2

Character Recognition with Metasets

Bartłomiej Starosta
Polish-Japanese Institute of Information Technology
Poland

1. Introduction

The chapter presents a new approach to the character recognition problem. It is based on metasets – a new concept of sets with partial membership relation. By the character recognition problem we understand determining the similarity degree of the given character sample to the defined character pattern. The discussed mechanism may be applied not only to characters (e.g. letters), but to arbitrary data represented on monochromatic images or even multi-dimensional figures.
The theory of metasets brings a new model of "fuzzy" membership relation for sets. A metaset may be a member of (or equal to) another metaset to variety of different degrees – contrary to classical sets where membership and equality are always either true or false.
The goal of the chapter is to present the application of the new, abstract theory to solving a practical, well-known problem. It develops the method which was partially introduced for some particular case in (Starosta, 2009). The proposed solution had been implemented as a computer program. The experiments made with the program confirm that the theoretical assumptions are correct and the obtained results properly reflect our perception of similarity of characters. It should also be stressed that the concept of metaset itself was partially inspired by another computer application for character recognition, based on neural networks.

1.1 The general idea

The process of determining the similarity degree consists in two stages. Initially, the compound character pattern must be prepared. It consists of several character samples accompanied by quality grades. The samples are depicted on rectangular matrices and they correspond to different forms of the same character. The pattern itself represents various possible approaches to the same character, as a single entity. In the second stage a testing character sample is matched against the pattern and the resulting similarity degree is calculated.
The character samples as well as the compound pattern are encoded as metasets. As the result of matching the testing sample against the pattern we obtain the membership degree of the sample metaset in the pattern metaset and additionally, the sequence of equality degrees of the sample metaset and the pattern elements. The membership degree measures how far the sample resembles the pattern. The equality degrees indicate the similarity of the input sample and each pattern element separately. The membership degrees as well as equality degrees for metasets are expressed as sets of nodes of the binary tree, which are finite binary sequences, and they may be evaluated as real numbers.

The quality grades of the samples in the pattern are membership degrees of the corresponding metasets, too. However, they are manually specified as areas of the matrix for depicting the characters, which contain valid pixels to be included in the matching process. This specification is interpreted as membership degrees of appropriate metasets. The quality grades show how close is a particular sample to the ideal. They may be supplied by experts together with the samples.
The most significant innovation here is treating the membership and equality degrees of metasets as similarity measures for characters provided they are properly encoded as metasets.

1.2 Basic terms and notation

The concept of binary tree plays the key role in the definition of metaset and related notions. Therefore, we start with establishing some well known terms and notation concerning it.
We use the symbol $\mathbb{T}$ for the infinite binary tree with the root $\mathbb{1}$. The nodes of the tree $\mathbb{T}$ are finite binary sequences, the root $\mathbb{1}$ is the empty sequence. For $p \in \mathbb{T}$ the symbol $|p|$ denotes the length of the sequence and $\#p$ denotes the natural number represented by the binary sequence p. Note, that $|\mathbb{1}| = 0$ and we assume $\#\mathbb{1} = 0$. The ordering of nodes in $\mathbb{T}$ is determined by reverse ordering of their lengths: $p \leq q$ whenever $|p| \geq |q|$. In particular the root $\mathbb{1}$ is the largest element in $\mathbb{T}$. The set of nodes of equal length n is called the n-th *level* in the tree: $\mathbb{T}_n = \{\, p \in \mathbb{T} \colon |p| = n \,\}$. The level 0 contains only the root. Nodes of the tree $\mathbb{T}$ are sometimes called *conditions*. If $p \leq q \in \mathbb{T}$, then we say that the condition p is *stronger* than the condition q, and q is *weaker* than p. Thus, the conditions 0 and 1 are stronger than the root $\mathbb{1}$ and they are weaker than the conditions 00, 01, 10, 11, which form the level $\mathbb{T}_2$.

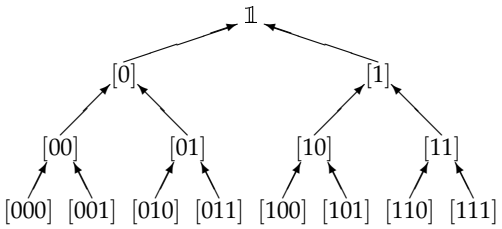

Fig. 1. The binary tree $\mathbb{T}$ and the ordering of nodes (conditions). Arrows point at the larger element, i.e., the weaker condition

A set of nodes $C \subset \mathbb{T}$ is called a *chain* in $\mathbb{T}$, whenever all its elements are pairwise comparable: $\forall_{p,q \in C}\,(p \leq q \vee q \leq p)$. A set $A \subset \mathbb{T}$ is called *antichain* in $\mathbb{T}$, if it consists of mutually incomparable elements: $\forall_{p,q \in A}\,(p \neq q \rightarrow \neg\,(p \leq q) \wedge \neg\,(p \geq q))$. On the Fig. 1, the elements $\{\, 00, 01, 100 \,\}$ form a sample antichain. A *maximal antichain* is an antichain which cannot be extended by adding new elements – it is a maximal element with respect to inclusion of antichains. Examples of maximal antichains on the Fig. 1 are $\{\, 0, 1 \,\}$ or $\{\, 00, 01, 1 \,\}$ or even $\{\, \mathbb{1} \,\}$. They are in fact maximal finite antichains (MFA). A *branch* is a maximal chain in the tree $\mathbb{T}$. Note that p is comparable to q only, if there exists a branch containing p and q simultaneously. Similarly, p is incomparable to q, when no branch contains both p and q.
To finish this section we prove a property of maximal finite antichains necessary for evaluating as numbers the degrees represented as sets of nodes. Clearly, there are 2^n nodes on the n-th level of the binary tree, so $\sum_{p \in \mathbb{T}_n} \frac{1}{2^{|p|}} = 1$. This property may be generalized to arbitrary MFA.

Lemma 1. *If $A \subset \mathbb{T}$ is a maximal finite antichain in $\mathbb{T}$, then* $\sum_{p\in A} \frac{1}{2^{|p|}} = 1$.

Proof. Each node $p \neq \mathbb{1}$ is a binary sequence which represents a natural number $\#p$. Therefore, each $p \neq \mathbb{1}$ corresponds to an interval $\bar{p} = [\frac{\#p}{2^{|p|}} \dots \frac{\#p+1}{2^{|p|}}) \subset [0 \dots 1]$ and $\mathbb{1}$ corresponds to $I = [0 \dots 1)$. The length of each interval is $\frac{1}{2^{|p|}}$. For incomparable p and q, the corresponding intervals are disjoint: $\bar{p} \cap \bar{q} = \emptyset$. Indeed, if $\bar{p} \cap \bar{q} \neq \emptyset$, then there must exist some $r \in \mathbb{T}$ such, that $\bar{r} \subset \bar{p} \cap \bar{q}$. Since $\bar{r} \subset \bar{p}$, then $r \leq p$, and similarly $r \leq q$. This implies $p \leq q$ or $q \leq p$, so they are comparable.
We now show, that the measure of $\bigcup_{p\in A} \bar{p}$ is equal 1. Clearly, it cannot be grater than 1, so if it is less, then let $u \subset I \setminus \bigcup_{p\in A} \bar{p}$ be an open interval. There must exist $s \in \mathbb{T}$ such, that $\bar{s} \subset u$. If s is comparable to some $p \in A$, then $\bar{s} \cap \bar{p} \neq \emptyset$, so $\bar{s} \cap \bigcup_{p\in A} \bar{p}$ is non-empty, what contradicts $\bar{s} \subset u$. Thus, assuming that the length of $\bigcup_{p\in A} \bar{p}$ is less than 1 we found s incomparable to all elements of A, what contradicts its maximality.
To complete the proof note, that the length of each $\bar{p}$ is $\frac{1}{2^{|p|}}$, the measure of $\bigcup_{p\in A} \bar{p}$ is 1 and they are all pairwise disjoint. □

2. Metasets

In the classical set theory a set either is an element of another set or it is not; there are no intermediate levels. This binary approach has many vital limitations which make it difficult to apply by representation of vague, imprecise data. Therefore, for the last decades there were several attempts to inventing a concept of set with partial membership relation. Among the most successful ones are fuzzy sets (Zadeh, 1965), intuitionistic fuzzy sets (Atanassov, 1986) and rough sets (Pawlak, 1982). The metaset idea is a new approach to the problem.
One of the most significant characteristics of the metaset concept is its computer oriented design. Definitions of fundamental notions – like membership, equality or algebraic operations – may be formulated in the way which makes them easily implementable using programming languages (Starosta & Kosiński, 2009). This facilitates fast and efficient computer representation and processing of vague data. Additionally, several important theoretical results may be obtained for the metasets which are representable in computers, because of their finite structure. Some of them – like the Lemma 3 – constitute the base for the discussed here mechanism.

