Information Fusion 103 (2024) 102101

Contents lists available at ScienceDirect

Information Fusion

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

Full length article

Ultrametrics for context-aware comparison of binary images

C. Lopez-Molina [a,b,*], S. Iglesias-Rey [a,b], B. De Baets [c]

[a] Dpto. Estadistica, Informatica y Matematicas, Universidad Publica de Navarra, 31006 Pamplona, Spain
[b] Idisna, NavarraBiomed, Hospital Universitario de Navarra, 31008 Pamplona, Spain
[c] KERMIT, Department of Data Analysis and Mathematical Modelling, Ghent University, Belgium

ARTICLE INFO

Keywords:
Image comparison
Binary image
Context awareness
Ultrametric

ABSTRACT

Quantitative image comparison has been a key topic in the image processing literature for the past 30 years. The reasons for it are diverse, and so is the range of applications in which measures of comparison are needed. Examples of image processing tasks requiring such measures are the evaluation of algorithmic results (through the comparison of computer-generated results to given ground truth) or the selection of loss/goal functions in a machine learning context. Measures of comparison in literature take different inspirations, and are often tailored to specific needs. Nevertheless, even if some measures of comparison intend to replicate how humans evaluate the similarity of two images, they normally overlook a fundamental characteristic of the way humans perform such evaluation: the context of comparison. In this paper, we present a measure of comparison for binary images that incorporates a sense of context. More specifically, we present a Methodology for the generation of ultrametrics for context-aware comparison of binary images. We test our proposal in the context of boundary image comparison on the BSDS500 benchmark.

1. Introduction

Binary images are common representations of feature information in image processing, often embodying intermediate or final results in image segmentation, or object/change detection [1]. The literature holds a large number of measures for the comparison of binary images, including (distance) metrics and various other types of functions. This rich variety is due to the many possible applications for such measures. For example, comparison measures are often used to compare computer-generated solutions with ground truth for quality evaluation [2,3], since many applications (e.g. segmentation) represent object/region detection results as one or more binary images [4–6]. Dedicated reviews can be found for generic contexts [7], as well as for specific ones [8,9]. Also, comparison measures play a role in shape comparison and matching, which is relevant for object recognition and tracking [10,11]. While those applications were long known in literature, the necessity for reliable comparison measures has been boosted by the development of CNN-based image processing systems [12,13]. The training phase of such systems heavily relies on loss functions, which are often based on comparisons between the predicted images and their labelled (ground-truth) solutions [14]. Despite some proposals to use non-binary ground truth [15], binary image comparison is still the standard in literature for goal functions in such detectors. By using context-aware comparison measures, which are more akin to human understanding than classical pixelwise comparison measures, all such applications shall produce more explainable and human-acceptable results.

Comparison measures take inspiration from a wide variety of mathematical tools, ranging from mathematical functions to information-theoretic measures. Some prominent examples are based on metrics, such as the well-known Hausdorff metric [16]. Other comparison measures are based on trigonometric functions and a vectorial interpretation of images [17]. Information-theoretic measures attempt to model the coinciding and diverging information in two binary images [18–20]. Most of these measures come in a parametric form, giving rise to a comprehensive range of instantiations, often exhibiting significantly different behaviour. Yet other approaches are based on the expectation–maximization algorithm in order to segment an image [18]. This idea has been further developed, introducing the concepts of *confusion* and *consensus* areas to the field of image segmentation [19]. In their comparison of edge detection algorithms, Fram and Deutsch [20] use information-theoretic measures to evaluate the capability of replicating human behaviour in the detection of edges in the presence of noise.

One of the few aspects generally left unexplored in the development of comparison measures is the incorporation of the context. The majority of comparison measures only consider the images to be compared in order to produce a quantitative evaluation of the (dis)similarity between them. This is usually the case with metrics,

* Corresponding author at: Dpto. Estadistica, Informatica y Matematicas, Universidad Publica de Navarra, 31006 Pamplona, Spain.
E-mail addresses: carlos.lopez@unavarra.es (C. Lopez-Molina), sara.iglesias@unavarra.es (S. Iglesias-Rey), bernard.debaets@ugent.be (B. De Baets).

https://doi.org/10.1016/j.inffus.2023.102101
Received 30 June 2023; Received in revised form 8 October 2023; Accepted 27 October 2023
Available online 30 October 2023
1566-2535/© 2023 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

which yield the distance between two elements in the generic universe of all binary images, fulfilling the corresponding metric axioms (*identity of indiscernibles*, *symmetry* and *triangle inequality*). Such approach might appear natural, and is definitely mathematically convenient, but is in fact divergent from how humans proceed. Humans implicitly incorporate a sense of context in their comparisons, arguably performing a multidimensional analysis of the characteristics for each object. This incorporation of context may even lead to not fulfilling the metric axioms when performing the comparison. The context can be explicit, but would be kept implicit by humans if they do not provide further information.

An illustrative example of such implicit context analysis performed by humans is the *Jamaica–Cuba–Russia* comparison, as presented by Tversky [21]. The example grounds on the idea that humans normally judge that $d(Cuba, Jamaica) + d(Cuba, Russia) < d(Jamaica, Russia)$, with d some dissimilarity measure between countries. Otherwise said, that the sum of the pairwise dissimilarity between *Jamaica* and *Cuba*, and that between *Cuba* and *Russia*, is in fact lower than the perceived dissimilarity between *Jamaica* and *Russia*, hence breaking the triangle inequality. The reason lies in the fact that each country is perceived as a multidimensional information object and, for the comparison of each pair of countries, humans implicitly determine the dimension of comparison, *i.e.*, the context of comparison. Even when the context is not explicit, it is implicitly determined by the human evaluator on the basis of prior experience. For example, the *Jamaica–Cuba* comparison is normally set in terms of geography, whereas the *Cuba–Russia* comparison is made on the basis of politics; humans selectively alter the role and configuration of the context in the comparisons. This example is used to discredit the necessity of imposing the triangle inequality in the modelling of human behaviour, as well as to shed light on the multidimensional nature of human interpretation. But, very interestingly, it also serves as an illustration of the relevant role of context in human comparisons.

In general, comparison measures in the literature make no use of the notion of context. Typically, they only consider the pair of images to be compared, assuming a global context and hence providing an absolute quantification of (dis)similarity. While this might be advantageous for some tasks, the process draws away from how humans perform comparisons. The aim of this work is to present a comparison measure for binary images in which the context plays a prominent role. Specifically, we intend to create an ultrametric that quantifies the dissimilarity between any two images within the context of comparison. The quantified distance between two images is obtained not only in terms of their coincidences and divergences, but also on basis of the characteristics of the other images within the context. Through the inclusion of a context, our proposed measure brings the process closer to how humans naturally perform comparisons.

In this work we present a comparison measure for binary images which attempts to produce human-coherent results by mimicking human understanding of context. This is done by relying on comparison models stem from psychological studies [22], which are able to accommodate and model the context of comparison. Also, our proposal is quantitatively tested in terms of comparison-based intra- and inter-class discrimination [23], a task for which humans are significantly more proficient than existing comparison measures in literature.

An early development of the ideas in this manuscript was presented in [24]. While [24] already depicted the idea of using a tree-based ultrametric for the comparison of binary images, the present manuscript features a number of enhancements. These enhancements involve, but are not restricted to, a better fit in literature, a cost-function-based algorithm for tree construction, the incorporation of N-to-M comparison measures in the tree construction and a qualitatively expanded experimental setup.

The remainder of this paper is organized as follows. Section 2 introduces the mathematical preliminaries that will be employed in this paper, while Section 3 presents the importance of context-aware comparison measures. The methodology for the construction of ultrametrics is presented in Section 4, which is put to the test in Section 5. Finally, Section 6 presents the conclusions of the paper.

2. Preliminaries

This section collects the mathematical definitions applied in upcoming sections.

Definition 2.1. A function $d : U \times U \to \mathbb{R}^+$ is called a metric on U if it satisfies the following properties:

1. *Identity of indiscernibles*: $d(x, y) = 0$ if and only if $x = y$.
2. *Symmetry*: $d(x, y) = d(y, x)$ for any $x, y \in U$.
3. *Triangle inequality*: $d(x, z) \leq d(x, y) + d(y, z)$ for any $x, y, z \in U$.

The function d is called an ultrametric if instead of the triangle inequality, it satisfies:

3'. *Ultrametric inequality* or *strong triangle inequality*: $d(x, y) \leq \max(d(x, z), d(y, z))$ for any $x, y, z \in U$.

Obviously, any ultrametric is a metric. Note that an ultrametric can be characterized [25] as a metric for which any three points can be relabelled as x, y, z such that

$$d(x, y) \leq d(x, z) = d(y, z).$$

Recall that a binary fuzzy relation $E : U^2 \to [0, 1]$ is called min-transitive if

$$\min(E(x, y), E(y, z)) \leq E(x, z)$$

for any $x, y, z \in U$. With a given bounded ultrametric d taking values in the unit interval $[0, 1]$, we can associate a binary fuzzy relation E defined by $E(x, y) = 1 - d(x, y)$. It is straightforwardly verified that the ultrametric inequality of d is equivalent to the min-transitivity of E [26,27].

