A property concept frame representation for flexible image_content retrieval in histopathology databases.

Marie-Christine Jaulent PhD., Christel Le Bozec MD., Yunyu Cao MS., Eric Zapletal and Patrice Degoulet MD. PhD.

SPIM, Faculté de Médecine, 115 rue de l'école de Médecine, 75005 Paris, France
jaulent@hegp.bhdc.jussieu.fr

Abstract

In histopathology databases, images descriptions are collections of properties provided by experts. Image content retrieval implies comparison of such properties. The objective of this work is to enrich the traditional attribute-value representation of properties in order to take into account the polymorphism and subjectivity of properties and to manage the comparison process. In this paper we define a property concept frame (PCF) representation based on fuzzy logic to handle both representation and comparison. Seven quantifiable morphological characteristics were selected from histopathological reports to illustrate the variety of fuzzy predicates and linguistic terms in properties. The PCF representation has been tested in the context of breast pathology. It is concluded that the PCF representation provides a unification scheme to retrieve in images morphological characteristics that are described in different ways. It may enhance the relevancy of applications in various contexts such as image content-based retrieval or case-based reasoning from images.

INTRODUCTION

In histopathology, the diagnostic making process relies on the subjective analysis of images and expertise that comes over time with the examination of a multitude of cases. Images databases are built in pathology to store this expertise [1][2] and retrieval tools are needed to reach relevant information inside these databases. Classically, information retrieval is based on indexation mechanisms. The indexation is often limited to information about images (medical data or diagnosis) rather than information contained within the images. Many researches attempt to carry image content based retrieval out through a matching process between relevant image descriptions and a query representation [3].

At the image representation level several works have proposed generic models to implement medical image databases [4]. Three layers are usually distinguished, the numerical level of pixels (the image), the symbolic level of primitives with their parameters (the features) and the highest level of semantic descriptions (the properties). Properties can either be obtained from features (e.g. the rounded aspect of a cell can be derived from surface and connectivity features) or from experts. In a previous work, we have proposed a coding framework for the representation of histopathological images at the semantic level in terms of a collection of properties that described morphological characteristics [5]. Experts provide descriptions in a standardized vocabulary and couples (Attribute Value) represent properties. Because of the human involvement and the intrinsic subjectivity of the specialty, uncertainty is manifest in the definitions and specifications of many properties [6]. We argue that fuzzy logic can be used to manage uncertainty and imprecision in the representation and processing of the subjective information provided by experts' [7].

At the matching level, classical or fuzzy approaches uses similarity matrix or distance calculus to compare properties and are often based on the paradigm that properties are of the same nature, that is represented in the same way (numbers, labels, fuzzy sets, etc) [8][9].

The objective of our work is to compare properties of different natures in order to perform flexible retrieval in images databases. In a recent paper, Dubitsky has defined a property concept frame (PCF) to represent polymorphous properties (numerical value, symbolic value, and fuzzy predicate) in the same data structure [10]. We propose in this paper an extension of the Dubitsky approach where fuzzy predicates are automatically constructed from linguistic labels and fuzzy quantifiers. The comparison between properties is based on the possibility theory.

The paper presents first the background on the IDEM environment and the different natures of properties to take into account. The property concept frame is then described at the level of property representation and at the level of property comparison. An implementation in Java with the ObjectStore Object-Oriented database has been done and applied in the specific context of seven usual attributes in the domain of histopathology concerning the description of cells, lobular gland, etc. Some examples of the use of the procedure are presented and the results are discussed.

BACKGROUND: the IDEM environment

IDEM is an integrated computerized environment dedicated to pathologists. It includes a case base in the domain of breast pathology. A case is composed of images, different examination's parameters, a structured description of the morphological characteristics present in images and relevant for the diagnosis (the properties),

1067—5027/00/$5.00 © 2000 AMIA, Inc.

some diagnosis information and the final textual report. Services that are integrated in IDEM include: 1) a case-based reasoning mechanism providing an "intelligent" retrieval of reference images and diagnostic clues [11], 2) a description tool to acquire from experts the description of a case and 3) a computerized consensus building tool to get consensual structured descriptions [5].

Structured case description in IDEM.

A case is considered as a collection of macroscopic areas and histological areas. A histological area can contain several histological areas as well as a cytological description. Areas are described by sets of properties. For instance, the following natural language expert description "a cell with granular eosinophil cytoplasm and small rounded nucleus" is embedded, using the description tool, in two histological areas each described by two properties (figure 1).