2.1 Fundamental concepts

The concept of metaset is strictly based on the classical Zermelo-Fraenkel set theory (ZFC). We define metaset as a set of ordered pairs. The first element of a pair is a member of the metaset, which is another metaset. The second element of the pair is a node of the binary tree which – informally speaking – specifies the membership degree of the first element in the metaset.

Definition 1. A metaset is a crisp set which is either the empty set $\emptyset$ or which has the form:

$$\tau = \{ \langle \sigma, p \rangle : \sigma \text{ is a metaset, } p \in \mathbb{T} \} \quad .$$

The definition is recursive, however it is founded by the empty set $\emptyset$, by the Axiom of Foundation in ZFC (Kunen, 1980). First elements of ordered pairs contained in the metaset are called its *potential elements*.

From the classical set theory point of view, a meta set is a relation between a crisp set of other meta sets and a set of nodes of the tree $\mathbb{T}$. Therefore, we adopt some terminology associated with relations. For the given metaset τ the set of its potential elements:

$$\mathrm{dom}(\tau) = \{\, \sigma\colon\ \langle \sigma, p\rangle \in \tau \,\} \tag{1}$$

is called the *domain* of the metaset τ. Its *range* is the following set:

$$\mathrm{ran}(\tau) = \{\, p\colon\ \langle \sigma, p\rangle \in \tau \,\} \quad . \tag{2}$$

The reader may confirm that $\tau \subset \mathrm{dom}(\tau) \times \mathrm{ran}(\tau) \subset \mathrm{dom}(\tau) \times \mathbb{T}$. For metasets τ and σ the set

$$\tau[\sigma] = \{\, p \in \mathbb{T}\colon\ \langle \sigma, p\rangle \in \tau \,\} \tag{3}$$

is called the *image* of the metaset τ at the metaset σ. The image $\tau[\sigma]$ is the empty set $\emptyset$, whenever σ is not a potential element of τ.

Example 1. The simplest metaset is the empty set $\emptyset$. It may be a potential element of other metasets:

$$\begin{array}{llll}
\tau = \{\, \langle \emptyset, p\rangle \,\}\ , & \tau[\emptyset] = \{\, p \,\}\ , & \mathrm{dom}(\tau) = \{\, \emptyset \,\}\ , & \mathrm{ran}(\tau) = \{\, p \,\}\ , \\
\sigma = \{\, \langle \emptyset, p\rangle\, , \langle \emptyset, q\rangle \,\}\ , & \sigma[\emptyset] = \{\, p, q \,\}\ , & \mathrm{dom}(\sigma) = \{\, \emptyset \,\}\ , & \mathrm{ran}(\sigma) = \{\, p, q \,\}\ . \\
\eta = \{\, \langle \tau, p\rangle\, , \langle \sigma, q\rangle \,\}\ , & \eta[\emptyset] = \emptyset\ , & \mathrm{dom}(\eta) = \{\, \tau, \sigma \,\}\ , & \mathrm{ran}(\eta) = \{\, p, q \,\}\ .
\end{array}$$

Clearly, $\eta[\tau] = p$, $\eta[\sigma] = q$ and since $\emptyset \notin \mathrm{dom}(\eta)$, then $\eta[\emptyset] = \emptyset$.

In this paper we do not deal with metasets in general. We focus here on very specific classes relevant to character recognition problem. Narrowing the domain of discourse simplifies formulations of some results too. We introduce now two classes of metasets used for representation of characters and patterns.
Let A be a maximal finite antichain in $\mathbb{T}$. A non-empty metaset of form

$$\chi \subset \{\, \emptyset \,\} \times A \tag{4}$$

is called *A-sample* metaset. Each non-empty subset $S \subset A$ determines A-sample metaset $\{\, \emptyset \,\} \times S$. A-sample metasets are used for representing character samples.
Let P be a finite set of A-sample metasets. A non-empty metaset of form

$$\pi \subset P \times A \tag{5}$$

is called *A-pattern* metaset. In other words, A-pattern metaset has the form

$$\pi = \bigcup_{i=1}^{i=n} \{\, \chi_i \,\} \times P_i \tag{6}$$

where χ_i are A-sample metasets and $P_i \subset A$, are not empty for $i = 1, \ldots, n$. A-pattern metasets are used for representing character patterns.
We now explain the fundamental technique of interpretation used for defining relations on metasets. Also, it allows to perceive a metaset as a "fuzzy" family of crisp sets. Each member of such family represents some specific, particular point of view on the metaset.

Definition 2. Let τ be a metaset and let $\mathcal{C}$ be a branch in the binary tree $\mathbb{T}$. The interpretation of the metaset τ, given by the branch $\mathcal{C}$, is the following crisp set:

$$\tau_{\mathcal{C}} = \{\, \sigma_{\mathcal{C}} \colon \langle \sigma, p \rangle \in \tau \wedge p \in \mathcal{C} \,\} \quad .$$

Thus, branches in $\mathbb{T}$ allow for producing crisp sets out of the metaset. The family of crisp sets $\{\, \tau_{\mathcal{C}} \colon \mathcal{C}$ is a branch in $\mathbb{T} \,\}$ consists of interpretations of the metaset τ. Properties of these interpretations determine properties of the metaset.
Any interpretation of the empty metaset is the empty set itself, independently of the branch: $\emptyset_{\mathcal{C}} = \emptyset$, for each $\mathcal{C} \subset \mathbb{T}$. The process of producing the interpretation of a metaset consists in two stages. In the first stage we remove all the ordered pairs whose second elements are conditions which do not belong to the branch $\mathcal{C}$. The second stage replaces the remaining pairs – whose second elements lie on the branch $\mathcal{C}$ – with interpretations of their first elements, which are other metasets. This two-stage process is repeated recursively on all the levels of the membership hierarchy. As the result we obtain a crisp set.

Example 2. Let $p \in \mathbb{T}$ and let $\tau = \{\, \langle \emptyset, p \rangle \,\}$. If $\mathcal{C}$ is a branch, then

$$\begin{aligned} &p \in \mathcal{C} \rightarrow \tau_{\mathcal{C}} = \{\, \emptyset_{\mathcal{C}} \,\} = \{\, \emptyset \,\} \ , \\ &p \notin \mathcal{C} \rightarrow \tau_{\mathcal{C}} = \emptyset \ . \end{aligned}$$

Depending on the branch the metaset τ acquires different interpretations.

An interpretation of A-sample metaset is either the empty set $\emptyset$ or the singleton $\{\, \emptyset \,\}$. An interpretation of A-pattern metaset $\eta = \{\, \langle \sigma, p \rangle \,\}$, where σ is A-sample metaset, is given by

$$\eta_{\mathcal{C}} = \begin{cases} \emptyset & p \notin \mathcal{C} \ , \\ \{\, \emptyset \,\} & p \in \mathcal{C} \wedge \mathrm{ran}(\sigma) \cap \mathcal{C} = \emptyset \ , \\ \{\, \{\, \emptyset \,\} \,\} & p \in \mathcal{C} \wedge \mathrm{ran}(\sigma) \cap \mathcal{C} \neq \emptyset \ . \end{cases} \tag{7}$$

Therefore, an interpretation of any A-pattern metaset is one of: $\emptyset$, $\{\, \emptyset \,\}$, $\{\, \{\, \emptyset \,\} \,\}$ or $\{\, \emptyset, \{\, \emptyset \,\} \,\}$. For instance, if $\nu = \{\, \langle \emptyset, 0 \rangle \,\}$, $\mu = \{\, \langle \emptyset, 111 \rangle \,\}$, $\tau = \{\, \langle \nu, 1 \rangle, \langle \mu, 11 \rangle \,\}$ and $\mathcal{C} = \{\, \mathbb{1}, 1, 11, 111, \dots \,\}$ is the rightmost branch, then $\nu_{\mathcal{C}} = \emptyset$, $\mu_{\mathcal{C}} = \{\, \emptyset \,\}$, so $\tau_{\mathcal{C}} = \{\, \emptyset, \{\, \emptyset \,\} \,\}$.
We introduce now basic set-theoretic relations for metasets. All the relations are defined using the same scheme – by referring to interpretations. We start with the membership.

Definition 3. Let τ, σ be metasets and let $p \in \mathbb{T}$. We say that σ belongs to τ under the condition p, if for each branch $\mathcal{C}$ containing p holds $\sigma_{\mathcal{C}} \in \tau_{\mathcal{C}}$. We use the notation $\sigma \, \epsilon_p \, \tau$.