In this paper, we consider images to have some fixed dimensions $\mathcal{M} \times \mathcal{N}$, so that $\Omega = \{1, \ldots, \mathcal{M}\} \times \{1, \ldots, \mathcal{N}\}$ represents the set of positions in an image. The set of all binary images, *i.e.*, the set of mappings $\Omega \mapsto \{0, 1\}$, is denoted as $\mathbb{B}$. We will refer to individual binary images with uppercase characters, like I or A, and to sets of images with bold-faced uppercase characters, like $\mathbf{I} = \{I_1, \ldots, I_k\}$.

The positive (resp. negative) information in binary images will be represented by 1s (resp. 0s). The classical set-theoretic operations on binary images will be used: intersection ($\wedge$), union ($\vee$), and inclusion ($\subseteq$, $\subset$). We reserve the symbols $\cap$ and $\cup$ for the intersection and union of sets of images, respectively. The dilation of a binary image A by some structuring element K is given by $D_K(A) = \{c \in \Omega \mid c = a + b \text{ for some } a \in A \text{ and } b \in K\}$.

3. Ultrametrics for object comparison

First, we present an initial classification of comparison measures and a discussion of their main differences in Section 3.1. Then, we focus on a network-based approach for the comparison of objects (Section 3.2) and present a methodology for the construction of ultrametrics derived from such approach (Section 3.3).

3.1. Classes of comparison measures

Comparison measures, such as metrics, ultrametrics or any other type of function, are crucial in most scientific fields. These measures are instrumental in tasks such as quality evaluation, optimization, and clustering. According to Tversky [22], comparison measures can be divided in two main classes: those based on geometrical interpretations, named *spatial models*, and those based on graph theory, named *network models*. Spatial models represent each object as a point in some coordinate space, so that the distance between points represents the proximity between objects. Most of the metrics and dissimilarity measures adhere to this strategy, including for example the different extensions of the Hausdorff metric [28,29]. Alternative to spatial models, network models generate a graph-like representation of the relationships between

the objects to be compared. Each object is defined as a node in a connected graph, usually a tree, while edges (and their weights) are used to represent the dissimilarity (or distance) between any two objects.

In Tversky's twofold taxonomy, network models are better suited to embody the notion of context than spatial models. Spatial models may potentially establish a coordinate space depending upon some definition of the context (nevertheless, this is absent from the image processing literature, except for image retrieval tasks [30] involving the comparison of distributions). In a sense, context is usually assumed to be large and unspecific enough, containing *all possible elements*. Network models allow for an easy adaptation to the context through the generation of a graph. When establishing the graph topology, all possible interrelations between elements are considered. Thus the set of objects to be compared affects the topology of the network, and hence the dissimilarity between any two nodes.

In this work, we present an ultrametric based on a network model, allowing for a simple, yet meaningful, definition of the *context of comparison*. The generation of a graph, upon which an ultrametric is built, allows to model relationships between images in the set. This alternative enables, in our opinion, a better modelling of the human behaviour in comparing and evaluating dissimilarities, since an implicit sense of context is included. Also, by satisfying the ultrametric properties, we guarantee that our comparison measure is also a metric, and can hence be applied in many different tasks for which metrics are required. The main drawback of this alternative is that the network topology cannot yield distances from or to elements that were not present at the time of defining the topology. Therefore, an important limitation is that all elements to be compared need to be known in advance.

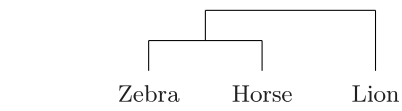
We recall that some comparison measures based on spatial models (partially) exhibit context-awareness. For example, we can consider that learned metrics [31,32] are in fact bounded to the context (training data) from which they were created. At the same time, since the learned object is a metric, it can be used to compare elements that were out of the original training data. Problems might arise, nevertheless, depending on how representative the training data is, potentially affecting the reliability of comparison. Also, the need for data labelling prior to the learning process makes learned metrics ineligible for some practical applications.

3.2. Network-based approach to comparison

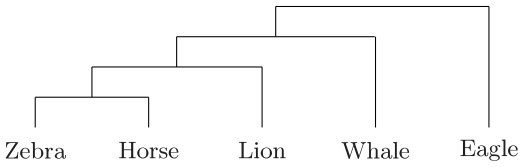
Network-based comparison measures are computed on the basis of a network topology on a given set of elements. Although this network topology can be any type of graph (the distances on often-cyclic road maps being an example of it), usually trees are the preferred choice, since they enforce a convenient hierarchical structure on the data. This choice is not only relevant in the current context, but also central to many other algorithms, notably hierarchical clustering algorithms. This work focuses on the use of trees as network topologies.

A graph is formed by *nodes*, connected through *edges* or *branches*. In the case of trees, we ensure the graph to be connected and acyclic. When using trees for the generation of hierarchical structures, lower-level nodes are referred to as *leafs* and the highest-level node is referred to as *root*. While edges in a tree can be directed, we exclusively consider undirected edges.

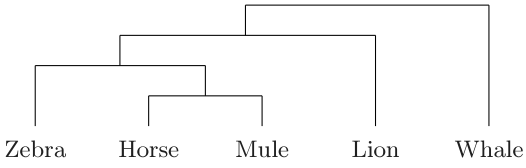
Fig. 1 displays an example of a dissimilarity-based, tree-like organization of linguistic terms. This graph contains, as leafs, the terms Zebra, Horse and Lion, and the resulting clusters of elements. In Fig. 1(a), {Zebra, Horse} are first grouped together, since they are the closest possible terms in the pool. Then, {Zebra, Horse} are grouped with {Lion}, forming the root of the tree. For now, we shall avoid numerical details, which will be provided in upcoming sections, so we assume that the dissimilarity between the elements is proportional to the height of the node in which they are first grouped together.



(a) Tree model of {Zebra, Horse, Lion}.



(b) Tree model of $\cup$ {Whale, Eagle}.



(c) Tree model of {Zebra, Horse, Lion} $\cup$ {Whale, Mule}.

Fig. 1. Tree model for the similarity representation of linguistic terms. Three different variants of the set of terms are presented.

Figs. 1(b) and 1(c) display the resulting trees when adding two different sets of terms to the candidate set in Fig. 1(a): Whale and Eagle, in the former, and Whale and Mule, in the latter . In the first case, by including terms perceptually distant from all elements in the original set, the tree topology does not change significantly. The new elements are grouped together with the previously-existing terms at the top level. In Fig. 1(c), we find a different situation. By including the terms Whale and Mule, the latter perceived as close to Horse, the relationships between the original terms is altered; again, not elaborating on numerical details, the dissimilarity between Zebra and Horse would be 2 instead of 1 (they merge at level 2, instead of at level 1). This represents the fact that humans understand Zebra and Horse as variably similar, depending on the context of comparison. Otherwise said, Zebra and Horse are perceived as not-so-close, once Mule is part of the candidates. A similar situation is found for the terms Horse and Lion, which are seen as variably similar depending on the appearance of perceptually close terms (as Mule).

3.3. Tree-based generation of comparison measures

The examples in Fig. 1 provide an idea of how to organize terms according to their perceived dissimilarity, but avoid details on how to generate an ultrametric from it, or any other comparison measure. This process requires certain mathematical tools.

The process of building the tree starts by considering all objects in the set to be compared, with the first merged node corresponding to the *closest pair* of terms among all candidate pairs. The choice of the closest pair is in fact not straightforward. Humans are relatively good at grouping, while, from a computational point of view, many different paradigms may be involved: minimum entropy of the resulting set, lowest dissimilarity using a similarity measure or metric, etc. In the process of further constructing the tree, we need to iteratively decide which is the next chosen pair, discarding it in future iterations. At each next step we will take into account only the remaining nodes in the set (including the new merged node) until the root is reached. It is relevant that, apart from the leafs, all remaining nodes will represent sets of elements, so that choices on node grouping at each stage need to be based on set (*i.e.*, M-to-N) comparisons.

Once the structure of the tree is set, it is important to establish weights for the edges or branches. Depending on the nature of the strategy used for node merging, we can distinguish two different situations. If the approach is qualitative, *i.e.*, it determines which is the next node to be generated but does not yield a numerical evaluation of its cost, each new node is taken at an increasing height in the tree. The first node created combining the items from the original set has height 1, the second has height 2, and so on. If the strategy used for node grouping is numerical in nature, the height of each node is unknown in advance. With the term *cost function*, we refer to the function yielding the numerical evaluation of the merging of two candidate nodes in the generation of the tree.

We consider the following definition of a cost function. The original set of nodes to be compared is denoted by A, with $\wp(A)$ the powerset of A. Given two nodes a_i, a_j, each representing a non-emtpy set of items, we write $a_j \subseteq a_i$ if all elements of a_j are contained in a_i.

Definition 3.1. A cost function on $\wp(A)$ is a function

$$\Gamma : \wp(A) \times \wp(A) \to \mathbb{R}^+$$

satisfying:

1. $\Gamma(a_i, a_j) = \Gamma(a_j, a_i)$ for any a_i, a_j;
2. $\Gamma(a_i, a_j) \leq \Gamma(a_i, a_k)$ for all $a_j \subseteq a_k$.