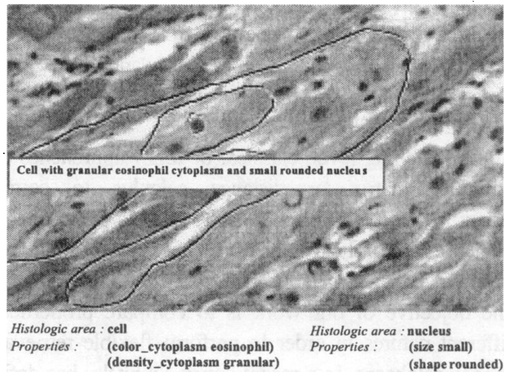

Cell with granular eosinophil cytoplasm and small rounded nucleus
Histologic area : cell
Properties : (color_cytoplasm eosinophil)
(density_cytoplasm granular)
Histologic area : nucleus
Properties : (size small)
(shape rounded)

Figure 1: Case description through properties

The nature of properties in IDEM.

The domain vocabulary is very rich, not well standardized and properties are polymorphous by nature. For instance, the property "two micro-calcification seats" is a numeric property, the property "cells with abundant cytoplasm" is a symbolic simple property and the property "numerous lobular glands of various sizes" is a symbolic complex property called further a fuzzy predicate. All these examples are extracted from daily reports and show that uncertainty and polymorphism is pervasive in the properties used to describe cases. In the context of this work, quantifiable subjective properties are considered. A quantifiable property is a property that can have a numerical value. Three types of properties are distinguished:

- Real numbers (*RN*). They are precise real values on $\Re$.
- Linguistic terms (*LT*). They are symbolic words that are part of a standardized universe of discourse. They are naturally imprecise.
- Fuzzy predicates (*FP*). They are symbolic expressions that are not part of a standardized universe of discourse. They are naturally imprecise. They are restricted to combination of *LT* or *RN* values on the same universe of discourse (e.g. "rather big", "around 2cm").

The comparison of properties in IDEM.

In the IDEM environment, the comparison of properties is based on similarity matrix. In the context of non-quantifiable properties, experts explicitly provide the different matrix. Concerning quantifiable properties, the matrix is partly automatically built up with the constraint that values are only linguistic terms. It is then necessary to extend the comparison of properties in IDEM to take into account the polymorphism.

METHODS

The representation and processing (similarity assessment) of properties being polymorphous instance values of attributes is based on a property concept frame (PCF). The property concept frame approach provides a representation platform to model the relationships between the various value formats, thus enabling the computation of cross-format similarity scores [10]. It serves as unifying representation formalism for three property value formats, *RN*, *LT* and *FP*.

In this section, we first present the modeling of the different value formats in the context of the fuzzy set theory and then describe the flexible comparison of the properties in the context of the possibility theory.

Multiple property representation

A property is an attribute/value couple. The « attribute » refers to a specific morphological dimension like, for instance, the size of a cell or its mitotic activity. The "value" is the instantiation of this attribute in a particular case. The representation of a property concerns the representation of all the possible values. The representation of a *RN* value is straightforward being just the exact value. The representation of a *LT* value is based on the fuzzy set theory. The fuzzy set theory allows interpreting and representing the imprecise sense of the words through the concept of fuzzy set. Briefly, a fuzzy set A, defined on a referential X is a set such that the membership function μ_A takes its value in the interval [0, 1]. $\forall x \in X$, $\mu_A(x)$ expresses to what extent the value x belongs to A. The value 0 corresponds to the absolute non-membership and the value 1 corresponds to the absolute membership. In the context of quantifiable attributes, the referential X is the set of reels $\Re$ and we adopt the classical trapezoidal representation of a fuzzy set (figure 2). For each attribute, a finite universe of discourse (Ω) is defined. Each element of Ω is a fuzzy set corresponding to a *LT* value. For example, the figure 2 shows the universe of discourse Ω={small, medium, large} defined on the referential X=[0, 30] with the unit 10^{-6}m for the attribute "size of an histological area".

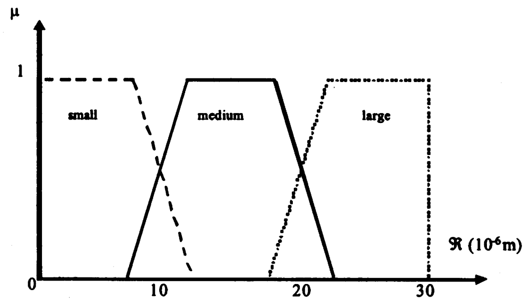


Figure 2: The universe of discourse Ω={small, medium, large} for the attribute "size of an histological area".