Note, that in fact we define an infinite number of membership relations here – each designated with different condition. The membership under the root condition $\sigma \, \epsilon_{\mathbb{1}} \, \tau$ corresponds to the crisp, classical membership. The $\mathbb{1}$ designates the highest membership degree, since it is the largest element in $\mathbb{T}$. Stronger conditions designate lower degrees of membership.
We also define an independent set of non-membership relations. The reason for this lies in the fact, that $\neg \, \sigma \, \epsilon_p \, \tau$ does not imply that for each branch $\mathcal{C}$ containing p holds $\sigma_{\mathcal{C}} \notin \tau_{\mathcal{C}}$. It merely means that not for each such branch holds $\sigma_{\mathcal{C}} \in \tau_{\mathcal{C}}$, however, there may still exist branches for which it is true.

Definition 4. Let τ, σ be metasets and let $p \in \mathbb{T}$. We say that σ is not a member of τ under the condition p, if for each branch $\mathcal{C}$ containing p holds $\sigma_{\mathcal{C}} \notin \tau_{\mathcal{C}}$. We use the notation $\sigma \notin_p \tau$.

It might occur strange to the reader that two metasets may be in membership and non-membership relations simultaneously. The relations must be qualified by incomparable conditions, though.

Example 3. Let $\tau = \{ \langle \emptyset, 0 \rangle \}$. We check that $\emptyset \ \epsilon_0 \ \tau \ \wedge \ \emptyset \notin_1 \tau$. Indeed, if $\mathcal{C}^0$ is a branch containing 0, then $\emptyset_{\mathcal{C}^0} = \emptyset \in \{ \emptyset \} = \tau_{\mathcal{C}^0}$. Similarly, if $\mathcal{C}^1$ is a branch containing 1, then $\emptyset_{\mathcal{C}^1} = \emptyset \notin \emptyset = \tau_{\mathcal{C}^1}$. Also, $\neg \emptyset \ \epsilon_{\mathbb{1}} \ \tau \ \wedge \ \neg \emptyset \notin_{\mathbb{1}} \tau$, since it is not true, that for each branch $\mathcal{C}$ containing $\mathbb{1}$ holds $\emptyset_{\mathcal{C}} \in \tau_{\mathcal{C}}$ or $\emptyset_{\mathcal{C}} \notin \tau_{\mathcal{C}}$.

As we see, $\neg \sigma \ \epsilon_p \ \tau$ does not completely exclude the membership of σ in τ, even for $p = \mathbb{1}$. The fact that $\neg \sigma \ \epsilon_{\mathbb{1}} \ \tau$ does not contradict $\sigma \ \epsilon_p \ \tau$ for some $p \in \mathbb{T}$. It merely says that σ cannot belong to τ under the condition $\mathbb{1}$. For incomparable conditions p, q it is possible that $\sigma \notin_p \tau$ and at the same time $\sigma \ \epsilon_q \ \tau$. But it is not true that $\sigma \notin_p \tau \wedge \sigma \ \epsilon_p \ \tau$ for any p.
Analogously – by referring to interpretations – we define two sets of equality relations.

Definition 5. Let $p \in \mathbb{T}$ and let τ, σ be metasets. We say that σ is equal to τ under the condition p, if for each branch $\mathcal{C}$ containing p holds $\sigma_{\mathcal{C}} = \tau_{\mathcal{C}}$. We use the notation $\sigma \approx_p \tau$.

Definition 6. Let $p \in \mathbb{T}$ and let τ, σ be metasets. We say that σ is different than τ under the condition p, if for each branch $\mathcal{C}$ containing p holds $\sigma_{\mathcal{C}} \neq \tau_{\mathcal{C}}$. We use the notation $\sigma \not\approx_p \tau$.

Similarly as for conditional membership, it is possible that $\sigma \approx_p \tau \wedge \sigma \not\approx_q \tau$ for some metasets σ, τ and $p, q \in \mathbb{T}$.

Example 4. Let $\tau = \{ \langle \emptyset, \mathbb{1} \rangle \}$ and $\eta = \{ \langle \emptyset, 1 \rangle \}$. For a branch $\mathcal{C}$ containing 0 we have $\tau_{\mathcal{C}} = \{ \emptyset \}$ and $\eta_{\mathcal{C}} = \emptyset$. On the other hand, if $\mathcal{C}$ contains 1, then we have $\tau_{\mathcal{C}} = \{ \emptyset \} = \eta_{\mathcal{C}}$. Thus, $\tau \not\approx_0 \eta$ and $\tau \approx_1 \eta$. However, $\neg \tau \approx_{\mathbb{1}} \eta \wedge \neg \tau \not\approx_{\mathbb{1}} \eta$.

The following lemma is the metaset version of the obvious fact known from the crisp set theory: $x = y \wedge y \in z \rightarrow x \in z$.

Lemma 2. *If $p \in \mathbb{T}$ and τ, σ, λ are metasets such, that $\tau \approx_p \sigma$ and $\sigma \ \epsilon_p \ \lambda$, then also $\tau \ \epsilon_p \ \lambda$.*

Proof. If $\mathcal{C}$ is an arbitrary branch containing p, then by the assumptions $\tau_{\mathcal{C}} = \sigma_{\mathcal{C}}$ and $\sigma_{\mathcal{C}} \in \lambda_{\mathcal{C}}$. Therefore, also $\tau_{\mathcal{C}} \in \lambda_{\mathcal{C}}$, what implies $\tau \ \epsilon_p \ \lambda$. □

The certainty grades for relations on metasets are represented by sets of nodes of the binary tree and they may be evaluated as real numbers. We do not develop the general theory here, the interested reader is referred to (Starosta, 2010). Instead, we show how to evaluate the degrees of membership, non-membership, equality and difference for A-sample metasets and A-pattern metasets, when the maximal finite antichain A is fixed. Let σ, η be A-sample metasets and let τ be A-pattern metaset. The following sets contained in A

$$\mathrm{M}(\sigma, \tau) = \{ \, p \in A \colon \sigma \ \epsilon_p \ \tau \, \} \ , \tag{8}$$

$$\mathrm{N}(\sigma, \tau) = \{ \, p \in A \colon \sigma \notin_p \tau \, \} \ , \tag{9}$$

$$\mathrm{E}(\sigma, \eta) = \{ \, p \in A \colon \sigma \approx_p \eta \, \} \ , \tag{10}$$

$$\mathrm{D}(\sigma, \eta) = \{ \, p \in A \colon \sigma \not\approx_p \eta \, \} \ , \tag{11}$$

are called *membership*, *non-membership*, *equality* and *difference sets* for σ and τ (or η) respectively. The values

$$\mathrm{m}(\sigma,\tau) = \sum_{p\in \mathrm{M}(\sigma,\tau)} \frac{1}{2^{|p|}} \ , \tag{12}$$

$$\mathrm{n}(\sigma,\tau) = \sum_{p\in \mathrm{N}(\sigma,\tau)} \frac{1}{2^{|p|}} \ , \tag{13}$$

$$\mathrm{e}(\sigma,\eta) = \sum_{p\in \mathrm{E}(\sigma,\eta)} \frac{1}{2^{|p|}} \ , \tag{14}$$

$$\mathrm{d}(\sigma,\eta) = \sum_{p\in \mathrm{D}(\sigma,\eta)} \frac{1}{2^{|p|}} \ , \tag{15}$$

are called the *membership*, *non-membership*, *equality* and *difference values* of σ in τ (or η) respectively. Clearly, by the Lemma 1 all they range between 0 and 1, inclusive.
It is worth stressing, that A-sample metasets and A-pattern metasets have the following important property.

