriptor taken separately in terms of uniqueness, distinctiveness or robustness towards noise, the second type of measures aims at evaluating the complementarity of a set of descriptors. Concerning the first type of measures, we first recalled the definitions of confusion matrices, recognition rate, precision, recall and Cumulative Match Characteristics (CMC) curves. Although some of these measures are already known by many researchers in our community, our originality is that we linked them to the notions of distinctiveness and uniqueness. Second, we introduce two measures that are new in the field of document analysis. These two measures are respectively the tolerance intervals, characterizing the robustness towards noise, and a qualitative measure characterizing the symmetries in the confusions. Concerning the measures of the second type, we introduce original measures to characterize upstream the complementarity between multiple descriptors. These measures may assist the researchers when selecting the descriptors to be combined, instead of selecting them by trial and error downstream.

We analysed experimentally a didactic case study (considering the widely-known descriptors ART and SC), to illustrate the effectiveness of the measures we defined. Even if the main objective of this experimental part is didactic and not directly to draw conclusions about the performance characteristics of ART and SC, it highlights the relevance of combining SC and ART for describing symbols.

It has to be noted that the complementarity measures may be additionally used for characterizing the complexity of a given database: the basic idea behind this is that, when a well-chosen set of mutually complementary descriptors gives poor results on a given database, we can consider that this database is highly complex. As a conclusion, our measures may therefore be helpful for various purposes concerning performance evaluation, in the field of document description and analysis. We are currently working on an ambitious performance evaluation campaign relying on our protocol and measures, dedicated to symbol description by shape descriptors.

References

1. Terrades, O.R., Tabbone, S., Valveny, E.: A review of shape descriptors for document analysis. In: Proceedings of the International Conference on Document Analysis and Recognition—ICDAR'07, pp. 227–231 (2007)
2. Phillips, I., Chhabra, A.: Empirical performance evaluation of graphics recognition systems. IEEE Trans. PAMI **21**(9), 849–870 (1999)
3. Chhabra, A., Phillips, I.: The second international graphics recognition contest—raster to vector conversion: A report. In: Tombre, K., Chhabra, A.K. (eds.) Graphics recognition: Algorithms and Systems. LNCS, vol. 1389, pp. 390–410. Springer (1998)
4. Chhabra, A., Phillips, I.: Performance evaluation of line drawing recognition systems. In: Proceedings of 15th. International Conference on Pattern Recognition, vol. 4, pp. 864–869. Barcelona, Spain (2000)
5. Wenyin, L., Zhai, J., Dori, D.: Extended summary of the arc segmentation contest. In: Blostein, D., Kwon, Y.B. (eds.) Graphics Recognition: Algorithms and Applications. LNCS, vol. 2390, pp. 343–349. Springer (2002)
6. Valveny, E., Dosch, P.: Symbol recognition contest: a synthesis. In: Lladós, J., Kwon, Y.B. (eds.) Graphics Recognition Recent Advances and Perspectives. LNCS, vol. 3088, pp. 368–385, Springer (2004)
7. Dosch, P., Valveny, E.: Report on the second symbol recognition contest. In: Liu, W., Lladós, J. (eds.) Graphics Recognition. Ten Years Review and Future Perspectives. LNCS, vol. 3926, pp. 381–397. Springer (2006)
8. Trier, O.D., Jain, A.K., Taxt, T.: Feature extraction methods for character recognition—a survey. Pattern Recognit. **29**(4), 41–662 (1996)
9. da Fontoura Costa, L., Cesar, R.M. Jr.: Shape Analysis and Classification: Theory and Practice. pp. 685 CRC Press, Boca Raton (2001)
10. Zhang, D., Lu, G.: Review of shape representation and description techniques. Pattern Recognit. **37**, 1–19 (2004)
11. Tumer, K., Ghosh, J.: Analysis of decision boundaries in linearly combined neural classifiers. Pattern Recognit. **29**(2), 314–348 (1996)
12. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: Proceedings of the Thirteenth International Conference on Machine Learning, pp. 148–156 (1996)
13. Skurichina, M., Duin, R.P.W.: Bagging, boosting and the random subspace method for linear classifiers. Int. J. Pattern Anal. Appl. **5**(2), 121–135 (2002)
14. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. Int. J. Data. Min. Knowl. Discov. **2**(2), 1–43 (1998)
15. Kittler, J.: A framework for classifier fusion: is it still needed? In: Proceedings of the Joint IAPR International Workshops on Advances in Pattern Recognition, pp. 45–56. Springer-Verlag (2000)