Fuzzy predicates are fuzzy sets not included in the universe of discourse. An example of a fuzzy predicate is "most cells are rather small but some are medium". "Most", "rather" and "some" are quantifiers while "small" and "medium" are *LT* values. The fuzzy set attached to a fuzzy predicate is dynamically built from the combination of linguistic terms with fuzzy quantifiers and fuzzy operators. The definition of the membership function is based on the fuzzy set theory.

Fuzzy quantifiers. Quantifiers are defined by fuzzy sets. Let T be an *LT* value with a membership function μ_T and Q a quantifier with a membership function μ_Q. The fuzzy predicate P= QxT is defined by:

$$\forall x \in \Re, \mu_p(x) = \mu_Q(x) * \mu_T(x).$$

Proportional quantifiers express intermediate situations between the universal quantifier and the existential quantifier like "some", "a few", "rarely", etc and are defined on the interval [0%, 100%]. For example, the quantifier "few" is defined around 25%. Semantic quantifiers express a modification in the meaning of an *LT* like for instance, "rather" or "very".

Fuzzy operators. In the context of a property, operators are disjunctive. Let P be a *FP* value such that P = LT1 and LT2, then

$$\forall x \in \Re, \mu_p(x) = \max(\mu_{LT1}(x), \mu_{LT2}(x)).$$

Comparison of polymorphous properties

Providing a powerful representation mechanism is not enough. One must also allow the systematic comparison of instances of such properties. Because of symmetry, a total of six possible value format combinations need to be considered namely *(RN,RN)*, *(RN,LT)*, *(RN,FP)*, *(LT,LT)*, *(LT,FP)*, *(FP, FP)*. They can be grouped into 1) crisp/crisp comparison, 2) crisp/fuzzy comparison and 3) fuzzy/fuzzy comparison. As the method for each group is in principle the same, only one combination per group has to be investigated. The method comes from the possibility theory.

The comparison process takes two properties as input and provides a compatibility degree between them as output. In the possibility theory, two scalar measures, the possibility and the necessity that the two properties describe the same morphological characteristic usually express the compatibility. These two measures are dual and we limit our presentation to the possibility measure Π. Let, d_1 and d_2 be two *RN* properties and F_1, F_2 two fuzzy sets with membership functions μ_{F1}, μ_{F2} associated to *LT* or *FP* properties. The expression of the possibility in each situation is given in table 1.

Table 1: Expression of the possibility degree

Comparison	**Possibility Π**
crisp/crisp	$\Pi(d_1,d_2) = 1$ if $d_1 = d_2$ $\Pi(d_1,d_2) = 0$ if $d_1 \neq d_2$
crisp/fuzzy	$\Pi(d_1,F_1) = \mu_{F1}(d_1)$
fuzzy/fuzzy	$\Pi(F_1,F_2) = \sup\min(\mu_{F1}(x), \mu_{F1}(x))$ for $x \in \Re$

RESULTS

Properties selection

We selected seven quantitative attributes from a set of 34 pathology reports. These reports describe cases in breast pathology from the Department of Pathology at the Institute Gustave Roussy, France. Table 2 reflects the polymorphism of properties in reports. It shows that mostly linguistic terms and fuzzy predicates were used. For many attributes, linguistic terms are preferred to fuzzy predicates. In the case of the size of a histological area, the properties are equally distributed between linguistic terms and fuzzy predicates.

Table 2: Selected quantitative attributes

Attribute	contexts	RN	LT	FP
Size	cells, lobules, cyst, seat		15	17
Mitotic_Activity	cell		8	2
Cytoplasm_Color	cell		3	1
Cytoplasm_Density	cell		9	0
Composite_Area_Density	cell		2	1
Composite_Area_Number	Cyst, lobule, seat	1	21	4
Composite_Area_Dispersion	seat		4	1

For instance, concerning the size of lobular glands, we found as fuzzy predicates, "diverse", "rather large", "often large" and "very diverse". Fuzzy predicates for Composite_Area_Number are "rather some", "rather numerous" and "one or two" while linguistic terms are "many", "numerous", "rare", "some".

PCF implementation

The property concept frame was developed in Java on a PC Pentium II with the ObjectStore object oriented database PSE Pro Software™. Many interfaces have been developed to test the PCF. Some of these interfaces are used to configure the referentials, the components of the universe of discourse and their associated fuzzy sets. Five proportional quantifiers and three semantic quantifiers have been defined to be able to describe all the fuzzy predicates in reports. The experts' membership functions of these quantifiers are shown in table 3.