Lemma 3. *Let A be a maximal finite antichain. Let σ, η be arbitrary A-sample metasets and let τ be arbitrary A-pattern metaset. The following equations hold:*

$$\mathrm{m}(\sigma,\tau) + \mathrm{n}(\sigma,\tau) = 1 \ , \tag{16}$$

$$\mathrm{e}(\sigma,\eta) + \mathrm{d}(\sigma,\eta) = 1 \ . \tag{17}$$

Proof. First, observe that $\mathrm{M}(\sigma,\tau) \cap \mathrm{N}(\sigma,\tau) = \emptyset$ and $\mathrm{E}(\sigma,\eta) \cap \mathrm{D}(\sigma,\eta) = \emptyset$. Indeed, it is not possible that for some $p \in A$ simultaneously hold $\sigma \ \epsilon_p \ \tau$ and $\sigma \notin_p \tau$ or $\sigma \approx_p \eta$ and $\sigma \not\approx_p \eta$. Therefore, by using the Lemma 1 we may reformulate the thesis as follows:

$$\mathrm{M}(\sigma,\tau) \cup \mathrm{N}(\sigma,\tau) = A \ , \tag{18}$$

$$\mathrm{E}(\sigma,\eta) \cup \mathrm{D}(\sigma,\eta) = A \ . \tag{19}$$

To prove (18) it is enough to show, that for each $p \in A$ either $\sigma \ \epsilon_p \ \tau$ or $\sigma \notin_p \tau$ is true. In other words, either for all branches $\mathcal{C}$ containing p holds $\sigma_{\mathcal{C}} \in \tau_{\mathcal{C}}$ or for all such branches holds $\sigma_{\mathcal{C}} \notin \tau_{\mathcal{C}}$. Clearly, for any branch $\mathcal{C}$ either $\sigma_{\mathcal{C}}$ is a member of $\tau_{\mathcal{C}}$ or not, the question is whether the (non-)membership is maintained for all interpretations determined by a $p \in A$. This is true for A-sample metaset σ and A-pattern metaset τ, since $\mathrm{ran}(\sigma) \subset A$ and $\mathrm{ran}(\tau) \subset A$ and also $\bigcup_{\eta\in\mathrm{dom}(\tau)} \mathrm{ran}(\eta) \subset A$. Therefore, there exist no conditions stronger than p which could affect the interpretations. In other words, if $\mathcal{C}'$ and $\mathcal{C}''$ are different branches containing $p \in A$, then $\tau_{\mathcal{C}'} = \tau_{\mathcal{C}''}$ and $\sigma_{\mathcal{C}'} = \sigma_{\mathcal{C}''}$. The proof of (19) is analogous. □

The lemma says that there is no hesitancy in membership or equality for such metasets. This is not true for metasets in general. There exist metasets α, β with infinite ranges such, that for any $p \in \mathbb{T}$ neither $\alpha \ \epsilon_p \ \beta$ nor $\alpha \notin_p \beta$ is true, see (Starosta, 2010) for details. When we translate this property into the language of character recognition, then it says that for each pixel of a character we may decide whether it matches some pattern (or another character) or not. There is not any doubt about it.

2.2 Properties relevant to character recognition

In this section we prove some technical facts strictly relevant to character recognition mechanism. We refer to them in the sequel. Proofs are not required for understanding the idea so they may be skipped on first reading. We supply them for mathematical completeness and clarity.

The following lemma tells that for two given A-sample metasets τ and σ, their conditional difference is determined by the elements of the symmetric difference of their ranges: $\mathrm{ran}(\tau) \triangle \mathrm{ran}(\sigma)$, whereas their conditional equality is determined by the complement to A of the symmetric difference: $A \setminus (\mathrm{ran}(\tau) \triangle \mathrm{ran}(\sigma))$.

We may express this property in terms of character recognition as follows. When comparing two characters, then not only the pixels belonging to them simultaneously affect the result of comparison, but also the pixels that belong to background of both. If a pixel belongs to one of the characters and for another character the same pixel forms the background, then such pixel asserts the difference between the characters.

Lemma 4. *Let A be a finite maximal antichain in $\mathbb{T}$ and let $S, T \subset A$ be not empty. Let $\tau = \{\emptyset\} \times T$ and $\sigma = \{\emptyset\} \times S$. For $R = S \cap T \cup (A \setminus S) \cap (A \setminus T)$ the following implications hold:*

$$r \in R \;\rightarrow\; \tau \approx_r \sigma \;, \tag{20}$$

$$r \in A \setminus R \;\rightarrow\; \tau \not\approx_r \sigma \;. \tag{21}$$

Proof. Assume that $r \in S \cap T$. If $\mathcal{C}$ is a branch containing r, then clearly $\tau_{\mathcal{C}} = \{\emptyset\} = \sigma_{\mathcal{C}}$, and therefore $\tau \approx_r \sigma$. If $r \in (A \setminus S) \cap (A \setminus T)$ and $\mathcal{C}$ is a branch containing r, then $\tau_{\mathcal{C}} = \emptyset = \sigma_{\mathcal{C}}$, so $\tau \approx_r \sigma$ holds too. This proves (20).

To prove (21) note, that:

$$\begin{aligned} A \setminus R &= A \setminus \big(T \cap S \cup (A \setminus T) \cap (A \setminus S)\big) \;, \\ &= (A \setminus T \cap S) \cap \big(A \setminus (A \setminus T) \cap (A \setminus S)\big) \;, \\ &= (A \setminus T \cap S) \cap (T \cup S) \;, \\ &= \big((A \setminus T) \cup (A \setminus S)\big) \cap (T \cup S) \;, \\ &= (A \setminus T) \cap S \cup (A \setminus S) \cap T \;, \\ &= (S \setminus T) \cup (T \setminus S) \;, \\ &= S \triangle T \;. \end{aligned}$$

If $r \in (A \setminus T) \cap S$, and $\mathcal{C}$ is a branch containing r, then $\tau_{\mathcal{C}} = \emptyset$ and $\sigma_{\mathcal{C}} = \{\emptyset\}$, so $\tau \not\approx_r \sigma$. Similarly, if $r \in (A \setminus S) \cap T$, then $\tau_{\mathcal{C}} = \{\emptyset\}$ and $\sigma_{\mathcal{C}} = \emptyset$, so $\tau \not\approx_r \sigma$. Thus, for $r \in A \setminus R$ we obtain $\tau \not\approx_r \sigma$. □

The set R is the equality set for τ and σ, and $A \setminus R$ is the difference set:

$$R = \mathrm{E}(\tau, \sigma) \;, \tag{22}$$

$$A \setminus R = \mathrm{D}(\tau, \sigma) \;. \tag{23}$$

The Lemma 4 enables evaluation of the equality degree of metasets representing character samples, i.e., the similarity of two characters.

We now prove the main result which shows the construction of the membership and non-membership sets for the given A-sample metaset and A-pattern metaset. In other words,

it allows for evaluation of the similarity degree of a character testing sample (CTS) to the compound character pattern (CCP).
In the following theorem the metaset σ represents the testing sample (CTS), ρ is the compound character pattern (CCP) built up of potential elements π^i representing characters. The sets P^i and S constitute the structures of the pattern samples and the input sample. The sets Q^i represent equality degrees of the CTS and CCP elements ans the sets R^i represent the qualities of CCP members.

Theorem 5. *Let A be a maximal finite antichain in $\mathbb{T}$ and let $i = 1, \ldots, k$. Let $P^i, R^i, S \subset A$ be not empty. Let $\sigma = \{\emptyset\} \times S$, $\pi^i = \{\emptyset\} \times P^i$ and $\rho = \bigcup_{i=1}^{k} \left\{ \pi^i \right\} \times R^i$ be metasets. For the sets $Q^i = S \cap P^i \cup (A \setminus S) \cap (A \setminus P^i)$ and $U = \bigcup_{i=1}^{k} Q^i \cap R^i$, the following holds:*

$$q \in Q^i \;\rightarrow\; \sigma \approx_q \pi^i \;, \tag{24}$$

$$q \in A \setminus Q^i \;\rightarrow\; \sigma \not\approx_q \pi^i \;, \tag{25}$$

$$u \in U \;\rightarrow\; \sigma \,\epsilon_u\, \rho \;, \tag{26}$$

$$u \in A \setminus U \;\rightarrow\; \sigma \notin_u \rho \;. \tag{27}$$

Proof. The Lemma 4 proves (24) and (25).
To prove (26) take $u \in U$. There exists $i \in \{1 \ldots k\}$ such, that $u \in R^i \cap Q^i$. By (24) this implies $\sigma \approx_u \pi^i$, since $u \in Q^i$. By the construction of ρ – $u \in R^i$, so $\left\langle \pi^i, u \right\rangle \in \rho$ – and by the Definition 3 we have $\pi^i \,\epsilon_u\, \rho$. Thus, by the Lemma 2 we obtain $\sigma \,\epsilon_u\, \rho$.
To prove (27) let $R = \bigcup_{i=1}^{k} R^i$ and let $\bar{Q}^i = A \setminus Q^i$. We may split each R^i into two parts: $R^i = \left(R^i \setminus Q^i\right) \cup \left(R^i \cap Q^i\right) = R^i \cap \bar{Q}^i \cup R^i \cap Q^i$. Therefore,

$$R \;=\; \bigcup_{i=1}^{k} R^i \cap \bar{Q}^i \;\cup\; \bigcup_{i=1}^{k} R^i \cap Q^i \;=\; \bigcup_{i=1}^{k} \left(R^i \cap \bar{Q}^i\right) \;\cup\; U \;. \tag{28}$$

