Gwenole Quellec[1,3], Mathieu Lamard[2,3], Guy Cazuguel[1,3],
Beatrice Cochener[2,3,4], Christian Roux[1,3]

1: Institut Telecom; TELECOM Bretagne;
UEB; Dpt ITI, Brest, F-29200 France;
2: Univ. Bretagne Occidentale, Brest, F-29200 France;
3: Inserm, U650, Brest, F-29200 France;
4: CHU Brest, Service d'Ophtalmologie, Brest, F-29200 France;

Multimodal information retrieval based on DSmT. Application to computer-aided medical diagnosis

Published in:
Florentin Smarandache & Jean Dezert (Editors)
Advances and Applications of DSmT for Information Fusion
(Collected works), Vol. III
American Research Press (ARP), Rehoboth, 2009
ISBN-10: 1-59973-073-1
ISBN-13: 978-1-59973-073-8
Chapter XVIII, pp. 471 - 502

Abstract: *We propose in this chapter a content-based information retrieval framework to select documents in a database, consisting of several images with semantic information. Information in these documents is not only heterogeneous, but also often incomplete. So,the method we propose may cover a wide range of applications. To select the most relevant cases in the database, for a query, a degree of match between two cases is defined for each case feature, and these degrees of match are fused. Two different fusion models are proposed: a Shafer's model consisting of two hypotheses and a hybrid DSm model consisting of N hypotheses, where N is the number of cases in the database. They allow us to model our confidence in each feature, and take it into account in the fusion process, to improve the retrieval performance. To include images in such a system, we characterize them by their digital content. The proposed methods are applied to two multimodal medical databases for computer aided diagnosis; a comparison with other retrieval methods we proposed recently is provided. A mean precision at five of 81.8% and 84.8% was obtained on a diabetic retinopathy and a mammography database, respectively: the methods are precise enough to be used in a diagnosis aid system.*

18.1 Introduction

Case-based reasoning (CBR) [1] was introduced in the early 1980s as a new decision support tool. It is based on the assumption that analogous problems have similar solutions: to help interpreting a new case, similar cases are retrieved from a database and returned to the user. In this chapter, we focus on CBR in multimodal databases. To retrieve heterogeneous information, some simple approaches, based on early fusion [21, 24] or late fusion [14, 26] have been introduced in the literature. Recently, an elaborate retrieval method, based on dissimilarity spaces and relevance feedback, has also been proposed [5]. In the same time, we proposed several other approaches that do not rely on relevance feedback, and can efficiently manage missing information and the aggregation of heterogeneous features (symbolic and multidimensional digital information). The first approach is based on decision trees [15]. The second one is based on Bayesian networks [16]. We introduce in this chapter a third approach, based on DSmT: information coming from each case feature F_i, $i = 1..M$, is used to derive an estimation of the degree of match between a query case and a case in the database. A case feature F_i can be either a nominal variable, an image acquired using a given modality, or any other type of signal. These estimations are then fused, in order to define a consensual degree of match, which is used to retrieve the most relevant cases for the query. In order to model our confidence in the estimation provided by each source of evidence, we propose two fusion models based on DSmT. The first one is based on a Shafer's model consisting of two hypotheses. The second one is based on a hybrid DSm model consisting of N hypotheses, where N is the number of cases in the database. Finally, the cases in the database maximizing the consensual degree of match with the query are returned to the user.

The proposed approach is applied to computer-aided diagnosis. In medicine, the knowledge of experts is a mixture of textbook knowledge and experience through real life clinical cases. So, the assumption that analogous problems have similar solutions is backed up by physicians' experience. Consequently, there is a growing interest in the development of medical decision support systems based on CBR [4], especially to assist the diagnosis of physicians. Such systems are intended to be used as follows: when a physician has doubts about his diagnosis, he sends the available patient data to the system. The most similar cases, along with their medical interpretations, are retrieved from the database and returned to the physician, who can then compare his case to these retrieved cases. Reasoning by analogy, the physician may so confirm or invalidate his diagnosis.

Medical cases often consist of digital information like images and symbolic information such as clinical annotations. Diabetic retinopathy experts, for instance, analyze multimodal series of images together with contextual information, such as the age, the sex and the medical history of the patient. So, to use all the information available, we have to manage both digital and semantic information. On one hand, there are some medical CBR systems designed to manage symbolic information [6].

On the other hand, some systems, based on Content-Based Image Retrieval [20], have been designed to manage digital images [13]. However, few attempts have been made to merge the two kinds of approaches. Actually, in some systems, it is possible to formulate both textual and digital queries in parallel [2, 11], but the two kinds of information are processed separately. In another system, a text based and an image based similarity measure are combined linearly into a common similarity measure [18]. Nevertheless, in our application, none of those solutions is suitable to use at best the relationships between symbolic and digital information. Our approaches are efficient solutions for information retrieval based on both clinical descriptors and digital image features. More, they take into account the fact that the information is often incomplete and uncertain.

The objectives are detailed in section 18.2. Shafer's model is presented in section 18.3 and the hybrid DSm model in section 18.4. The proposed approaches are applied to computer-aided diagnosis of diabetic retinopathy and mammography in section 18.5: we provide a comparison with the other two multimodal information retrieval methods we proposed [15, 16]. We end with a discussion and a conclusion in section 18.6.