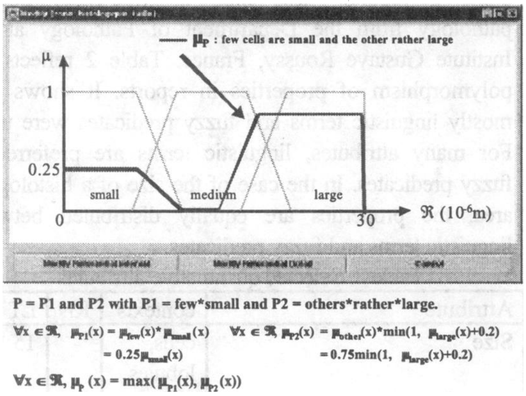
Table 3: Set of quantifiers in the IDEM context

Quantifier (Q)	fuzzy predicate P=Q*T
Proportional quantifiers	
all	$\mu_P(x) = \mu_T(x)$
many, a lot, most, often	$\mu_P(x) = 0.75\mu_T(x)$
on average	$\mu_P(x) = 0.5\mu_T(x)$
few, rare	$\mu_P(x) = 0.25\mu_T(x)$
none	$\mu_P(x) = 0$
Semantic quantifiers	
very, extremely	$\mu_P(x) = \mu_T(x + \lambda)^1$
rather	$\mu_P (x) = \min(1, \mu_T(x)+\nu)^2$

$^1\lambda$ is a function of T

$^2\nu = 0.2$

Figure 3 shows the fuzzy set for the *FP* value P = "few cells are small and others are rather large".



P = P1 and P2 with P1 = few*small and P2 = others*rather*large.

$\forall x \in \Re,\ \mu_{P1}(x) = \mu_{few}(x) * \mu_{small}(x) = 0.25\mu_{small}(x)$ $\forall x \in \Re\ \mu_{P2}(x) = \mu_{other}(x) * \min(1,\ \mu_{large}(x)+0.2) = 0.75\min(1,\ \mu_{large}(x)+0.2)$

$\forall x \in \Re,\ \mu_P(x) = \max(\mu_{P1}(x), \mu_{P2}(x))$

Figure 3 : The fuzzy set associated to the predicate "few cells are small and the other rather large"

Examples.

Table 3 shows the Π compatibility for the comparison between predicate and linguistic labels for the specific attribute "Mitotic_Activity of the cell".

Table 4: Possibility degrees between linguistic terms and fuzzy predicates for the Mitotic_activity

Property	a_1	a_2	a_3	a_4	a_5
a_1	1	0.5	0.37	0.2	0
a_2	0.5	1	0.5	0.33	0.5
a_3	0.37	0.5	0.75	0.5	0.2
a_4	0.2	0.33	0.5	0.75	0.5
a_5	0	0.5	0.2	0.5	1

a_1: low; a_2: moderate; a_3: not very high; a_4:on average high; a_5: high

DISCUSSION AND CONCLUSION

In domains without formal standardized semiology to describe images, it is necessary to manage imprecise and uncertain properties for developing image content-based retrieval tools. The representation of imprecise and uncertain information aims at reducing the ambiguity in the interpretation of knowledge. An important source of ambiguity is linked to the polymorphism of information. Although several works have been done on the modeling of imprecise and uncertain properties in the context of fuzzy logic's [8][9][12], the concept of fuzzy predicate is rarely taken into account. In a recent work, Dubitsky [10] has introduced this notion in knowledge representation formalism. In his framework, fuzzy sets are explicitly associated to fuzzy predicates. We propose an extension of this framework where fuzzy predicates can be defined dynamically through the use of fuzzy quantifiers and fuzzy operators. We define a compatibility degree between two properties based on the possibility theory.

Nowadays, properties comparison within the IDEM environment allows the crisp/crisp situation and label/label situation. The later makes use of predefined similarity matrix where all possible labels (linguistic terms and fuzzy predicates) have to be known. The PCF representation provides in addition a flexible mechanism to freely specify a value by introducing and defining a new concept and its semantics dynamically. It allows comparing this new value with the already known universe of discourse as well as other fuzzy predicates.

From the domain point of view, the PCF representation has several advantages. On the one hand, it allows the integration of the complex formalism of some properties found in reports thanks to the notion of fuzzy predicate. On the other hand, the use of fuzzy logic completes the traditional representation of attribute-value properties based on a single universe of discourse. Moreover, the fuzzy modeling allows taking into account the subjective sense of the terms used by one or several pathologists in their reports. Even if the description of images is going to be more and more standardized in the future with more linguistic values and less fuzzy predicates, the problem to consider complex properties is an important issue since such properties are likely to remain in queries.