Let $u \in A \setminus U$ and let $\mathcal{C}$ be a branch containing u. Note, that $U \subset R \subset A$, so we consider two cases: $u \in R \setminus U$ and $u \in A \setminus R$.
If $u \in R \setminus U$, then let $I \subset \{1 \ldots k\}$ be the set of all those i, for which $u \in R^i \cap \bar{Q}^i$. Since $u \in C$ and for each $i \in I$ the intersection $R^i \cap \mathcal{C}$ contains at most one element (which is u), then by the Definition 2

$$\rho_{\mathcal{C}} = \left\{ \pi^i_{\mathcal{C}} \colon 1 \le i \le k \wedge R^i \cap \mathcal{C} \neq \emptyset \right\} = \left\{ \pi^i_{\mathcal{C}} \colon 1 \le i \le k \wedge u \in R^i \right\} = \left\{ \pi^i_{\mathcal{C}} \colon i \in I \right\} \;. \tag{29}$$

The last equality is implied by the following (since $u \notin U$):

$$\left\{ i \colon u \in R^i \right\} = \left\{ i \colon u \in R^i \cap \bar{Q}^i \right\} \cup \left\{ i \colon u \in R^i \cap Q^i \right\} = I \cup \emptyset = I \;. \tag{30}$$

Thus, for $i \in I$ we have $\pi^i_{\mathcal{C}} \in \rho_{\mathcal{C}}$. However, for $i \in I$ we also have $u \notin Q^i$, so by (25) holds $\sigma \not\approx_u \pi^i$ and consequently $\sigma_{\mathcal{C}} \neq \pi^i_{\mathcal{C}}$. Since $\sigma_{\mathcal{C}}$ is different than all the members of $\rho_{\mathcal{C}}$, then $\sigma_{\mathcal{C}} \notin \rho_{\mathcal{C}}$ for any branch $\mathcal{C} \ni u$, what gives $\sigma \notin_u \rho$. This proves (27) for the case when $u \in R \setminus U$.
If $u \in A \setminus R$, then for $\mathcal{C} \ni u$ we have $\rho_{\mathcal{C}} = \emptyset$, so $\sigma \notin_u \rho$ for any σ, what implies the second case for (27). □

The set U is the membership set for σ in ρ, and $A \setminus U$ is the non-membership set:

$$U = \mathrm{M}(\sigma, \rho) \ , \tag{31}$$

$$A \setminus U = \mathrm{N}(\sigma, \rho) \ . \tag{32}$$

The sequence of equality sets $Q^i = \mathrm{E}(\sigma, \pi^i)$ enables evaluation of equality degrees of A-sample metaset σ and potential elements π^i of A-pattern metaset ρ. They show the distribution of the overall similarity degree among the pattern elements.

3. Character recognition with metasets

In this section we explain the core of the idea of applying metasets to recognition of characters. We show how to represent characters and compound character patterns as metasets. Then we we show how to calculate appropriate membership and equality degrees and interpret them as quality grades of the input samples.

The procedure we discuss here involves two stages. During the first stage we define the compound character pattern (CCP). It represents a single character and it is comprised of a number of different samples of the character. The samples are graded with quality grades.

In the second stage we supply character testing samples (CTS) and we calculate the result which is the similarity degree of CTS to CCP. The similarity degree tells how close is the CTS to the character represented by CCP. Besides the overall similarity degree we obtain also the sequence of similarity degrees of the CTS to each member of the CCP. These degrees show how close is the input sample to each element of the compound pattern.

The compound character pattern is represented by a metaset, whose potential elements represent particular character samples of the pattern. The testing sample is represented by a metaset too. The resulting similarity degree is the membership degree of CTS in CCP. The additional similarity degrees of CTS to pattern elements are partial equality degrees of CTS to potential elements of CCP.

One of the goals of this section is to convince the reader that partial membership and equality degrees of metasets encoding character samples properly reflect the human perception of similarity of characters.

3.1 Representing characters as metasets

Characters are displayed on the matrix X_r^c comprised of r rows and c columns (shortly: X). The natural numbers r and c may be arbitrary, however they must remain constant throughout the matching process: all the character samples in the CCP pattern as well as all the CTS input samples must use the same matrix dimensions. We focus on monochromatic images here, so the cells of the matrix acquire two states: selected ones belong to the character and deselected ones form the background. For the given character a displayed on the matrix, the set of selected cells is denoted by Xa.

Prior to defining character samples, a mapping $m\colon X \mapsto \mathbb{T}$ between matrix cells and nodes of the binary tree must be established. To each cell of the matrix a node of the binary tree must be assigned so that the set of assigned nodes – denoted by $m(X)$ – forms a maximal antichain A in the tree $\mathbb{T}$. The assignment of nodes to cells is arbitrary – no special ordering is required. The antichain and the mapping are constant for the whole character matching process – all the CTS and CCP samples use the same A and m. Note, that since the nodes assigned to cells form a MFA, then any branch in the tree contains exactly one assigned node.

The simplest example of such assignment is when $r \cdot c = 2^k$ for some k. In such case the nodes of the k-th level of the binary tree may be assigned in an arbitrary way to the cells. We call such one-to-one mapping of matrix and some level in $\mathbb{T}$ an *even* mapping. The Figure 2 demonstrates a sample 4×4 matrix with a mapping $m\colon X_4^4 \mapsto \mathbb{T}_4$ onto the level 4 of the tree. For simplicity, most examples will be based on this 4×4 matrix and the mapping.

0000	0001	0010	0011
0100	0101	0110	0111
1000	1001	1010	1011
1100	1101	1110	1111

Fig. 2. A standard mapping of the level 4 of the binary tree $\mathbb{T}$ to cells of the 4×4 matrix.

When $r \cdot c \neq 2^k$ for any $k \in \mathbb{N}$, then the cells of the matrix must be mapped to nodes from different levels of $\mathbb{T}$, since levels contain 2^k nodes. Anyway, the image $m(X_r^c)$ must be a MFA. We call such a mapping *uneven*. See Fig. 3 for an example of uneven 3×4 mapping.

11100	000	001	11110
1100	010	011	1101
11101	100	101	11111

Fig. 3. Mapping of some antichain in $\mathbb{T}$ to cells of the 3×4 matrix.

For an even mapping the placement of particular nodes is rather irrelevant. On the other hand, when the mapping is uneven, then the nodes from different levels assigned to cells impose the following interpretation. Parts of the matrix which are more important for the particular character, and which we want to stress somehow by distinguishing it from the rest, are associated with nodes which are closer to the root – the weaker conditions. The cells which are of less importance contain nodes from lower levels of the tree – the stronger conditions. Weaker conditions have more impact on the resulting membership and equality degrees than stronger ones (cf. Equations 12–15). For instance we might be particularly interested in proper recognizing of the dot over the letter 'i'. In such case we may use the assignment depicted on the Fig. 4. The reader is encouraged to check that the nodes form a maximal antichain. The cells containing the nodes 10 and 110 are more sensitive to errors than other cells and they influence the resulting similarity degree more than others.

0000	10	0100
0001	110	0101
0010	1110	0110
0011	1111	0111

Fig. 4. Simple assignment for stressing the dot over 'i'.

Note, that even when $r \cdot c = 2^k$ for some k, then the mapping might be uneven too, since we may assign nodes from different levels to cells in order to stress some areas of the matrix and diminish the influence of others. Anyway, the requirement that the range forms a maximal antichain must be fulfilled. The assignment on the Fig. 5 shows how to stress the upper-left corner of the X_2^2 matrix. The impact of the lower row is much less than the impact of the upper row in this case.

0	10
110	111

Fig. 5. Uneven assignment for 2×2 matrix.

We now construct the metaset χ representing the character denoted by a displayed on the matrix X. The domain of the metaset consists of the empty set only: $\mathrm{dom}(\chi) = \{\emptyset\}$. The set $m(Xa) \subset A$ of nodes corresponding to the marked cells of the matrix forms the range of the metaset representing the sample: $\mathrm{ran}(\chi) = m(Xa)$. Since the domain of χ contains exactly one element $\emptyset$, then $\mathrm{ran}(\chi) = \chi[\emptyset]$. Thus,

$$\chi = \{\emptyset\} \times m(Xa) \ . \tag{33}$$

Note, that we interpret the membership degree of $\emptyset$ in χ as the set of selected cells of the character. This membership degree is irrelevant by itself, however, it determines the equality degree of this sample and any other CTS supplied during the recognition phase. It also affects the overall result which is the membership degree of the CTS in the CCP.
As an example, let us represent the character 'c' on the 4×4 matrix with the standard assignment, like on the Fig. 6. The metaset representing this letter is

$$\begin{aligned} \chi = \{ \ & \langle \emptyset, 0001 \rangle, \langle \emptyset, 0010 \rangle, \langle \emptyset, 0011 \rangle, \langle \emptyset, 0100 \rangle, \\ & \langle \emptyset, 1000 \rangle, \langle \emptyset, 1101 \rangle, \langle \emptyset, 1110 \rangle, \langle \emptyset, 1111 \rangle \ \} \ . \end{aligned} \tag{34}$$

The set of nodes corresponding to the selected cells is

$$m(X\mathtt{c}) = \{\, 0001, 0010, 0011, 0100, 1000, 1101, 1110, 1111 \,\} \ . \tag{35}$$

	0001	0010	0011
0100			
1000			
	1101	1110	1111

Fig. 6. The character 'c' represented on the 4×4 matrix.