18.2 Objectives

As mentioned before, we have already proposed methods to manage databases with heterogeneous information. But they do not take into account the uncertainty of information and the possible conflicts between some feature values. We propose to evaluate the contribution of DSmT for medical CBR, in comparison with the other two multimodal retrieval methods we proposed. For this purpose, let us remind the evaluation procedure we use. Let $(x_j)_{j=1..N}$ be the cases in the database and x_q be a case placed as a query to the retrieval system. The system retrieves k cases from the database, where k is chosen by the users. The objective is to maximize the percentage of relevant cases, according to the users, among the k retrieved cases. This percentage is called the precision at k. For each method, we define a degree of match (or a similarity measure) between cases, and the k cases in the database maximizing the degree of match with x_q are retrieved. We tune the definition of the degree of match in order to maximize the percentage of relevant cases among the k retrieved cases, by training these methods. For this purpose, the cases in the database have to be classified by the users, in order to catch their perception of relevance between cases. Then, the database is divided into a training dataset $(x_j^T)_{j=1..N^T}$ and an evaluation dataset $(x_j^E)_{j=1..N^E}$.

18.3 Shafer's model for information retrieval

In order to select the k cases to retrieve for a query x_q, we compute the similarity of each case x_j in the database, $j = 1..N$, with x_q. For this purpose, we first estimate, for each case feature F_i, the degree of match between x_j and x_q according to F_i, denoted $dm_i(x_j, x_q)$. To compute these estimations, we define a finite number of states f_{is} for each feature F_i, $i = 1..M$, and we compute the membership degree of any case y to each state f_{is} of F_i, noted $\alpha_{is}(y)$. y denotes either x_q or x_j, $j = 1..N$. If F_i is a nominal variable, $\alpha_{is}(y)$ is Boolean; for instance, if y is a male then $\alpha_{``sex'',``male''}(y) = 1$ and $\alpha_{``sex'',``female''}(y) = 0$. If F_i is an image (or any type of signal), the definition of f_{is} and $\alpha_{is}(y)$ is given in section 18.3.1. The estimation of the degree of match between x_j and x_q according to F_i, namely $dm_i(x_j, x_q)$, is computed as described in section 18.3.2.

These estimations are then combined. The frame of discernment used in the fusion process is described in section 18.3.3. A belief mass function is first derived from each estimation of the degree of match, provided by a case feature (see section 18.3.4). It is designed in order to give more importance in the fusion process to sources of evidence in which we have more confidence. These belief mass functions are then fused (see section 18.3.5) and a consensual degree of match between x_j and x_q is derived: this consensual degree of match is used to find the k cases in the database maximizing the similarity with x_q (see section 18.3.6).

18.3.1 Image processing

If the case feature is a nominal variable, defining states f_{is} for F_i is straightforward, it is more difficult for images. To define states for images of a given type, we propose to follow the usual steps of Content-Based Image Retrieval (CBIR) [20], that is: 1) building a signature for each image (i.e. extracting a feature vector summarizing their digital content), and 2) defining a distance measure between two signatures. As a consequence, measuring the distance between two images comes down to measuring the distance between two signatures. Similarly, defining states for images of a given type comes down to defining states for the signatures of the corresponding images. For this purpose, we cluster similar image signatures, as described below, and we associate each cluster with a state. By this procedure, images can be processed by the retrieval method like any other feature.
In previous studies on CBIR, we used a customized wavelet decomposition to extract signatures from images [10]. These signatures characterize the distribution of the wavelet coefficients in each subband of the decomposition. Wouwer [25] showed that the wavelet coefficient distribution, in a given subband, can be modeled by a generalized Gaussian function. We define the signature as the juxtaposition of the maximum likelihood estimators of the wavelet coefficient distribution in each subband. To define a distance measure between signatures, we used a symmetric version of the Kullback-Leibler divergence between wavelet coefficient distributions [8]: the distance between two images is a weighted sum of these symmetrized divergences [10]. The ability to select a weight vector and a wavelet basis makes this image representation highly tunable.

In order to define states for images of a given type F_i, we cluster similar images with an unsupervised classification algorithm, thanks to the image signatures and the associated distance measure above. The Fuzzy C-Means algorithm (FCM) [3] was used for this purpose: each case y is assigned to each cluster s with a fuzzy membership $\alpha_{is}(y)$, $0 \leq \alpha_{is}(y) \leq 1$, such that $\sum_s \alpha_{is}(y) = 1$.

Other features can be discretized similarly: the age of a person, monodimensional signals, videos, etc.