Property 1 expresses the symmetry of the cost function. Property 2 ensures that any grouping of nodes in time will have an increasing cost, *i.e.*, that it is impossible to merge two nodes in the tree at a height lower than that of previous nodes.

A cost function (in the sense of Definition 3.1) can be created from any 1-to-1 comparison measure c, as long as it is positive and symmetric, by computing the maximum over all possible pairs of elements as follows

$$\Gamma_c(a_i, a_j) = \max_{x \in a_i, y \in a_j} c(x, y) \, . \tag{1}$$

We use the symbol C to refer to the set of candidate nodes in each iteration of the tree construction, *i.e.*, a subset of $\wp(A)$. Given a cost function Γ, a tree is constructed by selecting, iteratively, the two non-equal nodes that yield the lowest cost when merged forming a prospective node:

$$\begin{aligned} &\arg\min_{a_i, a_j \in C} \Gamma(a_i, a_j) \\ &\quad = \{(a_i, a_j) \in C^2 \mid i \neq j, \forall u, v \in C : \Gamma(a_i, a_j) \leq \Gamma(u, v)\}. \end{aligned}$$

Each new node is considered to have a height equivalent to the cost of its generation, *i.e.*, the cost of merging the two original nodes. Note that different pairs of candidate nodes might have the same cost during the merging process; there are many different strategies to overcome this issue, e.g., creating several nodes with the same cost simultaneously or selecting one at random.

Whichever strategy is employed to design the tree, the graph can be used to create an ultrametric. The distance between any two items in the original set, in terms of the ultrametric, is the height of the lowest node including both of them. This concurs with the *strong inequality* axiom in Definition 2.1, since the distance between these two elements fulfils $d(x, y) \leq \max(d(x, z), d(y, z))$ for any $x, y, z \in U$. This measure is context-dependent, in the sense that different pools of elements will lead to distinct tree topologies, and hence to the ultrametric yielding different values.

In the case of Fig. 1, and assuming that the cost of each node is equal to its height, the ultrametric is explicitly defined as in Table 1.

Table 1
Ultrametric resulting from the tree in Fig. 1(a).

	Zebra	Horse	Lion
Zebra	0	1	2
Horse	1	0	2
Lion	2	2	0

4. Ultrametrics for binary image comparison

In this section, we tackle the construction of an ultrametric for binary image comparison. First, we review the existing literature (Section 4.1) and introduce ultrametric trees (Section 4.2), whose construction is supported by the Twofold Consensus Ground Truth (TCGT) (Section 4.3). Then, we present our proposal for the construction of a tree-based ultrametric for binary images (Section 4.4).

4.1. Binary image comparison

Binary images, while having very little use for image acquisition, are recurrent in literature for the representation of intermediate/final result. For example, it is typical in medical imaging to use binary images for the representation of segmented/detected areas [13], despite recent proposals to use richer representations to account for inter-expert variability, uncertainty and confidence [15]. Binary images are also widely used in photogrammetry, both representing intermediate and final results [1]. In fact, binary images have been the standard representation for low-level features, including boundaries and ridges, on top of which complex tasks are built. Examples of such tasks range from boundary-aware inpainting [33], 3D building reconstruction using pre-segmented areas [34] and ridge-based fungal growth modelling and control [35].

The extensive use of binary images led to the proliferation of a broad variety of measures for binary image comparison. A large part of the literature is dedicated to confusion-matrix-based measures [18,19,36], but there is also a number of measures adopting a geometrical interpretation of the space of binary images [17]. Popular options in this regard are the Hausdorff metric [16], which has led to different task-specific generalizations [28,29], or the Symmetric Difference [7]. Other measures, although not satisfying the metric axioms, are to some extent based on metrics. Examples are Pratt's FoM (PFoM) [37] and Haralick's Measure [38]. Yet other measures are based on ground truth composition and segmentation comparison [39].

All of the mentioned measures fulfil certain properties, metric axioms or other ones, but do not offer an easy and intuitive way to incorporate the context of comparison. As far as we know, there are no references in the literature introducing context-aware comparison measures, meaning that each comparison is performed while taking into account all the binary images within the set. In fact, when used, the term *context* is related to the idea that different regions of an image present local and global discriminative surroundings. Hence, *context* refers to a collection of points in the neighbourhood of the object of interest; examples can be found in a large variety of applications like shape matching [40] and edge [41] or saliency detection [42]. Combining object saliency and segmentation, a notion of context may also be incorporated through the analysis of multi-scale superpixels [43].

The only measures that consider the notion of context-awareness, *i.e.*, taking into account all the binary images within the set as presented in this work, are related to concept comparison [44,45] and to context-dependent models in choice theory [46], incorporating different weights to simulate the effect of background and local context. The majority of these measures arise as a fundamental notion from theories of cognition in psychology [47]. Although considering the notion of context-awareness as advocated in this paper, these measures are not suitable for image comparison. In fact, there are no examples in the image processing literature applying this notion of context to binary image comparison.

We propose a context-aware comparison measure for binary images, using a tree-based representation to obtain dissimilarity quantifications by means of a derived ultrametric.

4.2. Tree-based analysis of sets of binary images

The tree-based construction of an ultrametric presented in Section 3 can be extended to the case of binary images, given a suitable cost

function Γ. A cost function shall represent the effort needed to group together two sets of images in the process of creating a new node. Hence, the construction of Γ is intuitively related to N-to-M comparisons of binary images.

Despite the large variety of 1-to-1 comparison measures for binary images, very few of them allow for comparisons between sets of images. Some strategies have been proposed for 1-to-N comparisons for specific tasks, e.g., for the quantification of the dissimilarity between a computer-generated solution and multiple ground truth images. While dedicated strategies have been designed (see [48], only applicable to 1-to-N comparisons), most authors in literature simply rely on 1-to-1 comparisons and compute the lowest dissimilarity to the N candidates. This solution is numerically convenient, but fails at capturing the spirit of set comparison, since usually only two images are actively used in the quantification.

In our proposal, N-to-M comparisons will be performed on the basis of the Twofold Consensus Ground Truth (TCGT) [23], a tool for binary image representation and fusion that can be further evolved to measure the heterogeneity of sets of images. This method provides a quantification of the distance between images in the set, and stands as a valid strategy to design a tree-based dissimilarity measure. It also allows for the creation of an ultrametric, since the distance between two images in the original set is the height of the lowest node including both of them, as explained in Section 3.3. Addition of new binary images to the original set will lead to different quantifications of the dissimilarity, and consequently, to different configurations of the ultrametric tree, due to its context-aware nature.

4.3. Twofold Consensus Ground Truth

The Twofold Consensus Ground Truth (TCGT) is a fusion method for binary images [23]. Originally, it was designed to fuse ground truth boundary images in datasets in which each image comes with diverse solutions (see, e.g., [49,50]). This is a very common case, due to the existing divergence among experts when labelling the ground truth, and also among results obtained by the same human at different times. Note that the TCGT can be used for purposes other than the original one; in fact, it can be seen as a general method to fuse binary images.

In the literature there are many different techniques to fuse images, *i.e.*, to produce a single binary image (consensus) from diverse binary images (candidates). Examples can be found using both geometrical [51,52] and statistical [18] approaches to the problem. A critical difference between the TCGT and those techniques is that the TCGT is robust to spatial imprecisions in the images. Another one is that the TCGT represents the result of the fusion process as a set. This set is a compact representation of the information contained in the input image set, based on the spatial interpretation of coincidences and divergences of the images to fuse, and also of some of the characteristics and properties of the set.

The TCGT is supported by two different consensus operators, namely the *strong* and the *weak consensus*.

Definition 4.1. The strong consensus image of a set of binary images $\mathbf{I} = \{I_1, \dots, I_k\}$ is the binary image $s_T(\mathbf{I})$ defined as

$$s_T(\mathbf{I}) = \mathcal{D}_T(I_1) \wedge \mathcal{D}_T(I_2) \wedge \cdots \wedge \mathcal{D}_T(I_k), \tag{2}$$

where $\mathcal{D}_T(I_i)$ denotes the dilation of image I_i by the structuring element T.

Definition 4.2. The weak consensus image of a set of binary images $\mathbf{I} = \{I_1, \dots, I_k\}$ is the binary image $w_T(\mathbf{I})$ defined as

$$w_T(\mathbf{I}) = \mathcal{D}_T(I_1) \vee \mathcal{D}_T(I_2) \vee \dots \vee \mathcal{D}_T(I_k), \tag{3}$$

where $\mathcal{D}_T(I_i)$ denotes the dilation of image I_i by the structuring element T.