3.2 Defining the compound pattern

Defining the compound character pattern (CCP) is the essential step in the process of character recognition with metasets. The CCP consists of a number of character samples accompanied by quality grades. The samples describe some point of view on the character. The different

shapes collected together give an idea of how the character should look like. They may be supplied by independent experts or they may be samples of hand-writing retrieved from different persons.

The most important factor here is defining the quality grades for samples included in the CCP as sets of nodes of the binary tree. Instead of giving them numerical values – as it is usually done – initially we define quality grades to be the parts of the character matrix which contain valid data, for each sample separately. Thus, the quality grade of a character sample is the set of cells containing correct, necessary pixels of the character or its background. For each sample, the cells of the matrix which are considered bad, missing or not important are excluded by the quality grade area and therefore, they are not taken into account during the recognition process. The mapping m transforms this quality set into a subset of the maximal finite antichain A, which may be evaluated as a number then, however some part of information is lost this way (i.e., which exactly cells are taken into account and which are considered invalid).

Associating character samples represented as A-sample metasets with the corresponding quality grades represented as subsets of A we create the A-pattern metaset representing the CCP. Then, testing character samples represented by other A-sample metasets are matched against the A-pattern metaset.

If we denote characters included in the CCP with the variables $c_1, c_2, \ldots$, then $Xc_1, Xc_2, \ldots$ are the sets of cells of the matrix which contain their pixels – the selected cells. The metasets corresponding to the characters are denoted with $\chi_1, \chi_2, \ldots$ (cf. Equation 33):

$$\chi_i = \{\, \emptyset \,\} \times m(Xc_i) \ . \tag{36}$$

The corresponding quality grades $q_1, q_2, \ldots$, when expressed as sets of cells of the matrix X are denoted with $Xq_1, Xq_2, \ldots$. Thus, $m(Xq_1), m(Xq_2), \ldots$ are subsets of A specifying quality grades of characters $c_1, c_2, \ldots$, or – in other words – they are membership degrees of the A-sample metasets χ_i in the A-pattern metaset π representing the CCP, which is defined as follows (n is the number of samples in the pattern):

$$\pi = \bigcup_{i=1}^{i=n} \{\, \chi_i \,\} \times m(Xq_i) \ . \tag{37}$$

The complete structure of the A-pattern metaset π representing the compound character pattern comprised of the characters $c_1, \ldots, c_n$ accompanied by the quality grades $q_1, \ldots, q_n$ is depicted by the following equation

$$\pi = \bigcup_{i=1}^{i=n} \{\, \{\, \emptyset \,\} \times m(Xc_i) \,\} \times m(Xq_i). \tag{38}$$

We illustrate the above formulas with an example. We use the X_4^4 matrix with the standard mapping m onto the antichain $A = \mathbb{T}_4$ which is the 4th level of the tree (cf. Fig.2). The Figure 7 depicts three different samples of the letter 'c'. Pixels of characters are those containing binary sequences. Invalid cells are marked gray; the cells without background form the quality grades.

We understand that the areas of the matrix with gray background contain pixels which are either unreadable or we are not sure whether they are selected or not, or they are distorted

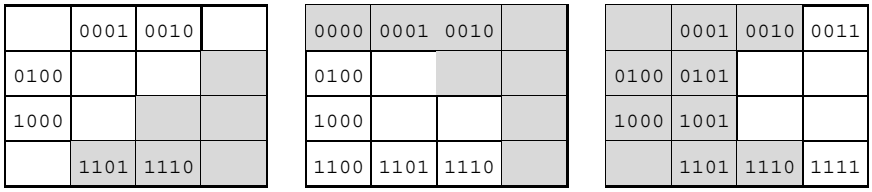

Fig. 7. Three samples c_1, c_2, c_3 of letter 'c'. Cells without gray background make up quality grades.

somehow, and therefore they cannot be included in the representation of the character without causing any doubt. They may also be treated as a mask for excluding parts of the matrix from the matching process. Anyway, when calculating equality degrees the whole matrix area is taken into account, so excluded parts play role in determining the membership (similarity) degree to the CCP only.
The sets of nodes corresponding to the selected cells of the characters c_1, c_2 and c_3 on the Fig. 7 are shown by the following equations:

$$m(Xc_1) = \{\, 0001, 0010, 0100, 1000, 1101, 1110 \,\} \ , \tag{39}$$
$$m(Xc_2) = \{\, 0000, 0001, 0010, 0100, 1000, 1100, 1101, 1110 \,\} \ , \tag{40}$$
$$m(Xc_3) = \{\, 0001, 0010, 0011, 0100, 0101, 1000, 1001, 1101, 1110, 1111 \,\} \ . \tag{41}$$

The A-sample metasets χ_1, χ_2, χ_3 representing these characters have form (cf. Equation 36):

$$\chi_1 = \{\ \langle \emptyset, 0001 \rangle, \langle \emptyset, 0010 \rangle, \langle \emptyset, 0100 \rangle, \langle \emptyset, 1000 \rangle, \langle \emptyset, 1101 \rangle, \langle \emptyset, 1110 \rangle \ \} \ , \tag{42}$$
$$\chi_2 = \{\ \langle \emptyset, 0000 \rangle, \langle \emptyset, 0001 \rangle, \langle \emptyset, 0010 \rangle, \langle \emptyset, 0100 \rangle, \langle \emptyset, 1000 \rangle, \langle \emptyset, 1100 \rangle, \langle \emptyset, 1101 \rangle, \langle \emptyset, 1110 \rangle \ \} \ , \tag{43}$$
$$\chi_3 = \{\ \langle \emptyset, 0001 \rangle, \langle \emptyset, 0010 \rangle, \langle \emptyset, 0011 \rangle, \langle \emptyset, 0100 \rangle, \langle \emptyset, 0101 \rangle, \langle \emptyset, 1000 \rangle, \langle \emptyset, 1101 \rangle, \langle \emptyset, 1101 \rangle, \langle \emptyset, 1110 \rangle, \langle \emptyset, 1111 \rangle \ \} \ . \tag{44}$$

They comprise the domain of the A-pattern metaset π: $\mathrm{dom}(\pi) = \{\, \chi_1, \chi_2, \chi_3 \,\}$. The quality grades q_i of the samples c_i – represented by the cells without gray background – when mapped to subsets of A with the mapping m, make up the membership degrees $m(Xq_i)$ of χ_i in the π:

$$\pi[\chi_i] = m(Xq_i), \text{ for } i = 1, 2, 3 \ . \tag{45}$$

From the Fig. 7 we may read that

$$m(Xq_1) = \{\, 0000, 0001, 0010, 0011, 0100, 0101, 0110, 1000, 1001, 1100 \,\} = \mathbb{T}_4 \setminus \{\, 0111, 1010, 1011, 1101, 1110, 1111 \,\} \ , \tag{46}$$
$$m(Xq_2) = \{\, 0100, 0101, 1000, 1001, 1010, 1100, 1101, 1110 \,\} = \mathbb{T}_4 \setminus \{\, 0000, 0001, 0010, 0011, 0110, 0111, 1011, 1111 \,\} \ , \tag{47}$$
$$m(Xq_3) = \{\, 0011, 0110, 0111, 1010, 1011, 1111 \,\} = \mathbb{T}_4 \setminus \{\, 0000, 0001, 0010, 0100, 0101, 1000, 1001, 1100, 1101, 1110 \,\} \ . \tag{48}$$

Thus, the A-pattern metaset π representing the CCP has the following complete structure:

$$\begin{aligned}\pi \;&=\; \{\,\chi_1\,\} \times m(Xq_1) \;\cup\; \{\,\chi_2\,\} \times m(Xq_2) \;\cup\; \{\,\chi_3\,\} \times m(Xq_3) \\ &= \{\{\,\emptyset\,\} \times \{\,0001, 0010, 0100, 1000, 1101, 1110\,\}\} \\ &\quad \times \{\,0000, 0001, 0010, 0011, 0100, 0101, 0110, 1000, 1001, 1100\,\} \\ &\cup \{\{\,\emptyset\,\} \times \{\,0000, 0001, 0010, 0100, 1000, 1100, 1101, 1110\,\}\} \\ &\quad \times \{\,0100, 0101, 1000, 1001, 1010, 1100, 1101, 1110\,\} \\ &\cup \{\{\,\emptyset\,\} \times \{\,0001, 0010, 0011, 0100, 0101, 1000, 1001, 1101, 1110, 1111\,\}\} \\ &\quad \times \{\,0011, 0110, 0111, 1010, 1011, 1111\,\} \quad .\end{aligned} \tag{49}$$

For each $\chi_i \in \mathrm{dom}(\pi)$ the numerical values of the membership degrees of χ_i in π, equal to their numerical quality grades, are given by the formula (cf. Equation 12):

$$\mathrm{m}(\chi_i, \pi) = \sum_{p \in \pi[\chi_i]} \tfrac{1}{2^{|p|}} \;, \tag{50}$$

and they are equal 0.62, 0.5 and 0.38, for the characters c_1, c_2, c_3 respectively (the reader is encouraged to verify it). The numerical representation of the degree looses information concerning the particular cells taken into account. For instance, there are many combinations of cells for which the above formula gives the result of 0.5. The numerical value is more human-friendly, however.