18.3.2 Estimation of the degree of match for a feature F_i

We have to define a similarity measure between two cases from their membership degree to each state of a feature F_i. We could assume that cases with membership degrees close to that of x_q are the most liable to be relevant for x_q. So, a similarity measure between x_j and x_q, according to a case feature F_i, may be $\sum_s \alpha_{is}(x_j)\alpha_{is}(x_q)$. However, this assumption is only appropriate if all cases in a given class tend to be at the same state for F_i. Another model, more general, is used: we assume that cases in the same class are predominantly in a subset of states for F_i. So, to estimate the degree of match, we define a correlation measure S_{ist} between couples of states (f_{is}, f_{it}) of F_i, regarding the class of the cases at these states. S_{ist} is computed using the cases $(x_j^T)_{j=1..N^T}$ in the training dataset. Let $c = 1..C$ be the possible classes for a case in the database. We first compute the mean membership D_{isc} (resp. D_{itc}) of cases x_j^T in a given class c to the state f_{is} (resp. f_{it}):

$$D_{isc} = \beta \frac{\sum_j \delta(x_j^T, c)\alpha_{is}(x_j^T)}{\sum_j \delta(x_j^T, c)} \tag{18.1}$$

$$\sum_{c=1}^{C} D_{isc}^2 = 1, \forall(i,j) \tag{18.2}$$

where $\delta(x_j^T, c) = 1$ if x_j^T belongs to class c, $\delta(x_j^T, c) = 0$ otherwhise, and β is a normalizing factor chosen so that equation 18.2 holds. S_{ist} is given by equation 18.3:

$$S_{ist} = \sum_{c=1}^{C} D_{isc}D_{itc} \tag{18.3}$$

So we estimate the degree of match between the two cases x_j and x_q, with respect to a case feature F_i, as follows:

$$dm_i(x_j, x_q) = \sum_s \sum_t \alpha_{is}(x_j)S_{ist}\alpha_{it}(x_q) \tag{18.4}$$

18.3.3 Designing the frame of discernment

In order to estimate the relevance of a case x_j for the query x_q, as a consensus between all the sources of evidence, we define two hypotheses: Q="x_j is relevant for

x_q" and $\bar{Q}$="x_j is not relevant for x_q". The following frame of discernment is used in the fusion problem: $\Theta^{(1)} = \{Q, \bar{Q}\}$. To define the belief mass function associated with a given source of evidence, i.e. a feature F_i, we assign a mass to each element in $D^{\Theta^{(1)}} = \{\emptyset, Q, \bar{Q}, Q \cap \bar{Q}, Q \cup \bar{Q}\}$. In fact, it is meaningless to assign a mass to $Q \cap \bar{Q}$, as a consequence, we only assign a mass to elements in $D^{\Theta^{(1)}} \setminus Q \cap \bar{Q} = \{\emptyset, Q, \bar{Q}, Q \cup \bar{Q}\} = 2^{\Theta^{(1)}}$. We are thus designing a Shafer's model consisting of two hypotheses.

18.3.4 Defining the belief mass functions

To compute the belief mass functions, we define a test T_i on the degree of match $dm_i(x_j, x_q)$: T_i is true if $dm(x_j, x_q) >= \tau_i$ and false otherwise, $0 \leq \tau_i \leq 1$. The belief masses are then assigned according to T_i:

- if T_i is true:
  - $m_i(Q) = P(T_i | x_j\ is\ relevant\ for\ x_q) \rightarrow$ the sensitivity of T_i
  - $m_i(Q \cup \bar{Q}) = 1 - m_i(Q)$
  - $m_i(\bar{Q}) = 0$
- else
  - $m_i(\bar{Q}) = P(\bar{T}_i | x_j\ is\ not\ relevant\ for\ x_q) \rightarrow$ the specificity of T_i
  - $m_i(Q \cup \bar{Q}) = 1 - m_i(\bar{Q})$
  - $m_i(Q) = 0$

The sensitivity (resp. the specificity) represents the degree of confidence in a positive (resp. negative) answer to test T_i; $m_i(Q \cup \bar{Q})$, the belief mass assigned to the total ignorance, represents the degree of uncertainty: the higher this term, the lower our confidence in the case feature F_i. The sensitivity and the specificity of T_i, for a given threshold τ_i, are estimated using each pair of cases (x_a^T, x_b^T) in the training dataset, one playing the role of x_q, the other playing the role of x_j. The sensitivity (resp. the specificity) is estimated by the average number of pairs for which T_i is true (resp. false) among the pairs of cases belonging to the same class (resp. to different classes). T_i is appropriate if it is both sensitive and specific. As τ_i increases, sensitivity increases and specificity decreases. So, we set τ_i as the intersection of the two curves "sensitivity according to τ_i" and "specificity according to τ_i". A binary search is used to find the optimal τ_i.

18.3.5 Fusing the belief mass functions

If the i^{th} case feature is available for both x_j and x_q, the degree of match $dm_i(x_j, x_q)$ is estimated (see section 18.3.2) and the belief mass function is computed according to test T_i (see section 18.3.4). The computed belief mass functions are then fused. Let

$M' \leq M$ be the number of belief mass functions to fuse. Usual rules of combination have a time complexity exponential in M', which might be a limitation. So we propose a rule of combination for problems consisting of two hypotheses (Q and $\bar{Q}$ in our application), adapted from the Proportional Conflict Redistribution (PCR) rules [19], with a time complexity evolving polynomially with M' (see appendix 18.8).