Definition 4.3. The consensus of a set of binary images $\mathbf{I}$ is the set of images $c_T(\mathbf{I})$ defined as

$$c_T(\mathbf{I}) = \{B \in \mathbb{B} \mid B \subseteq w_T(\mathbf{I}) \text{ and } s_T(\mathbf{I}) \subseteq D_T(B)\}. \tag{4}$$

The consensus, which is a set itself, satisfies some practical properties as presented in [23]. One of these properties is that of *information combination*. This property refers to the ability to combine information from different images, meaning that the resulting set selectively fuses information from each image. An example can be found in Fig. 2, which depicts a scenario of ground truth gathering from human labellers. From the original image in Fig. 2(a), two labellers have created the ground truth images S_1 and S_2 (Figs. 2(b)–(c)). In this situation, the solution D (Fig. 2(f)) is not exactly like S_1 and S_2 despite its resemblance. However, $D \in c_T(\{S_1, S_2\})$, as we can infer from observing the strong and weak consensus of $\{S_1, S_2\}$ (Figs. 2(d)–(e)). This example shows that images similar to or composed of parts from the original set are actually gathered in the consensus set of the TCGT, *i.e.*, we are indeed obtaining combined information from the different solutions. Observing Fig. 2, we can derive that, even with simplistic examples, thanks to the TCGT, the consensus set provides much more information than just a list of images. It builds up knowledge from the existing morphological relations between images.

The TCGT is useful not only to generate consensus sets, but also enables the analysis of the original set of images. That is, to produce metadata about the images involved in the process. An example is the *scaled heterogeneity* of a set of images.

Definition 4.4. Let $\mathbf{I} = \{I_i, \dots, I_n\}$ be a set of binary images. The *scaled heterogeneity* of $\mathbf{I}$ is given by

$$H_T^*(\mathbf{I}) = \frac{|w_T(\mathbf{I}) \setminus s_T(\mathbf{I})|}{|\Omega|},$$

where w_T and s_T are the weak and strong consensus.

We intend to use the TCGT as a basis for the quantitative N-to-M comparison of binary images, which leads to the construction of a hierarchical tree of binary images. In order to achieve it, we propose to use a cost function based on H_T^*. Specifically, we consider

$$\Psi_T(\mathbf{I}, \mathbf{J}) = \mathrm{H}_T^*(\mathbf{I} \cup \mathbf{J}). \tag{5}$$

The cost function Ψ_T fulfils the properties of Definition 3.1. The proof of Property 1 (symmetry) is straightforward. With regard to Property 2, adding more images to one of the sets of binary images will only extend w_T, while reducing s_T; therefore, Ψ_T will increase in each iteration, fulfilling Property 2.

By means of Definition 3.1 and the scaled heterogeneity, we are able to perform N-to-M comparisons, avoiding the usual reuse of 1-to-1 comparisons. Applying this strategy, we will construct an ultrametric tree, calculating the scaled heterogeneities between images within a set in a context-aware manner. Since there are not many N-to-M comparison measures in the literature as stated in Section 4.2, a performance comparison will be presented in Section 5, applying the cost function presented in Eq. (1) using different conventional 1-to-1 binary image comparison measures.

4.4. Ultrametric tree construction

We present an ultrametric tree construction procedure for a set of binary images based on N-to-M comparisons using Ψ_T. First, one set is created for each binary image, so as to produce as many nodes (leaves) as images. The node grouping costs between all such nodes are calculated. The closest pair, *i.e.*, the two sets having the lowest cost, are merged in a new node. Then, the distances between the remaining nodes, including the new one, are recalculated. Once again, the two nodes having the lowest scaled heterogeneity are merged together and assembled in a new node. This process is repeated until a single node,

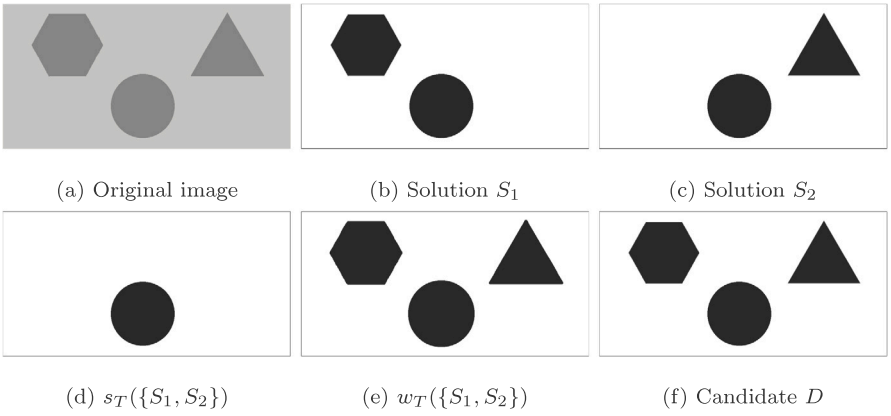


Fig. 2. Example of information fusion using the TCGT. We have (a) an image, (b,c) two hand-made segmentations from it, (d,e) the strong and weak consensus images and (f) a candidate image. The candidate image belongs to $c_T(\{S_1, S_2\})$, although it does not match any of the original images. The structuring element T used for the dilation is a disk of radius 5. 1-valued (resp. 0-valued) pixels are represented in black (resp. white).

containing all binary images from the original set, is obtained; that is, until the root of an ultrametric tree is attained. The designated distance between any two of the original binary images is the height of the lowest node including both of them, thus yielding an ultrametric. In the unlikely case of two or more pairs of nodes having the same grouping cost during the ultrametric tree construction, either choice would be acceptable.

In Algorithm 1 we present the computational procedure of our proposal based on a cost function Γ. The input is a set of binary images $\mathbf{I} = \{I_1, \ldots, I_N\}$, with N the number of images in the set. The output is composed of three different elements. First, a dissimilarity array D including all the cost values between every two nodes; in our case, for a branching factor of 2 (two children merging at each level of the ultrametric tree), the total number of elements is $2N - 1$, so the dimensions of D are $(2N - 1)\times(2N - 1)$. Second, a $(2N - 1)\times 2$ array called *nodeArray* is created, including the height of each node in the ultrametric tree. Third, a $(2N - 1)\times 1$ array called *treeArray* is obtained, including, for each node, which binary images are grouped together.

As presented in Section 3, one of the main advantages of using tree-based ultrametrics as comparison measures is the proportionality among the distances between elements and the tree topology. Setting a vertical axis proportional to the cost of the merging, the distance between any two images in the original set is equal to the height of the lowest node containing both images. This approach provides a graphical and intuitive representation of the proximity of the elements to be compared from the original set.

Algorithm 1: Ultrametric tree based on a cost function Γ.

input : Set of binary images **I**
output: Dissimilarity array D, node array *nodeArray* and tree array *treeArray*

$N \leftarrow$ number of images
$D \leftarrow (2N-1)\times(2N-1)$ dissimilarity array of ∞s
$nodeArray \leftarrow (2N-1)\times 2$ array of 0s
$treeArray \leftarrow (2N-1)\times 1$ array of merged nodes
for $n \leftarrow \{1, \ldots, N\}$ **do**
  $nodeArray(n, 1) \leftarrow n$
  $nodeArray(n, 2) \leftarrow 0$
  $treeArray(n) \leftarrow \{I_n\}$
for $i \leftarrow \{1, \ldots, N\}$ **do**
  for $j \leftarrow \{1, \ldots, N\}$ **do**
    $D(i, j) \leftarrow \Gamma(treeArray(i) \cup treeArray(j))$
    $D(j, i) \leftarrow D(i, j)$
for $k \leftarrow \{N+1, \ldots, 2N-1\}$ **do**
  $(x, y) \leftarrow \arg\min D(1 : k, 1 : k)$
  // Updating graph
  $nodeArray(k, 1) \leftarrow k$
  $nodeArray(k, 2) \leftarrow D(x, y)$
  $treeArray(k) \leftarrow treeArray(x) \cup treeArray(y)$
  // Discarding the combined nodes
  // setting their dissimilarity to ∞
  $D(x, :) \leftarrow \infty$
  $D(:, x) \leftarrow \infty$
  $D(y, :) \leftarrow \infty$
  $D(:, y) \leftarrow \infty$
  // Updating the new node
  for $l \leftarrow \{1, \ldots, k\}$ **do**
    $D(k, l) \leftarrow \Gamma(treeArray(k) \cup treeArray(l))$
    $D(l, k) \leftarrow D(k, l)$

5. Experiments

In this section, we conduct extensive experiments aimed at providing insights into two different aspects of our proposal: (a) whether tree-based ultrametrics are solid tools for binary image comparison and (b) whether the choice of the cost function has a severe impact on the performance of an ultrametric. In this regard, we first present the binary image dataset employed for the experiments (Section 5.1). Then, we introduce as alternative binary image cost functions Baddeley's Delta Metric (Δ) and the Symmetric Difference (SD) (Section 5.2). Finally, we present a detailed analysis based on quantitative data (Section 5.3).

5.1. Experimental data

The Berkeley Segmentation Data Set and Benchmark 500 (BSDS500) [49] is a popular dataset for image segmentation and boundary detection tasks [53]. It holds a large set of images, each linked to a collection of hand-labelled ground truth segmentations, which are themselves presented as boundary images. Fig. 3 displays the ground truth images associated with two different original images in the BSDS500 dataset. We refer to the boundary images taken as ground truth for a given original one as a *class*.