3.3 Evaluating similarity degrees

Once the compound character pattern (CCP) is prepared we are ready to supply testing character samples (CTS) and evaluate their similarity degrees. The CTS is represented as a metaset in exactly the same manner as CCP elements, i.e., we use the same matrix X with the same mapping m of cells to some maximal finite antichain A.

The process of matching the input character sample represented by the A-sample metaset τ against the prepared compound character pattern represented by the A-pattern metaset π involves calculation of the membership degree of τ in π and the sequence of equality degrees of τ and potential elements χ_i of π. The membership degree tells us to what measure the CTS resembles the character defined by the CCP and is represented by the set $\mathrm{M}(\tau, \pi)$ (see Equation 8). The equality degrees play supplemental role and they show the similarity of τ and each pattern element separately, which – contrary to the CCP – are single characters. They are represented by the sets $\mathrm{E}(\tau, \chi_i)$ (see Equation 10). We apply here the Theorem 5 for determining the similarity degrees and also the Equations 12–15 for numerical evaluation of the degres.

Let us make calculations for the sample letter 'c' shown on the Fig. 6. The metaset χ representing the character is defined by the Equation (34). First, we establish the notation. The left hand sides of the following equations correspond to variables used in the Theorem 5 and in the right hand sides we use metasets defined in previous sections, $i = 1, 2, 3$.

$$\sigma = \chi \qquad \text{the CTS, see Fig. 6 and Equation 34} \;, \tag{51}$$

$$\pi^i = \chi_i \qquad \text{the pattern elements, see Equations 42–44} \;, \tag{52}$$

$$\rho = \pi \qquad \text{the CCP, see Fig. 7 and Equation 49} \;, \tag{53}$$

$$S = m(X\mathrm{c}) \qquad \text{the CTS selected cells, see Equation 35} \;, \tag{54}$$

$$P^i = m(Xc_i) \quad \text{the selected cells of CCP elements, see Fig. 7 and Equations 39–41 ,} \tag{55}$$

$$R^i = m(Xq_i) \quad \text{the CCP quality marks, see Equations 46–48 ,} \tag{56}$$

$$Q^i = \mathrm{E}(\chi, \chi_i) \quad \text{the equality sets, see Equations 10 and 22–23 ,} \tag{57}$$

$$U = \mathrm{M}(\chi, \pi) \quad \text{the membership set, see Equations 8 and 31–32 .} \tag{58}$$

We start with calculating the sets Q^i, Recall, that the antichain A is equal to the level $\mathbb{T}_4$ of the tree and the mapping m is shown of the Fig. 2.

$$\begin{aligned}
\mathrm{E}(\chi_1, \chi) = Q^1 &= S \cap P^1 \cup (A \setminus S) \cap (A \setminus P^1) \\
&= \{\, 0001, 0010, 0011, 0100, 1000, 1101, 1110, 1111 \,\} \\
&\quad \cap \{\, 0001, 0010, 0100, 1000, 1101, 1110 \,\} \\
&\quad \cup \{\, 0000, 0101, 0110, 0111, 1001, 1010, 1011, 1100 \,\} \\
&\quad \cap \{\, 0000, 0011, 0101, 0110, 0111, 1001, 1010, 1011, 1100, 1111 \,\} \\
&= P^1 \cup (A \setminus S) \\
&= \{\, 0001, 0010, 0100, 1000, 1101, 1110 \,\} \\
&\quad \cup \{\, 0000, 0101, 0110, 0111, 1001, 1010, 1011 \,\} \\
&= \mathbb{T}_4 \setminus \{\, 0011, 1111 \,\} \quad .
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}(\chi_2, \chi) = Q^2 &= S \cap P^2 \cup (A \setminus S) \cap (A \setminus P^2) \\
&= \{\, 0001, 0010, 0011, 0100, 1000, 1101, 1110, 1111 \,\} \\
&\quad \cap \{\, 0000, 0001, 0010, 0100, 1000, 1100, 1101, 1110 \,\} \\
&\quad \cup \{\, 0000, 0101, 0110, 0111, 1001, 1010, 1011, 1100 \,\} \\
&\quad \cap \{\, 0011, 0101, 0110, 0111, 1001, 1010, 1011, 1111 \,\} \\
&= P^2 \setminus \{\, 0000, 1100 \,\} \ \cup \ (A \setminus S) \setminus \{\, 0000, 1100 \,\} \\
&= \{\, 0001, 0010, 0100, 1000, 1101, 1110 \,\} \\
&\quad \cup \{\, 0101, 0110, 0111, 1001, 1010, 1011 \,\} \\
&= \mathbb{T}_4 \setminus \{\, 0000, 0011, 1100, 1111 \,\} \quad .
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}(\chi_3, \chi) = Q^3 &= S \cap P^3 \cup (A \setminus S) \cap (A \setminus P^3) \\
&= \{\, 0001, 0010, 0011, 0100, 1000, 1101, 1110, 1111 \,\} \\
&\quad \cap \{\, 0001, 0010, 0011, 0100, 0101, 1000, 1001, 1101, 1110, 1111 \,\} \\
&\quad \cup \{\, 0000, 0101, 0110, 0111, 1001, 1010, 1011, 1100 \,\} \\
&\quad \cap \{\, 0000, 0110, 0111, 1010, 1011, 1100 \,\} \\
&= S \ \cup \ (A \setminus P^3) \\
&= \{\, 0001, 0010, 0011, 0100, 1000, 1101, 1110, 1111 \,\} \\
&\quad \cup \{\, 0000, 0110, 0111, 1010, 1011, 1100 \,\} \\
&= \mathbb{T}_4 \setminus \{\, 0101, 1001 \,\} \quad .
\end{aligned}$$

From the equation 14 we obtain the numerical values of the equality degrees.

$$\mathrm{e}(\chi_1,\chi) = 1 - \frac{2}{2^4} = \frac{7}{8}$$

$$\mathrm{e}(\chi_2,\chi) = 1 - \frac{4}{2^4} = \frac{6}{8}$$

$$\mathrm{e}(\chi_3,\chi) = 1 - \frac{2}{2^4} = \frac{7}{8} \ .$$

The results show, that the character on the Fig. 6 resembles the characters c_1 and c_3 on the Fig. 7 equally well, whereas the character c_2 a bit worse. Note, that we do not take into account the qualities q_i of the samples c_i when calculating the equality degrees.
Now we calculate the membership set $\mathrm{M}(\chi,\pi)$ (U in terms of Theorem 5) and the membership value $\mathrm{m}(\chi,\pi)$ of χ in π. We apply the equations 46–48.

$$\begin{aligned}
\mathrm{M}(\chi,\pi) = U &= Q^1 \cap R^1 \ \cup \ Q^2 \cap R^2 \ \cup \ Q^3 \cap R^3 \\
&= Q^1 \cap m(Xq_1) \ \cup \ Q^2 \cap m(Xq_2) \ \cup \ Q^3 \cap m(Xq_3) \\
&= \ (\mathbb{T}_4 \setminus \{\, 0011, 1111 \,\}) \\
&\quad \cap (\mathbb{T}_4 \setminus \{\, 0111, 1010, 1011, 1101, 1110, 1111 \,\}) \\
&\quad \cup (\mathbb{T}_4 \setminus \{\, 0000, 0011, 1100, 1111 \,\}) \\
&\quad \cap (\mathbb{T}_4 \setminus \{\, 0000, 0001, 0010, 0011, 0110, 0111, 1011, 1111 \,\}) \\
&\quad \cup (\mathbb{T}_4 \setminus \{\, 0101, 1001 \,\}) \\
&\quad \cap (\mathbb{T}_4 \setminus \{\, 0000, 0001, 0010, 0100, 0101, 1000, 1001, 1100, 1101, 1110 \,\}) \\
&= m(Xq_1) \setminus \{\, 0011 \,\} \ \cup \ m(Xq_2) \setminus \{\, 1100 \,\} \ \cup \ m(Xq_3) \\
&= \mathbb{T}_4 \ .
\end{aligned}$$

Clearly, by the Equation 12 and by the Lemma 1,

$$\mathrm{m}(\chi,\pi) = 1 \ . \tag{59}$$

This means, that the sample χ perfectly matches the pattern π.