18.3.6 Identifying the most similar cases

Once the sources available for x_q are fused by the proposed rule of combination, a decision function is used to compute the consensual degree of match between x_j and x_q. We express this consensual degree of match either by the credibility (cf. equation 18.5), the plausibility (cf. equation 18.6), or the pignistic probability of Q (cf. equation 18.7).

$$Bel(A) = \sum_{B \in D^\Theta, B \subseteq A, B \not\equiv \emptyset} m(B) \tag{18.5}$$

$$Pl(A) = \sum_{B \in D^\Theta, A \cap B \not\equiv \emptyset} m(B) \tag{18.6}$$

$$BetP(A) = \sum_{B \in D^\Theta, B \not\equiv \emptyset} \frac{\mathcal{C}_\mathcal{M}(A \cap B)}{\mathcal{C}_\mathcal{M}(B)} m(B) \tag{18.7}$$

The notation $B \not\equiv \emptyset$ means that $B \neq \emptyset$ and B has not been forced to be empty through the constraints of the model $\mathcal{M}$; $\mathcal{C}_\mathcal{M}(B)$ denotes the number of parts of B in the Venn diagram of the model $\mathcal{M}(\Theta)$ [7, 22]. It emerges from our applications that using the pignistic probability of Q leads to a higher mean precision at k (more elaborate decision functions might improve the retrieval performance). The pignistic probability of Q, $BetP(Q)$, is computed according to equation 18.8.

$$BetP(Q) = m_f(Q) + \frac{m_f(Q \cup \bar{Q})}{2} \tag{18.8}$$

The k cases maximizing $BetP(Q)$ are then returned to the user.

18.3.7 Managing missing values

The proposed method works even if some features are missing for x_j and x_q: we simply take into account the sources of evidence available for both x_j and x_q. However, it may be more efficient to take also into account information available for only one of the two cases.

A solution is to use a Bayesian network modeling the probabilistic dependencies between the features F_i, $i = 1..M$. The Bayesian network is built from the training dataset automatically [16]. We use it to infer the posterior probability of the features F_i unavailable for x_j, but available for x_q. As a consequence, all the features available for x_j are used to infer the posterior probability of the other features. And all the features available for x_q are involved in the fusion process: a belief mass function is defined for each feature available for x_q.

18.4 Hybrid DSm model for information retrieval

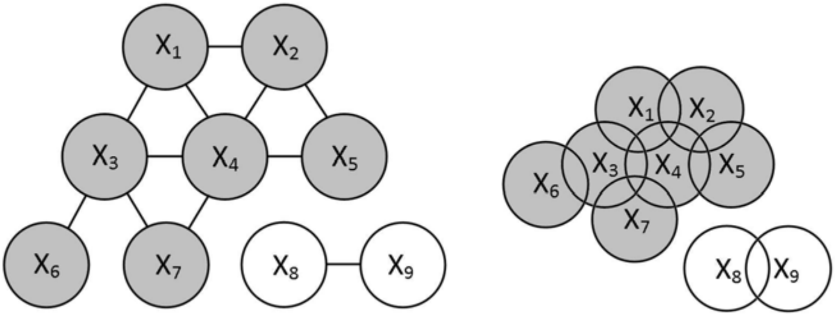
In the model presented in section 18.3, we have estimated the probability that each case x_j in the database is relevant for the case x_q placed as a query, $j = 1..N$. In this second model, we slightly reformulate the retrieval problem: we estimate the probability that x_q is relevant for each case x_j in the database, $j = 1..N$. The interest of this new formulation is that we can include in the model the similarity between cases x_j. To find the cases maximizing the similarity with the query in the database, we assess the following hypotheses X_j="x_q is relevant for x_j", $j = 1..N$, and we select the k most likely: the k corresponding cases x_j are thus returned to the user. As a consequence, a different frame of discernment is used (see section 18.4.1). The likelihood of each hypothesis X_j, $j = 1..N$, is estimated for each feature F_i, $i = 1..M$. These estimations are based on the same degree of match that was used in the previous model (see section 18.3.2).

Since we use a new frame of discernment, a new belief mass function is defined for each feature F_i (see section 18.4.2). These belief mass functions are then fused (see section 18.4.3). And a consensual estimation of the likelihood of X_j is derived: this consensual estimation of the likelihood is used to find the cases in the database maximizing the similarity with x_q (see section 18.4.4).

18.4.1 Designing the frame of discernment

The following frame of discernment is used in the new fusion problem: $\Theta^{(2)} = \{X_1, X_2, ..., X_N\}$. The cardinal of $D^{\Theta^{(2)}}$ is hyper-exponential in N. As a consequence, to solve the fusion problem, it is necessary to set some constraints in the model. We are thus designing a hybrid DSm model. These constraints are also justified from a logical point of view: *a priori*, if two cases x_a and x_b are dissimilar, or if x_a and x_b belong to different classes (as indicated by the users), then the two hypotheses X_a and X_b are incompatible.

To design the frame of discernment, we first build an undirected graph $G_c = (V, E)$, that we call compatibility graph. Each vertex $v \in V$ in this graph represents an hypothesis, and each edge $e \in E$ represents a couple of compatible hypotheses. To build the compatibility graph, each case x_j in the database, $j = 1..N$, is connected in the compatibility graph G_c to its l nearest neighbors. The distance measure we used to find the nearest neighbors is simply a linear combination of heterogeneous distance functions (one for each case feature F_i), managing missing information [24]. The complexity of the fusion process mainly depends on the cardinality of the largest clique in G_c (a clique is a set of vertices V such that for every pair of vertices $(u, v) \in V^2$, there is an edge connecting u and v). The number l is closely related to the cardinality of the largest clique in G_c and consequently to the complexity of the fusion process. l was set to five in the application (see section 18.5). The Venn diagram of the model $\mathcal{M}(\Theta^{(2)})$ is then built: for this purpose, we identify the cliques in G_c, as described in figure 18.1.