Although labellers do incur in a high variability when analysing or segmenting images, humans are generally able to group those images produced after the same original image [48]. Otherwise said, any human would be able to differentiate which boundary images are created from the same original image (intra-class comparisons) and which are not (inter-class comparisons). Any comparison measure should be able

(a) Ground truth associated with image 14092 from the BSDS500 Test set

(b) Ground truth associated with image 14085 from the BSDS500 Test set

Fig. 3. Boundary ground truth images from the BSDS500 Test set. Different human labellers lead to a high variability among images belonging to the same class.

to replicate this human capability, yielding lower values for inter-class images than for intra-class ones. We intend, in fact, to evaluate different comparison measures on the basis of how well they are able to discriminate intra-class from inter-class pairs of boundary images, as is further detailed in Section 5.3.

5.2. Alternative cost functions for the generation of ultrametric trees

Cost functions for the construction of ultrametric trees can be based on the comparison of sets of images, regardless of their cardinality. However, very few measures in literature allow for N-to-M comparison of binary images. Nevertheless, 1-to-1 comparison measures can be used for the generation of N-to-M comparison measures, further used as cost functions. Specifically, in this experiment, we will use two 1-to-1 comparison measures to generate a cost function as in Eq. (1):

(i) Baddeley's Delta Metric (Δ^k) [54,55] is a popular comparison measure derived from the Hausdorff metric [11,16]. Let $I, J \in \mathbb{B}$ be two binary images on Ω and m a metric on Ω. The distance between I and J, as measured by BDM, is then given by

$$\Delta^k(I,J) = \left[\frac{1}{|\Omega|} \sum_{p \in \Omega} \left| w(\mathcal{T}_m[I](p)) - w(\mathcal{T}_m[J](p)) \right|^k \right]^{\frac{1}{k}}, \tag{6}$$

where $w : \mathbb{R}^+ \mapsto \mathbb{R}^+$ is a concave function with $w(x) = 0$ if and only if $x = 0$, $k \in \mathbb{R}^+$ and $\mathcal{T}_m$ is an image distance transformation defined by

$$\mathcal{T}_m[I](p) = \min_{p' \in I} m(p, p'), \tag{7}$$

for all $p \in \Omega$.

In our experiments, we set $w(x) = x$, $k = 2$, and $m = m_t$, so that $m_t(p,p') = \min(t, d_{\text{euc}}(p,p'))$ for any $p, p' \in \Omega$, with d_{euc} the Euclidean metric and $t \in \{2.5, 5, 10\}$. We refer to this comparison measure as Δ_t.

(ii) The Symmetric Difference (SD) [7], a distance-based error measure between binary images. Let $I, J \in \mathbb{B}$ be two binary images on Ω and m a metric on Ω. The distance between I and J, as measured by SD, is then given by

$$\text{SD}^k(I,J) = \frac{\left(\sum_{p \in J} \mathcal{T}_m^k[I](p) + \sum_{p \in I} \mathcal{T}_m^k[J](p)\right)^{1/k}}{|I \cup J|^{1/k}}, \tag{8}$$

where $k \in \mathbb{R}^+$ and $\mathcal{T}_m$ is an image distance transformation defined by Eq. (7).

In our experiments, we set $k = 2$, and $m = m_t$, so that $m_t(p,p') = \min(t, d_{\text{euc}}(p,p'))$ for any $p, p' \in \Omega$, with d_{euc} the Euclidean metric and $t \in \{2.5, 5, 10\}$. We refer to this comparison measure as SD_t.

5.3. Separability analysis

Evaluating the performance of a comparison measure is far from trivial. In general, it is unclear which numerical values a comparison measure shall yield in the comparison of complex objects. In the case of binary image comparison, it is complicated to determine the numerical values comparison measures are meant to produce. The most obvious option would be to ask humans to perform numerical evaluations of similarity or dissimilarity between binary images; this is a challenging quest, since humans do not rate well in numerical terms, and are both unstable and inconsistent when performing such evaluations.

In this work, we analyse whether different comparison measures for binary images are able to discriminate when two boundary images in the BSDS500 are generated from the same original image or not; that is, whether the comparison measure generates larger (dissimilarity) values for inter-class comparisons than for intra-class comparisons. Since this is a task humans can perform with little effort, it can be used as a bare measurement of compliance with human behaviour.

The candidate comparison measures for the separability analysis are the following:

(i) Direct comparison measures: HD_T, Δ_t and SD_t. Here, HD_T is a one-to-one comparison measure derived from the scaled heterogeneity measure: $\text{HD}_T(A,B) = \Psi_T(\{A\},\{B\}) = H_T^*(\{A,B\})$;

(ii) Tree-based ultrametrics: Ultrametrics with cost functions Ψ_T and Γ_c (with c either Δ_t or SD_t). These ultrametrics will be referred to as UMT-Ψ_T, UMT-Γ_{Δ_t} and UMT-Γ_{SD_t}, respectively.

First, we compute the intra-class and inter-class comparisons between each ever two boundary images in the BSDS500 dataset, using the six candidate comparison measures (three of them based on ultrametric trees). The BSDS500 Test set contains 200 different classes, with 1063 images in total. This implies that the number of inter-class comparisons is multiple orders of magnitude larger than the intra-class ones. We also computed the accuracy (Acc) of discrimination of the distributions for each possible threshold. Ideally it should hold that $\text{Acc} = 1$ for at least one threshold, if both distributions were non-overlapping, resulting in total separability.

In Fig. 4 we present the distributions and the accuracy for the BSDS500 Test set of the comparison measures presented in Section 5.2. The upper rows contain the distributions obtained with direct comparison measures (HD_T, Δ_t, SD_t), whereas the lower rows contain those obtained with ultrametric-tree-based comparison measures (UMT-Ψ_T, UMT-Γ_{Δ_t}, UMT-Γ_{SD_t}). Results are replicated with $t \in \{2.5, 5, 10\}$, where t is both the parameter in Δ_t and SD_t and the radius of the structuring element T (a disc) in HD_T and UMT-Ψ_T. Note that due to the large difference between the number of intra- and inter-class comparisons, the distributions are recorded in percentage terms. Note also that distributions take up the left axis, while Acc is expressed on the right one, taking values between 0.5 and 1.

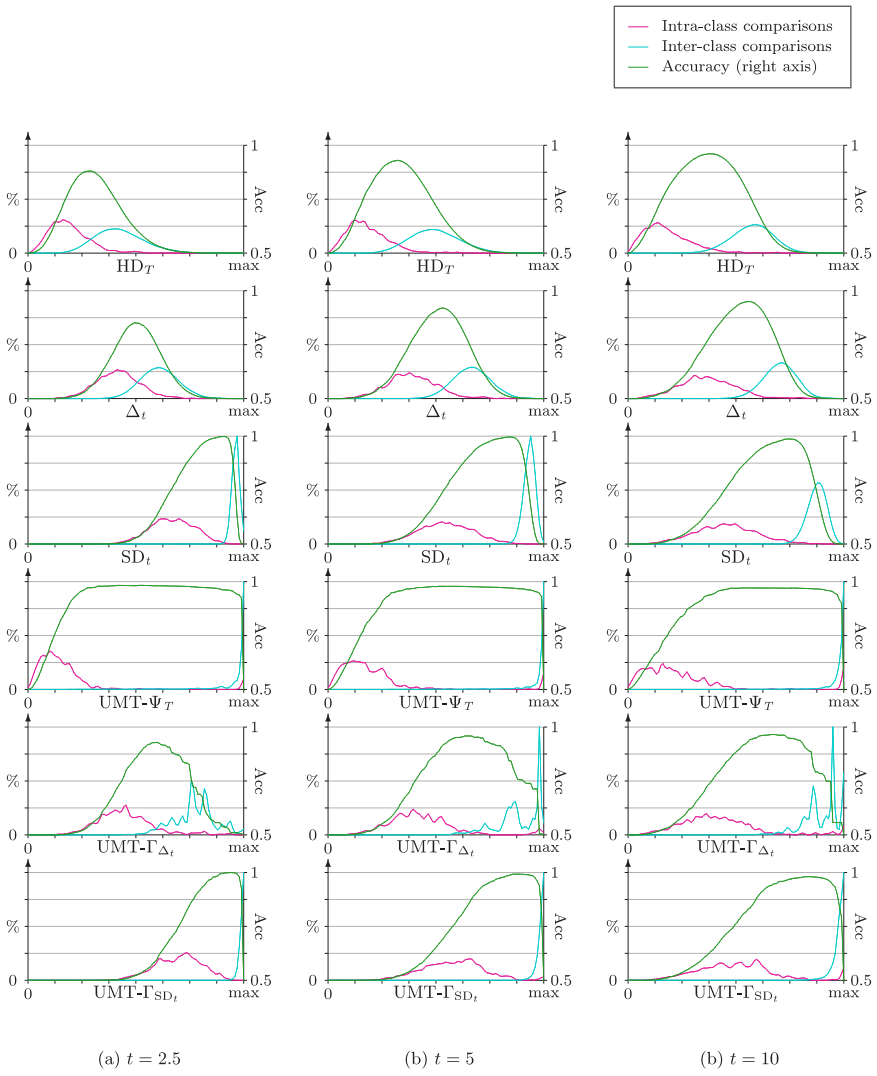


Fig. 4. Distributions and accuracy of the values yielded by HD$_T$, Δ_t, SD$_t$, UMT-Ψ_T, UMT-Γ_{Δ_t} and UMT-Γ_{SD_t} with $t \in \{2.5, 5, 10\}$, for inter- and intra-class pairs of ground truth images in the BSDS500 Test set [49]. Distributions configured with 100 bins.