16. Ramos, O., Valveny, E., Tabbone, S.: Optimal classifiers fusion in a non-Bayesian probabilistic framework. IEEE Tran. PAMI **31**(9), 1630–1644 (2009)
17. Terrades, O.R., Valveny, E., Tabbone, S.: On the combination of ridgelets descriptors for symbol recognition. In: Liu, W., Lladós, J., Ogier, J.-M. (eds.) Graphics Recognition. Recent Advances and New Opportunities. LNCS, vol. 5046, pp. 40–50. Springer (2008)
18. Valveny, E., Dosch, P., Winstanley, A., Zhou, Y., Yang, S., Yan, L., Wenyin, L., Elliman, D., Delalandre, M., Trupin, E., Adam, S., Ogier, J.M.: A general framework for the evaluation of symbol recognition methods. Int. J. Doc. Anal. Recognit. **9**(1), 59–74 (2007)
19. Delalandre, M., Pridmore, T., Valveny, E., Locteau, H., Trupin, E.: Building synthetic graphical documents for performance evaluation. In: Liu, W., Lladós, J., Ogier, J.-M. (eds.) Graphics Recognition. Recent Advances and New Opportunities. LNCS vol. 5046, pp. 288–298. Springer (2008)
20. Valveny, E., Tabbone, S., Terrades, O.R., Philippot, E.: Performance characterization of shape descriptors for symbol representation. In: Liu, W., Lladós, J., Ogier, J.-M. (eds.) Graphics Recognition. Recent Advances and New Opportunities. LNCS vol. 5046, pp. 278–287. Springer (2008)
21. Jouili, S., Tabbone, S.: Evaluation of graph matching measures for documents retrieval. In: Eighth IAPR International Workshop on Graphics Recognition (GREC 09), La Rochelle (2009)
22. Kim, W.Y., Kim, Y.S.: A new region-based shape descriptor. ISO/IEC MPEG99/M5472 Maui, Hawaii (1999)
23. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE Trans. PAMI **24**(4), 509–522 (2002)
24. Mori, G., Belongie, S., Malik, J.: Efficient shape matching using shape contexts. IEEE Trans. PAMI **27**(11), 1832–1837 (2005)
25. Visani, M., Garcia, C., Laurent, C.: Comparing robustness of two-dimensional PCA and eigenfaces for face recognition. In: Proceedings of the International Conference on Image Analysis and Recognition (ICIAR 04) Springer LNCS 3212, 2:717–724. Porto, Portugal (2004)
26. Doddington, G., Liggett, W., Martin, A., Przybocki, M., Reynolds, D.: Sheep, Goats, Lambs and Wolves: a statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. In: International Conference on Spoken Language Processing (ICSLP), Sydney, USA (1998)
27. Kanungo, T., Haralick, R.M., Phillips, I.: Nonlinear global and local document degradation models. Int. J. Imaging Syst. Technol. **5**, 220–230 (1994)
28. Jouili, S., Tabbone, S.: Graph matching using node signatures. In: Proceedings of the 7th workshop on graph-based representations in pattern recognition—GbRPR 2009, pp. 154–163. Venice, Italy May (2009)
29. Robles-Kelly, A., Hancock, E.R.: Graph edit distance from spectral seriation. IEEE Trans. PAMI **27**(3), 365–378 (2005)
30. Papadopoulos, A.N., Manolopoulos, Y.: Structure-based similarity search with graph histograms. In: Proceedings of International Workshop on Similarity Search (DEXA IWOSS 99), pp. 174–178 Sep (1999)