(a) Compatibility graph

(b) Venn diagram

Figure 18.1: Building the frame of discernment from the compatibility graph. Hypotheses associated with cases in the same class are represented with the same color.

18.4.2 Defining the belief mass functions

For each feature F_i, the belief mass function m_i is defined as follows. We first identify the set of cases $(x_j)_{j=1..N' \leq N}$ such that $dm_i(x_j, x_q)$ is greater than a threshold τ_i', $j = 1..N' \leq N$:

- a belief mass m_{i1} is assigned to the set $\bigcup_{j=1}^{N' \leq N} X_j$,
- and a belief mass $m_{i2} = 1 - m_{i1}$ is assigned to the total ignorance $\bigcup_{j=1}^{N} X_j$.

τ_i' is searched similarly to threshold τ_i (see section 18.3.4) with the following test: X_j is true if $dm_i(x_j, x_q) \geq \tau_i'$, otherwise X_j is false; we perform a binary search to find the threshold maximizing the minimum of the sensitivity and of the specificity of that test, whatever X_j (a single threshold τ_i' is searched for each case feature). m_{i1} is defined as the sensitivity of that test.

18.4.3 Fusing the belief mass functions

Once the Venn diagram of the model $\mathcal{M}(\Theta^{(2)})$ has been designed, we associate a unique number with each element in this diagram. The belief mass function m_i defined above is then encoded as follows:

- a binary string denoted $e_i(A)$ is assigned to each set $A \in D^{\Theta^{(2)}}$ such that $m_i(A) \neq 0$,

- the j^{th} character in the string $e_i(A)$ is 1 if and only if the j^{th} set in the Venn diagram is included in A.

In memory, the binary strings are encoded as byte strings: we associate each element in the diagram with a bit, and bits are grouped eight by eight into bytes. The elements of the Venn diagram form a partition of $\Omega = \bigcup_{j=1}^{N} X_j$, as a consequence, the following equation holds:

$$e_i(A \cap B) = e_i(A) \cap e_i(B) \tag{18.9}$$

Let us consider the following three-source problem, illustrated in figure 18.2.

The frame of discernment consists of five elements: $\Theta^{(2)} = \{X_1, X_2, X_3, X_4, X_5\}$, where $X_1 = \{0, 5, 6, 9\}$, $X_2 = \{1, 5, 7, 9\}$, $X_3 = \{2, 6, 7, 9\}$, $X_4 = \{3, 8\}$ and $X_5 = \{4, 8\}$.

These belief mass functions are fused sequentially:

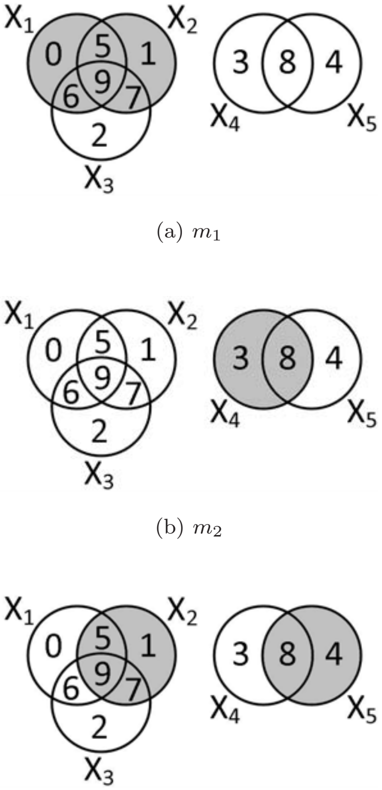
- fusion of m_1 and m_2 by the PCR5 rule → m_{12} [19],

- fusion of m_{12} and m_3 by the PCR5 rule → m_{123},

- etc.

As we fuse the belief mass functions, the number of elements $A \in D^{\Theta^{(2)}}$ satisfying $m_j(A) \neq 0$ increases. To access these elements and update their mass, we rank them in alphabetical order of $e_i(A)$: we can thus access them quickly with a binary search algorithm.

Detecting conflicts between two sources is made easier with this representation: if $e_i(A) \cap e_i(B) = 0$, $A \in D^{\Theta^{(2)}}$, $B \in D^{\Theta^{(2)}}$, then A and B are conflicting. On the example above, the fused belief mass function we obtain is illustrated in figure 18.3.