Overall, the distributions of intra- and inter-class comparisons in Fig. 4 are fairly separable for all comparison measures, whether based on ultrametric trees or not. Intra-class comparisons normally yield lower values than inter-class comparisons for all configurations, replicating human behaviour correctly. Nevertheless, it can be observed at first sight that some comparison measures produce intra- and inter-class distributions that are, visually, more distant than others. It is difficult to appoint the best configuration from the data in Fig. 4. While comparisons based on SD$_t$ peak, in terms of Acc, higher than any other contending comparison measure, other measures (noteworthingly, UMT-Ψ_T) produce a larger area under the curve of Acc.

Focusing on the comparison between standard (direct) and tree-based comparison measures, we find different interesting facts. At first sight, each of the direct comparison measures behaves differently when combined with a tree-based strategy. While UMT-Ψ_T presents a notable improvement w.r.t. HD$_T$, the results by UMT-Γ_{SD_t} and UMT-Γ_{Δ_t} remain almost equal to those by SD$_t$ and Δ_t, respectively. This might indicate that the use of ultrametric trees leads to better or worse results depending on the suitability of the cost function it is based on. Hence, comparison measures such as HD$_T$ (also, Ψ_T), which are intrinsically suitable for N-to-M comparison, perform better when used to produce cost functions within the tree-based strategy, while unsuitable comparison measures (such as Δ_t and SD$_t$) do not benefit clearly from such use.

A more exhaustive analysis of class separability can be based on the *separability criteria* presented in [56], namely the weak, moderate, strong and total separability. These criteria provide an evaluation of the efficiency of the comparison measure mimicking the human behaviour, *i.e.*, their capability to differentiate between intra- and inter-class images.

A dataset for image processing can be modelled as a triplet $\mathbb{D} = (\mathbf{I}, \mathbf{E}, \lambda)$ such that

(i) $\mathbf{I} = \{I_1, \ldots, I_k\} \subseteq \mathbb{G}$ is the set of original images in the dataset;
(ii) $\mathbf{E} = \{E_1, \ldots, E_n\} \subseteq \mathbb{B}$ is the set of ground truth images in the dataset;
(iii) $\lambda : \{1, \ldots, n\} \to \{1, \ldots, k\}$ is a mapping such that $\lambda(i) = j$ if the image E_i was created by a human from image I_j.

Let $\mathbb{D}$ be a dataset and q be a metric or dissimilarity measure used to compare the binary images. The four separability criteria are defined as follows.

(S_1) *Weak separability:* The pair $(\mathbb{D}, q)$ is weakly separable if

$$\min_{\substack{\lambda(i)=\lambda(j)\\ i\neq j}} q(E_i, E_j) \leq \min_{\lambda(i)\neq\lambda(r)} q(E_i, E_r),$$

for all $i \in \{1, \ldots, n\}$.

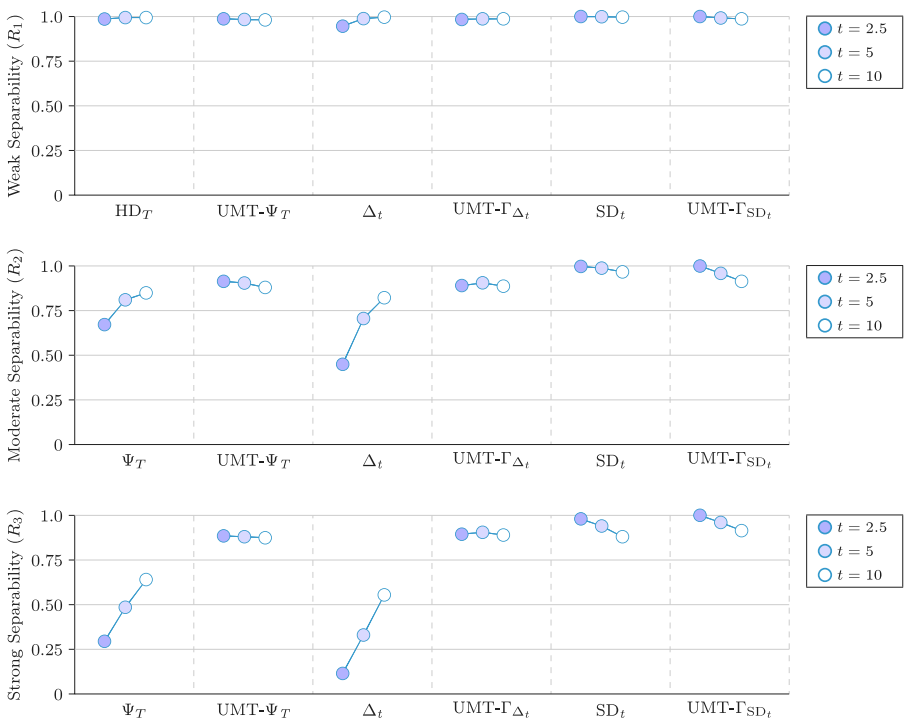


Fig. 5. Separability ratios [56] obtained by different binary image comparison measures on the BSDS500 Test set [49]. Each plot represents a different ratio of separability, while each circle represents one configuration of each specific comparison measure.

(S_2) *Moderate separability:* The pair $(\mathbb{D}, q)$ is moderately separable if

$$\max_{\lambda(i)=\lambda(j)} q(E_i, E_j) \leq \min_{\lambda(i)\neq\lambda(r)} q(E_i, E_r),$$

for all $i \in \{1, \ldots, n\}$.

(S_3) *Strong separability:* The pair $(\mathbb{D}, q)$ is strongly separable if

$$\max_{\lambda(i)=\lambda(j)=m} q(E_i, E_j) \leq \min_{\substack{\lambda(r)=m \\ \lambda(s)\neq m}} q(E_r, E_s),$$

for all $m \in \{1, \ldots, k\}$.

(S_4) *Total separability:* The pair $(\mathbb{D}, q)$ is totally separable if

$$\max_{\lambda(i)=\lambda(j)} q(E_i, E_j) \leq \min_{\lambda(r)\neq\lambda(s)} q(E_r, E_s).$$

The criteria are presented in increasing severity. In this way, for any pair $(\mathbb{D}, q)$, it holds that

$$S_4 \Rightarrow S_3 \Rightarrow S_2 \Rightarrow S_1 .$$

In Fig. 5, we present the separability ratios on the BSDS500 Test set, using the six comparison measures in Fig. 4, each configured with $t \in \{2.5, 5, 10\}$. We observe that increasing the severity of the criteria entails an expected decrease of the associated ratio. Overall, the results in Fig. 5 are in line with those in Fig. 4: Well-tailored cost functions (such as Ψ_T) lead to an improvement in the separability criteria when using ultrametric trees. Cost functions artificially prepared for N-to-M comparisons can lead to no or little gain (UMT-Γ_{SD_t}, UMT-Γ_{Δ_t}). Still, considering the separability analysis in its entirety, we can state that UMT-Ψ_T is preferred over measures built on ultrametric tress using comparison measures that are unsuitable for N-to-M comparisons (such as Δ_t and SD_t), on the basis of the greater area-under-the-curve in terms of Acc in Fig. 4. The results seem to confirm the intuition that comparison measures that actually model information at each non-leaf node, as HD_T, perform more according to human behaviour, and hence obtain better results in the task of class discrimination. However, those that produce no intermediate representation at such nodes fail to make use of coincidental and divergent information in the images at each node.