The main difficulty of this approach is to know the referentials and universe of discourse associated to the properties. Indeed, many properties are not quantifiable. In that case, the concept frame is reduced to a discrete referential whose elements are the labels of the universes of discourse. Fuzzy predicates can be defined on such frames. In the realized work, the knowledge about the referential and the knowledge about the possible quantifiers come from the experts and must be validated through time and use of the system. We provided default values for the quantifiers in our specific demonstration context. One perspective would be to automatically define quantifiers.

One important aspect in the comparison of fuzzy predicates is to retrieve compatibility between descriptions of similar properties that use opposite labels. For instance, the comparison of the predicate "cells are often large" with the predicate "cells are rarely small" returns a non-null compatibility. An advantage is the ability to define a same property in different contexts and to be able to return a null compatibility. Let's take for example the attribute "Mitotic_Activity". In the context of smooth tissue, the associated referential is [0, 50] while in the context of breast tumors, the associated referential is [0, 15]. The comparison of the property "Mitotic_Activity moderate" in the first context and the property "Mitotic_Activity moderate" in the second one returns a null compatibility which is intuitively correct since "moderate" corresponds to different linguistic terms (fuzzy sets) in the two cases.

The next steps of this work will be to aggregate the result of comparison at the level of the case description and to integrate this approach within the case based reasoning module of the IDEM environment.

References

[1] Klossa J., Cordier JC., Flandrin G., Got C., Hemet J.A. European de facto standard for image folders applied to telepathology and teaching. Int J Med Inf 1998 ;48 (1-3) : 207-16.

[2] Berman, J. J.; Moore, G. W. SNOMED-encoded surgical pathology databases: a tool for epidemiologic investigation Mod Pathol. 1996 ; 9 (9) : 944-50

[3] Kostomanolakis S.' Lourakis M. Chronaki C. Kavaklis Y. Orphanoudakis S.C. The Architecture of a System for the Indexing of Images by Content. In : Proceedings of the International Symposium CAR'93. Computer Assisted Radiology. H.U. Lemke, K.Inamura, C.C. Jaffe and R Felix (eds). Berlin : Springer-Verlag. 1993; pp279- 282.

[4] Aubry F., Chameroy V. Lavaire F. Ramond J.P. Saidane I.E., Giron A., Bizais Y. Todd-Pokropek A., Di Paola R. Medical Image Management Using a Semantic Approach: Image Description. In : Proceedings of the International Symposium CAR'93. Computer Assisted Radiology. H.U. Lemke, K.Inamura, C.C. Jaffe and R Felix (eds). Berlin: Springer-Verlag 1993; pp265- 271.

[5] Le Bozec C, Jaulent MC, Zapletal E, Heudes D, Degoulet P. A visual coding system in histopathology and its consensual acquisition. *JAMIA* 99, N. M. Lorenzi (ed), Hanley & Belfus, Inc.: Philadelphia; pp 306-310.

[6] Klir Gj, Folger TA. Fuzzy sets, uncertainty and information. Prentice Hall, Englewood Cliffs, NJ, 1988.

[7] Zadeh L.A. Fuzzy Sets. Information and control, 1965; 8:338-353.

[8] Dubois D, Prade H, etc. Fuzzy set-based models in case-based reasoning. *Un cadre formel pour le raisonnement interpolatif et le raisonnement par cas.* Prade H eds. Papport IRIT (Institut de recherche en informatique de Toulouse)/96-54-R, 1996 ; pp. 1-26

[9] Salotti S. Filtrage flou et représentation centrée-objet pour raisonner par analogie : le système FLORAN. Thèse du Doctorat en science, Université Paris VI, 1992

[10] Dubitzky W, Schuster A, Bell DA, Hughes JG, Adamson K. How Similar is VERY YOUNG to 43 Years of Age? On the Representation and Comparison of Polymorphic Properties, in (eds) Proc. *15th Int. Joint Conf. on Artificial Intelligence*. Japan, 1997 : pp226-231.

[11] Jaulent MC, Le Bozec C, Zapletal E, Degoulet P. Case based diagnosis in histopathology of Breast Tumours.MEDINFO 98, B. Cesnik et al. (Eds), Amsterdam: IOS Press. 1998; pp. 544-548.

[12] Pivert O. Expression et évaluation des requêtes floues dans les bases de données. *Les applications des ensembles fous.* 1992 ; 285-92.

[13] Sztandera LM, Goodenday LS, Cios KJ. A neuro-fuzzy algorithm for diagnosis of coronary artery stenosis. Comput Biol Med. 1996 Mar;26(2):97-111.