(a) m_1

$$m_1(X_1 \cup X_2) = 0.6,\ m_1(\Omega) = 0.4$$

$$\begin{aligned} &{*}{*}{*}{*}{*}{*}98, 76543210 \\ e_1(X_1 \cup X_2) &= 00000010, 11100011 \\ e_1(\Omega) &= 00000011, 11111111 \end{aligned}$$

(b) m_2

$$m_2(X_4) = 0.7,\ m_2(\Omega) = 0.3$$

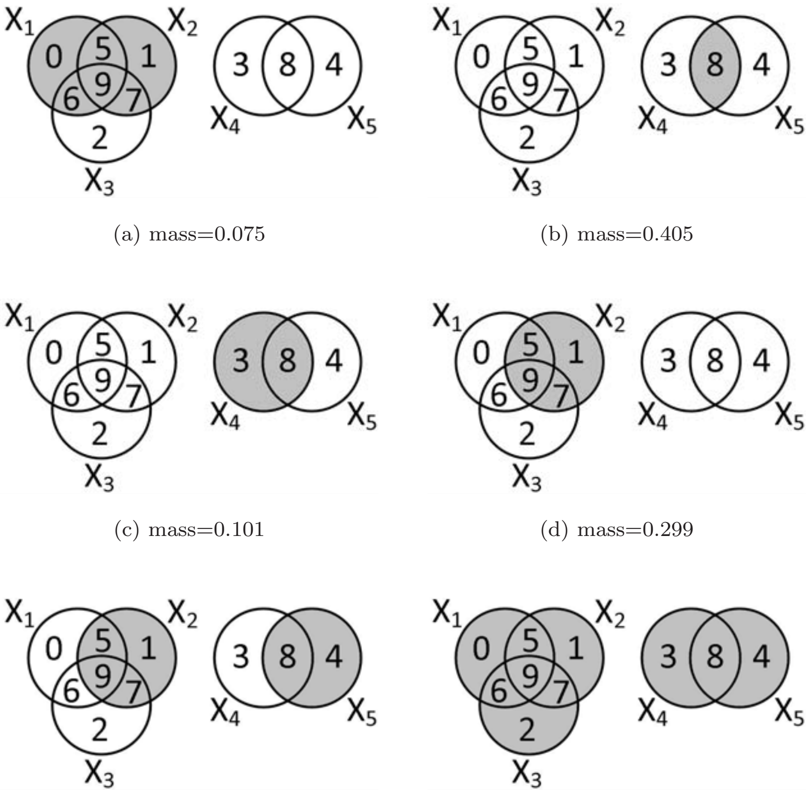
$$\begin{aligned} &{*}{*}{*}{*}{*}{*}98, 76543210 \\ e_2(X_4) &= 00000001, 00001000 \\ e_2(\Omega) &= 00000011, 11111111 \end{aligned}$$

(c) m_3

$$m_3(X_2 \cup X_5) = 0.8,\ m_3(\Omega) = 0.2$$

$$\begin{aligned} &{*}{*}{*}{*}{*}{*}98, 76543210 \\ e_3(X_2 \cup X_5) &= 00000011, 10110010 \\ e_3(\Omega) &= 00000011, 11111111 \end{aligned}$$

Figure 18.2: Encoding the belief mass functions.



(a) mass=0.075

(b) mass=0.405

(c) mass=0.101

(d) mass=0.299

(e) mass=0.096

(f) mass=0.024

Figure 18.3: Fused belief mass function: this figure represents the sets to which a non-zero belief mass has been assigned.

18.4.4 Identifying the most similar cases

Once the belief mass functions are fused, the pignistic probability of each element $X_j \in \Theta^{(2)}$ is computed (see equation 18.7), like in the previous model (see section 18.3.6): it is used as the consensual estimation of the likelihood of X_j. For instance, the computation of $BetP(X_4)$ is given below:

$$BetP(X_4) = \frac{1}{1} \cdot 0.405 + \frac{2}{2} \cdot 0.101 + \frac{1}{6} \cdot 0.096 + \frac{2}{10} \cdot 0.024 = 0.527 \qquad (18.10)$$

Then, the k cases x_j in the database maximizing $BetP(X_j)$ are returned to the user.

18.4.5 Managing missing values

Like in the previous model, we can use a Bayesian network to better manage missing information. The Bayesian network described in section 18.3.7 is used to infer the posterior probability of the features F_i unavailable for the query x_q.

18.5 Application to computer-aided medical diagnosis

The proposed methods have been evaluated on two multimodal medical databases, for computer-aided medical diagnosis. The first one (DRD) is being built at the Inserm U650 laboratory in collaboration with ophthalmologists of Brest University Hospital. The second one (DDSM) is a public access database.

18.5.1 Diabetic retinopathy database (DRD)

The diabetic retinopathy database contains retinal images of diabetic patients, with associated anonymized information on the pathology. Diabetes is a metabolic disorder characterized by a sustained high sugar level in the blood. Progressively, blood vessels are affected in many organs, which may lead to serious renal, cardiovascular, cerebral and also retinal complications. In the latter case, the pathology, namely diabetic retinopathy, can cause blindness. Patients have been recruited at Brest University Hospital since June 2003.

The database consists of 67 patient records containing 1112 photographs altogether. The disease severity level, according to ICDRS classification [23], was assessed by experts for each patient. The distribution of the disease severity among the 67 patients is given in table 18.1. Images have a definition of 1280 pixels/line for 1008 lines/image. They are lossless compressed images, acquired by experts using a Topcon Retinal Digital Camera (TRC-50IA) connected to a computer.