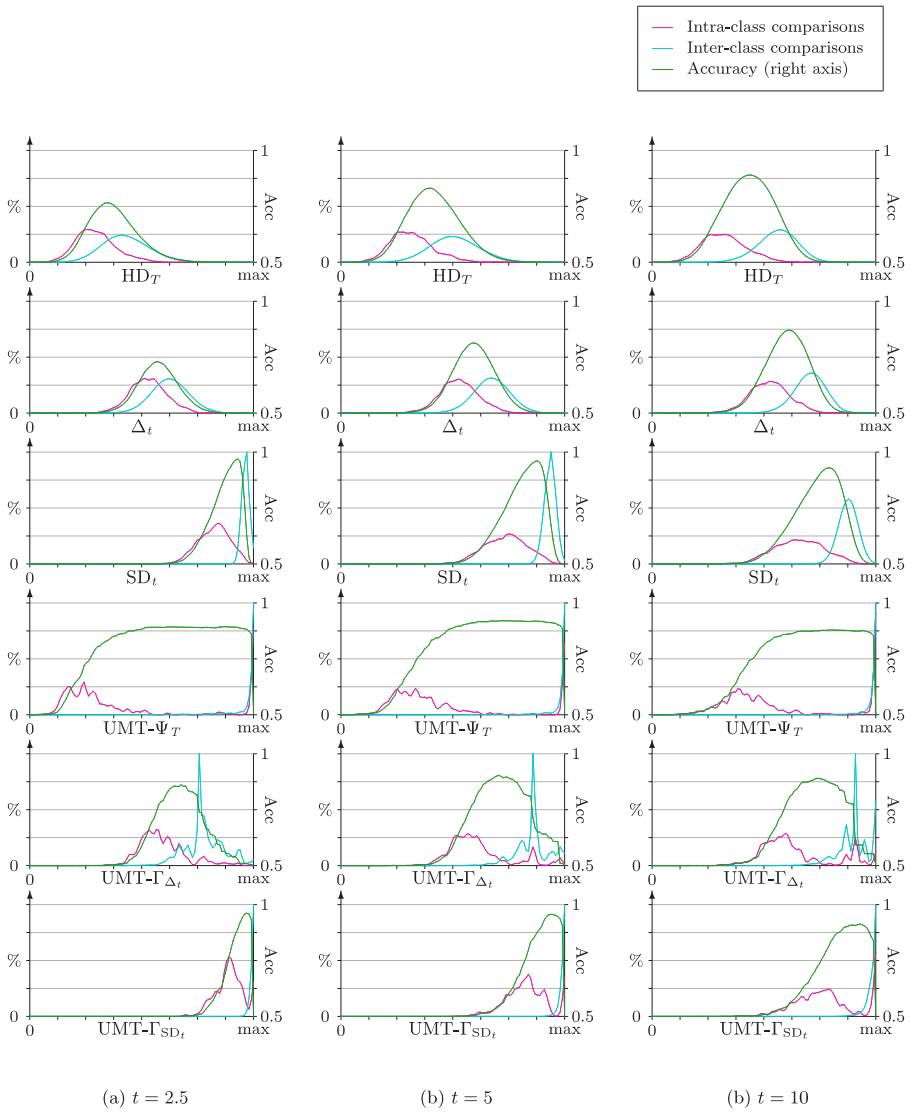
5.4. Experiments on computer-generated boundary images

The results obtained so far have been based on the BSDS500 Test set, specifically on human-made ground truth images. This sheds doubt on whether the conclusions can be ported to scenarios in which the images to be compared are computer-generated. Note that, while edge detection and segmentation methods produce boundary images relatively similar to those by humans, they normally incur in different types of errors. While humans are prone to missing objects or (slightly) displacing the marked boundaries from their actual positions, computer-based methods often fall into texture misinterpretation and false positives on irrelevant objects.

In order to confirm that our conclusions are valid regardless of the origin of the images to be compared, the experiments have been repeated using computer-generated boundary images. Specifically, we have used the state-of-the-art, superpixel-based method by Lei et al. (SFFCM, [57]). This method generates a hierarchical interpretation of the regions in an image, so it can produce different boundary images (more precisely, boundary images at different levels of clustering/refinement). We have generated six different boundary images per original image in the BSDS500 Test set, totalling 1200 computer-generated images.

Fig. 6 replicates the configuration in Fig. 4, but displays the results for the comparison of human-made to computer-generated images. The intra-class distribution is composed of the comparisons between two images: a human-made and a computer-generated image, both created from the same original image. Equivalently, the inter-class distribution is composed of the comparisons between two images: a human-made and a computer-generated image, created from different original images. Note that, as in Fig. 4, due to the large difference between the number of intra- and inter-class comparisons, the distributions are presented in percentage terms on the left axis. Note also that the Acc is displayed on the right axis, taking values between 0.5 and 1.

In Fig. 6, we observe that the results are coincident with those derived from Fig. 4. Due to the increased number of imperfections in computer-generated images (w.r.t. human-made images), the distributions are (visually) more overlapping. In fact, Acc is generally



(a) $t = 2.5$ (b) $t = 5$ (b) $t = 10$

Fig. 6. Distributions and accuracy of the values yielded by HD_T, Δ_t, SD_t, UMT-Ψ_T, UMT-Γ_{Δ_t} and UMT-Γ_{SD_t} with $t \in \{2.5, 5, 10\}$, for inter- and intra-class pairs of ground truth images in the BSDS500 Test set [49] and automatically-generated boundary images using the SFFCM method [57]. Distributions are configured with 100 bins.

lower in Fig. 6 than in Fig. 4. Nevertheless, the main conclusions still stand. Incorporating context-awareness, through the use of ultrametric trees, is advisable when the cost functions are prepared for N-to-M comparisons. Applying comparison measures ill-prepared for N-to-M comparisons using ultrametric trees, and producing a context-aware ultrametric, provides little or no gain in performance. In the case of HD_T, however, the separability improves significantly when applying ultrametric trees (UMT-Ψ_T) instead of direct comparisons (HD_T), reaching a similar, if not higher peak in Acc and, very importantly, a larger area under its curve.

6. Conclusions

In this work we introduced a context-aware comparison measure based on ultrametric trees for binary images. In order to obtain quantitative N-to-M comparisons, we applied the Twofold Consensus Ground Truth (TCGT) [23], more specifically the resulting scaled heterogeneity, avoiding the strategy of reusing 1-to-1 comparison measures. We applied our algorithm to the BSDS500 Test set [49], and performed a separability analysis, obtaining the corresponding accuracy and ratios.

As a conclusion, we can affirm that the construction of an ultrametric tree applying simple mathematical notions allows for context modelling in binary image comparison. Nevertheless, such construction must be supported by a cost function that provides a quantification for the selection of node pairs and shall be carefully studied. Using the TCGT allows for comparisons involving different cardinalities, while providing high separability values. Hence we can state that our algorithm not only provides a context-aware comparison measure, but also a better replication of human behaviour.

Context-aware ultrametrics could be used to compare multichannel images, such as colour or multi/hyperspectral ones. Such application should, in fact, leverage the use of tools that are barely studied for binary images. For example, the modelling of the information at each node, supporting M-to-N comparisons, could be based on image fusion (hence representing each node with one or several fused images). As stated by Ghassemian, *the number of scientific papers [on image fusion] published in the international journals increases dramatically since 2010* [58]. While many of such works are context-specific (e.g., remote sensing [58,59]), the diversity of solutions for colour image fusion [60, 61] hints at a vast potential for modelling node information and subsequent generation of M-to-N comparison operators. This potential could render into improvements in tasks as image retrieval [62], which is fundamentally based on image comparison.

CRediT authorship contribution statement

C. Lopez-Molina: Conceptualization, Validation, Methodology, Formal analysis, Writing – review & editing, Supervision. **S. Iglesias-Rey:** Writing – review & editing, Software, Investigation. **B. De Baets:** Conceptualization, Methodology, Formal analysis, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

The authors gratefully acknowledge the financial support of the Spanish Research Agency, project PID2019-108392GB-I00 (AEI/10.13039/501100011033), as well as that of Navarra de Servicios y Tecnologías, S.A. (NASERTIC).