3.4 Discussion of the results

The interesting question that arises is what are the similarity degrees of each pattern element χ_i to the pattern π itself? It turns out that the pattern samples do not have to be of the best quality in order to assure that other input samples result in perfect matches.
We show that the membership sets $\mathrm{M}(\chi_i,\pi)$ are proper subsets of $\mathbb{T}_4$ and therefore, the membership values $\mathrm{m}(\chi_i,\pi)$ are less than 1 for all $i = 1,2,3$. We present the results of calculations only, leaving the details to the reader. Let us start with equality sets:

$$\mathrm{E}(\chi_i,\chi_j) = \mathbb{T}_4 \setminus \mathrm{D}(\chi_i,\chi_j) \ , \tag{60}$$

where $\mathrm{D}(\chi_i,\chi_j)$ are the difference sets depicted on the Table 1.

$\mathrm{D}(\chi_i, \chi_j)$	χ_1	χ_2	χ_3
χ_1	$\emptyset$	0000, 1100	0011, 0101, 1001, 1111
χ_2	0000, 1100	$\emptyset$	0000, 0011, 0101, 1001, 1100, 1111
χ_3	0011, 0101, 1001, 1111	0000, 0011, 0101, 1001, 1100, 1111	$\emptyset$

Table 1. Difference sets $\mathrm{D}(\chi_i, \chi_j)$ for compound pattern elements

The matrix on the Table 1 is symmetric since $\chi_i \approx_p \chi_j$ is equivalent to $\chi_j \approx_p \chi_i$. Empty sets on the diagonal confirm, that $\chi_i \approx_p \chi_i$, for each $p \in \mathbb{T}_4$. We conclude that the equality values are as depicted on the Table 2.

$\mathrm{e}(\chi_i, \chi_j)$	χ_1	χ_2	χ_3
χ_1	1	0.88	0.75
χ_2	0.88	1	0.62
χ_3	0.75	0.62	1

Table 2. Equality values $\mathrm{e}(\chi_i, \chi_j)$ for compound pattern elements

Based on the above sets we calculate the membership sets and the membership values, similarly as before.

$$\mathrm{M}(\chi_1, \pi) = \mathbb{T}_4 \setminus \{\, 1111 \,\} \qquad \mathrm{m}(\chi_1, \pi) = 0.94 \ ,$$

$$\mathrm{M}(\chi_2, \pi) = \mathbb{T}_4 \setminus \{\, 0000, 1111 \,\} \qquad \mathrm{m}(\chi_2, \pi) = 0.88 \ ,$$

$$\mathrm{M}(\chi_3, \pi) = \mathbb{T}_4 \setminus \{\, 0101, 1001 \,\} \qquad \mathrm{m}(\chi_3, \pi) = 0.88 \ .$$

Thus, the similarity values of the characters c_1, c_2 and c_3 to the CCP built on top of them are 0.94, 0.88 and 0.88, respectively. None of them matches the pattern to the highest degree. Even though the membership values of χ_i in π are less than 1, there exist samples which match the CCP to the highest possible degree, with the membership value equal 1. Besides the character on the Fig. 6, the are three more – shown on the Fig.8 – for which the similarity degree reaches the maximal value.

	0001	0010	0011
0100			
1000			
1100	1101	1110	1111

	0001	0010	
0100			
1000			
1100	1101	1110	1111

	0001	0010	
0100			
1000			
	1101	1110	1111

Fig. 8. Three remaining samples with the best similarity to the pattern represented by χ.

Note, that the samples on the Fig. 6 and Fig. 8 differ only in pixels 0011 and 1100, which in at least one of the CCP elements on Fig. 7 belong to the sample and in at least another one belong to the background – being not excluded by the quality area at the same time.
The character samples c_i and their quality grades q_i were intentionally chosen so that they do not match the CCP to the highest degree, in order to demonstrate interpolation capabilities of the new mechanism. In typical cases, one constructs the CCP based on the good samples, which reflect most characteristics of the modelled pattern.

When creating a CCP one should bear in mind the following rule. Each pixel of the matrix must be covered by the foreground or background of at least one sample in the pattern. By covering we understand that it is included in at least one quality area. The reader may confirm that this rule is preserved in our example. If there exist a cell which is contained in exclusion area of each sample, then reaching the similarity value of 1 is not possible for any sample.

4. Conclusions

We demonstrated the method for character recognition based on metasets. The core of the idea lies in representing character samples and character patterns directly as metasets, as well as interpreting the membership and the equality degrees of corresponding metasets as the similarity degrees of characters. Although the idea is quite simple and straightforward, it seems to work fine. The experiments carried out with the computer application[1] implementing this model confirm that it adequately reflects human perception of similarity of characters. As we have seen, the mechanism requires some laborious calculations, however they are to be carried out by machines.

So far, no comparisons with other techniques for character recognition have been made. It must be stressed that the presented method is not by itself competitive to commercial solutions yet. It is rather a sketch of an idea which – when applied in cooperation with other techniques used for data processing, like centering and sharpening of character images – may turn out to have some advantages over other solutions.

The main goal of this chapter was to convince the reader, that the idea of metaset is applicable to solving problems related to processing of vague, imprecise data. And moreover, that modelling of real world using metasets is quite natural and simple. We showed that metaset membership correctly mimics similarity when characters are appropriately encoded.

It should be clear, that the discussed method has much wider scope of applications than recognition of letters. Although we presented the version for monochromatic (binary) images, it is not difficult to generalize it to color (many-valued) ones. The next step in research on the subject will focus on determining the characteristics of graphical data for which this method gives the best results.

5. References

Atanassov, K. T. (1986). Intuitionistic fuzzy sets, *Fuzzy Sets and Systems* 20: 87–96.

Kunen, K. (1980). *Set Theory, An Introduction to Independence Proofs*, number 102 in *Studies in Logic and Foundations of Mathematics*, North-Holland Publishing Company.

Pawlak, Z. (1982). Rough sets, *International Journal of Computer and Information Sciences* 11: 341–356.

Starosta, B. (2009). Application of metasets to character recognition, *Proc. of 18th International Symposium, ISMIS 2009*, Vol. 5722 of *Lecture Notes in Artificial Intelligence*, pp. 602–611.

Starosta, B. (2010). Representing intuitionistic fuzzy sets as metasets, *Developments in Fuzzy Sets, Intuitionistic Fuzzy Sets, Generalized Nets and Related Topics. Volume I: Foundations*, pp. 185–208.

Starosta, B. & Kosiński, W. (2009). *Views on Fuzzy Sets and Systems from Different Perspectives. Philosophy and Logic, Criticisms and Applications*, Vol. 243 of *Studies in Fuzziness and*

[1] The application is available as Java applet under the URL:
`http://www.pjwstk.edu.pl/~barstar/Research/MSOCR/index.html`

Soft Computing, Springer Verlag, chapter Meta Sets. Another Approach to Fuzziness, pp. 509–522.
Zadeh, L. A. (1965). Fuzzy sets, *Information and Control* 8: 338–353.

Recent Advances in Document Recognition and Understanding

Edited by Dr. Minoru Mori

ISBN 978-953-307-320-0
Hard cover, 94 pages
Publisher InTech
Published online 17, October, 2011
Published in print edition October, 2011

In the field of document recognition and understanding, whereas scanned paper documents were previously the only recognition target, various new media such as camera-captured documents, videos, and natural scene images have recently started to attract attention because of the growth of the Internet/WWW and the rapid adoption of low-priced digital cameras/videos. The keys to the breakthrough include character detection from complex backgrounds, discrimination of characters from non-characters, modern or ancient unique font recognition, fast retrieval technique from large-scaled scanned documents, multi-lingual OCR, and unconstrained handwriting recognition. This book aims to present recent advances, applications, and new ideas that are relevant to document recognition and understanding, from technical topics such as image processing, feature extraction or classification, to new applications like camera-based recognition or character-based natural scene analysis. The goal of this book is to provide a new trend and a reference source for academic research and for professionals working in the document recognition and understanding field

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821