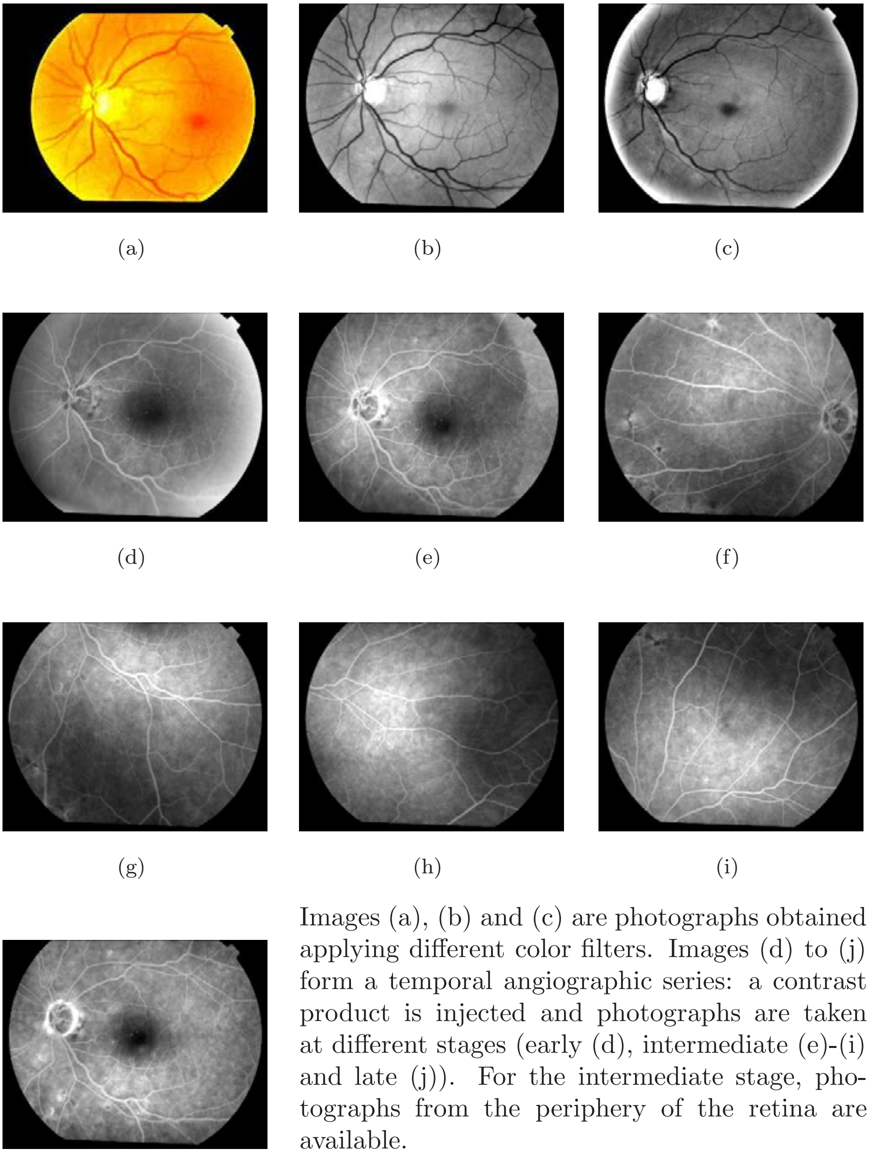
Database	Disease severity	Number of patients
DRD	no apparent diabetic retinopathy	7
	mild non-proliferative	9
	moderate non-proliferative	22
	severe non-proliferative	9
	proliferative	9
	treated/non active diabetic retinopathy	11
DDSM	normal	695
	benign	669
	cancer	913

Table 18.1: Patient disease severity distribution.

An example of image sequence is given in figure 18.4. The contextual information available is the age, the sex and structured medical information about the patient (see table 18.2). Patient records consist of 10 images per eye (see figure 18.4) and 13 contextual attributes at most; 12.1% of these images and 40.5% of these contextual attribute values are missing.

Attributes	Possible values
General clinical context	
family clinical context	diabetes, glaucoma, blindness, misc.
medical clinical context	arterial hypertension, dyslipidemia, protenuria, renal dialysis, allergy, misc.
surgical clinical context	cardiovascular, pancreas transplant, renal transplant, misc.
ophthalmologic clinical context	cataract, myopia, AMD, glaucoma,unclear medium, cataract surgery, glaucoma surgery, misc.
Examination and diabetes context	
diabetes type	none, type I, type II
diabetes duration	$<$ 1 year, 1 to 5 years, 5 to 10 years,$>$ 10 years
diabetes stability	good, bad, fast modifications, glycosylated hemoglobin
treatments	insulin injection, insulin pump, anti-diabetic drug + insulin, anti-diabetic drug, pancreas transplant
Eye symptoms before the angiography test	
ophthalmologically symptomatic	none, systematic ophthalmologic screening-known diabetes, recently diagnosed diabetes by check-up, diabetic diseases other than ophthalmic ones
ophthalmologically asymptomatic	none, infection, unilateral decreased visual acuity (DVA), bilateral DVA, neovascular glaucoma, intra-retinal hemorrhage, retinal detachment, misc.
Maculopathy	
maculopathy	focal edema, diffuse edema, none, ischemic

Table 18.2: Structured contextual information for diabetic retinopathy patients.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j)

Images (a), (b) and (c) are photographs obtained applying different color filters. Images (d) to (j) form a temporal angiographic series: a contrast product is injected and photographs are taken at different stages (early (d), intermediate (e)-(i) and late (j)). For the intermediate stage, photographs from the periphery of the retina are available.