References

[1] F.J. Cardama, D.B. Heras, F. Argüello, Consensus techniques for unsupervised binary change detection using multi-scale segmentation detectors for land cover vegetation images, Remote Sens. 15 (11) (2023) 2889.
[2] R. Unnikrishnan, C. Pantofaru, M. Hebert, Toward objective evaluation of image segmentation algorithms, IEEE Trans. Pattern Anal. Mach. Intell. 29 (6) (2007) 929–944.
[3] J. Pont-Tuset, F. Marques, Measures and meta-measures for the supervised evaluation of image segmentation, in: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, 2013, pp. 2131–2138.
[4] F. Bardozzo, B. De La Osa, L. Horanská, J. Fumanal-Idocin, M. delli Priscoli, L. Troiano, R. Tagliaferri, J. Fernandez, H. Bustince, Sugeno integral generalization applied to improve adaptive image binarization, Inf. Fusion 68 (2021) 37–45.
[5] S. Guo, X. Liu, H. Zhang, Q. Lin, L. Xu, C. Shi, Z. Gao, A. Guzzo, G. Fortino, Causal knowledge fusion for 3D cross-modality cardiac image segmentation, Inf. Fusion (2023) 101864.
[6] C. Yu, S. Li, D. Ghista, Z. Gao, H. Zhang, J. Del Ser, L. Xu, Multi-level multi-type self-generated knowledge fusion for cardiac ultrasound segmentation, Inf. Fusion 92 (2023) 1–12.
[7] C. Lopez-Molina, B. De Baets, H. Bustince, Quantitative error measures for edge detection, Pattern Recognit. 46 (4) (2013) 1125–1139.
[8] J.K. Udupa, V.R. LeBlanc, Y. Zhuge, C. Imielinska, H. Schmidt, L.M. Currie, B.E. Hirsch, J. Woodburn, A framework for evaluating image segmentation algorithms, Comput. Med. Imaging Graph. 30 (2) (2006) 75–87.
[9] A.A. Taha, A. Hanbury, Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool, BMC Med. Imaging 15 (1) (2015) 1–28.
[10] D. Huttenlocher, G. Klanderman, W. Rucklidge, Comparing images using the Hausdorff distance, IEEE Trans. Pattern Anal. Mach. Intell. 15 (9) (1993) 850–863.
[11] D. Brunet, D. Sills, A generalized distance transform: Theory and applications to weather analysis and forecasting, IEEE Trans. Geosci. Remote Sens. 55 (3) (2016) 1752–1764.
[12] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, 2015, arXiv preprint 1505.04597.
[13] N. Ibtehaz, M.S. Rahman, MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation, Neural Netw. 121 (2020) 74–87.
[14] D. Karimi, S.E. Salcudean, Reducing the hausdorff distance in medical image segmentation with convolutional neural networks, IEEE Trans. Med. Imaging 39 (2) (2019) 499–513.
[15] C. Gros, A. Lemay, J. Cohen-Adad, SoftSeg: Advantages of soft versus binary training for image segmentation, Med. Image Anal. 71 (2021) 102038.
[16] D.-G. Sim, O.-K. Kwon, R.-H. Park, Object matching algorithms using robust Hausdorff distance measures, IEEE Trans. Image Process. 8 (3) (1999) 425–429.
[17] J. Gimenez, J. Martinez, A.G. Flesia, Unsupervised edge map scoring: A statistical complexity approach, Comput. Vis. Image Underst. 122 (2014) 131–142.
[18] S.K. Warfield, K.H. Zou, W.M. Wells, Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation, IEEE Trans. Med. Imaging 23 (7) (2004) 903–921.
[19] A. Asman, B. Landman, Robust statistical label fusion through consensus level, labeler accuracy and truth estimation (COLLATE), IEEE Trans. Med. Imaging 30 (2011) 1779–1794.
[20] J. Fram, E.S. Deutsch, On the quantitative evaluation of edge detection schemes and their comparison with human performance, IEEE Trans. Comput. 24 (6) (1975) 616–628.
[21] A. Tversky, Features of similarity, Psychol. Rev. 84 (4) (1977) 327–352.
[22] A. Tversky, Preference, Belief, and Similarity : Selected Writings By Amos Tversky; Edited By Eldar Shafir, MIT Press, Cambridge, Massachusetts, 2004.
[23] C. Lopez-Molina, B. De Baets, H. Bustince, Twofold consensus for boundary detection ground truth, Knowl.-Based Syst. 98 (2016) 162–171.
[24] S. Iglesias-Rey, A. Castillo-Lopez, C. Lopez-Molina, B. De Baets, On the role of context-awareness in binary image comparison, in: Proc. of the Hawaii International Conference on System Sciences, HICSS, 2022.
[25] T. Faver, K. Kochalski, M.K. Murugan, H. Verheggen, E. Wesson, A. Weston, Roundness properties of ultrametric spaces, Glasg. Math. J. 56 (3) (2014) 519–535.
[26] B. De Baets, R. Mesiar, Pseudo-metrics and T-equivalences, Fuzzy Math 5 (1997) 471–481.
[27] B. De Baets, R. Mesiar, Metrics and T-equalities, Math. Anal. Appl. 267 (2) (2002) 531–547.
[28] M. Dubuisson, A.K. Jain, A modified hausdorff distance for object matching, in: Proc. on Pattern Recognition, Vol. 1, 1994, pp. 566–568.
[29] B. Takács, Comparing face images using the modified Hausdorff distance, Pattern Recognit. 31 (12) (1998) 1873–1881.
[30] F. Perronnin, Y. Liu, J.-M. Renders, A family of contextual measures of similarity between distributions with application to image retrieval, 2009, pp. 2358–2365.
[31] C. Wang, G. Peng, B. De Baets, Deep feature fusion through adaptive discriminative metric learning for scene recognition, Inf. Fusion 63 (2020) 1–12.
[32] V.E. Liong, J. Lu, Y. Ge, Regularized local metric learning for person re-identification, Pattern Recognit. Lett. 68 (2015) 288–296.
[33] Y. Yamashita, K. Shimosato, N. Ukita, Boundary-aware image inpainting with multiple auxiliary cues, in: Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 2022, pp. 619–629.
[34] J. Mahmud, T. Price, A. Bapat, J.-M. Frahm, Boundary-aware 3D building reconstruction from a single overhead image, in: Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 2020.
[35] G. Vidal-Diez de Ulzurrun, J. Baetens, J. Van den Bulcke, B. De Baets, Modelling three-dimensional fungal growth in response to environmental stimuli, J. Theoret. Biol. 414 (2017) 35–49.
[36] D. Crevier, Image segmentation algorithm development using ground truth image data sets, Comput. Vis. Image Underst. 112 (2008) 143–159.
[37] I. Abdou, W. Pratt, Quantitative design and evaluation of enhancement/thresholding edge detectors, in: Proc. of the IEEE, Vol. 67, 1979, pp. 753–763.
[38] R.M. Haralick, Digital step edges from zero crossing of second directional derivatives, IEEE Trans. Pattern Anal. Mach. Intell. 6 (1) (1984) 58–68.
[39] B. Peng, L. Zhang, Evaluation of image segmentation quality by adaptive ground truth composition, in: Proc. of the European Conference on Computer Vision, Springer, 2012, pp. 287–300.
[40] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE Trans. Pattern Anal. Mach. Intell. 24 (4) (2002) 509–522.
[41] R.M. Haralick, J.S.J. Lee, Context dependent edge detection and evaluation, Pattern Recognit. 23 (1–2) (1990) 1–19.
[42] S. Goferman, L. Zelnik-Manor, A. Tal, Context-aware saliency detection, IEEE Trans. Pattern Anal. Mach. Intell. 34 (10) (2012) 1915–1926.
[43] H. Jiang, J. Wang, Z. Yuan, T. Liu, N. Zheng, S. Li, Automatic salient object segmentation based on context and shape prior, in: Proc. British Machine Vision Conference, BMVC, 2011, pp. 110.1–110.12.
[44] S. Bordag, A comparison of co-occurrence and similarity measures as simulations of context, in: A. Gelbukh (Ed.), Computational Linguistics and Intelligent Text Processing, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 52–63.
[45] C. Keßler, Similarity measurement in context, in: B. Kokinov, D.C. Richardson, T.R. Roth-Berghofer, L. Vieu (Eds.), Modeling and using Context, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007, pp. 277–290.
[46] A. Tversky, I. Simonson, Context-dependent preferences, Manage. Sci. 39 (10) (1993) 1179–1189.
[47] M. Bazire, P. Brézillon, Understanding context before using it, in: A. Dey, B. Kokinov, D. Leake, R. Turner (Eds.), Modeling and using Context, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005, pp. 29–40.
[48] D. Martin, C. Fowlkes, J. Malik, Learning to detect natural image boundaries using local brightness, color, and texture cues, IEEE Trans. Pattern Anal. Mach. Intell. 26 (5) (2004) 530–549.
[49] P. Arbelaez, M. Maire, C. Fowlkes, J. Malik, Contour detection and hierarchical image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2011) 898–916.
[50] D.A. Mély, J. Kim, M. McGill, Y. Guo, T. Serre, A systematic comparison between visual cues for boundary detection, Vis. Res. 120 (2016) 93–107.
[51] Y. Yitzhaky, E. Peli, A method for objective edge detection evaluation and detector parameter selection, IEEE Trans. Pattern Anal. Mach. Intell. 25 (8) (2003) 1027–1033.

[52] N. Fernández-García, A. Carmona-Poyato, R. Medina-Carnicer, F. Madrid-Cuevas, Automatic generation of consensus ground truth for the comparison of edge detection techniques, Image Vis. Comput. 26 (4) (2008) 496–511.
[53] G. Papari, N. Petkov, Edge and line oriented contour detection: State of the art, Image Vis. Comput. 29 (2–3) (2011) 79–103.
[54] A.J. Baddeley, An error metric for binary images, in: W. Förstner, S. Ruwiedel (Eds.), Robust Computer Vision: Quality of Vision Algorithms, Wichmann Verlag, Karlsruhe, 1992, pp. 59–78.
[55] A.J. Baddeley, Errors in binary images and an L^p version of the Hausdorff metric, Nieuw Arch. Wiskd. 10 (1992) 157–183.
[56] C. Lopez-Molina, H. Bustince, B. De Baets, Separability criteria for the evaluation of boundary detection benchmarks, IEEE Trans. Image Process. 25 (3) (2016) 1047–1055.
[57] T. Lei, X. Jia, Y. Zhang, S. Liu, H. Meng, A.K. Nandi, Superpixel-based fast fuzzy C-means clustering for color image segmentation, IEEE Trans. Fuzzy Syst. 27 (9) (2019) 1753–1766.
[58] H. Ghassemian, A review of remote sensing image fusion methods, Inf. Fusion 32 (2016) 75–89.
[59] G. Vivone, Multispectral and hyperspectral image fusion in remote sensing: A survey, Inf. Fusion 89 (2023) 405–417.
[60] S. Li, X. Kang, J. Hu, Image fusion with guided filtering, IEEE Trans. Image process. 22 (7) (2013) 2864–2875.
[61] H. Zhang, H. Xu, X. Tian, J. Jiang, J. Ma, Image fusion meets deep learning: A survey and perspective, Inf. Fusion 76 (2021) 323–336.
[62] U. Sharif, Z. Mehmood, T. Mahmood, M.A. Javid, A. Rehman, T. Saba, Scene analysis and search using local features and support vector machine for effective content-based image retrieval, Artif. Intell. Rev. 52 (2019) 901–925.