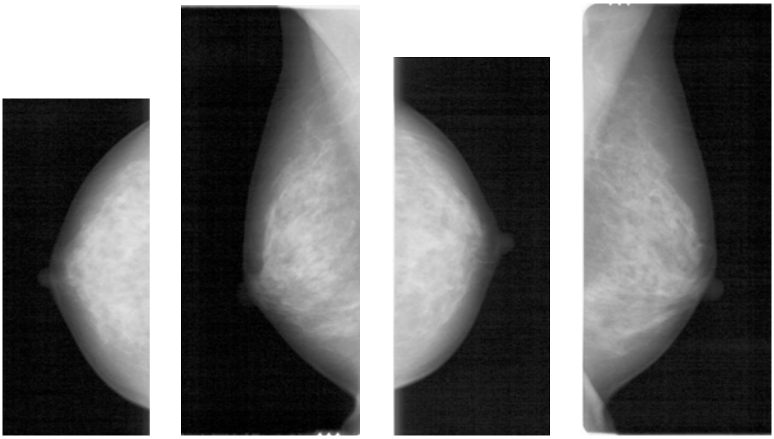
Figure 18.4: Photograph sequence of a patient's eye.

18.5.2 Digital database for screening mammography

The Digital Database for Screening Mammography (DDSM) project [9], involving the Massachusetts General Hospital, the University of South Florida and the Sandia National laboratories, has led to the setting-up of a mammographic image database for research on breast cancer screening. This database consists of 2277 patient records. Each one includes two images of each breast, associated with some information about the patient (the age, rating for abnormalities, American College of Radiology breast density rating and keyword description of abnormalities) and information about images (the scanner, the spatial resolution, etc). The following contextual attributes are taken into account in the system:

- the age of the patient,
- the breast density rating.

Images have a varying definition, of about 2000 pixels/line for 5000 lines/image. An example of image sequence is given in figure 18.5.



(a) LCC (b) LMLO (c) RCC (d) RMLO

Figure 18.5: Mammographic image sequence of the same patient. (a) and (b) are two views of the left breast, (c) and (d) are two views of the right one.

Each patient record has been graded by a physician. Patients are then classified into three groups: 'normal', 'benign' and 'cancer'. The distribution of grades among the patients is given in table 18.1.

18.5.3 Objective of the computer-aided diagnosis system

For each case placed as a query by a user, we want to retrieve the most similar cases from a given database. In DRD, the number of cases selected by the system is set to $k = 5$, at ophthalmologist's request; they consider this number sufficient for time reasons and in view of the good results provided by the system. For comparison purposes, the same number of cases is displayed in DDSM. The satisfaction of the user's needs can thus be assessed by the precision at five, the percentage of cases relevant for the query among the topmost five results.

18.5.4 Features of the patient records

In those databases, each patient record consists of both digital images and contextual information. Contextual features (13 in DRD, 2 in DDSM) are processed as in the CBR system. Images need to be processed in order to extract digital features. A usual solution is to segment images and extract domain specific information (such as the number of lesions). For DRD, we use the number of microaneurysms (the most frequent lesion of diabetic retinopathy) detected by the algorithm described in [17], in conjunction with other features. However, this kind of approach requires expert knowledge for determining pertinent information in images, and a robust segmentation of images, which is not always possible because of acquisition variability. This is the reason why we characterized images as described in section 18.3.1. An image signature is thus computed for each kind of image (10 for DRD, 4 for DDSM).

18.5.5 Training and evaluation datasets

Both databases are divided randomly into a training dataset (80% of the database) and an evaluation dataset (20% of the database). To assess the system, each case in the evaluation dataset is placed sequentially as a query to the system, and the five closest cases within the training dataset, according to the retrieval system, are retrieved. The precision at five is then computed. Because the datasets are small, especially for DRD, we use a 5-fold cross-validation procedure, so that each case in the database appears once in the evaluation dataset.

18.5.6 Results

The mean precision at five obtained with each method, on the two medical databases, is given in table 18.3; the proposed methods were compared to an early [24] and a late fusion method [14], and to the other two multimodal information retrieval methods we proposed [15, 16], as well. For both databases, we obtain a mean precision at five greater than 80%: it means that, on average, more than four cases out of the five cases retrieved by the system are relevant for the query.

Model	DRD	DDSM
Early fusion [24]	*42.8%*	*71.4%*
Late fusion [14]	*39.4%*	*70.3%*
Decision trees [15]	*81.0%*	*92.9%*
Bayesian networks [16]	*70.4%*	*82.1%*
Shafer's model	74.3%	77.3%
Shafer's model + Bayesian networks	80.8%	80.3%
Hybrid DSm model	78.6%	82.1%
Hybrid DSm model + Bayesian networks	81.8%	84.8%

Table 18.3: Mean precision at five for each method.

Clearly, simple early or late fusion methods are inappropriate to retrieve patient records efficiently: in the rest of the section, we will focus on the other methods.

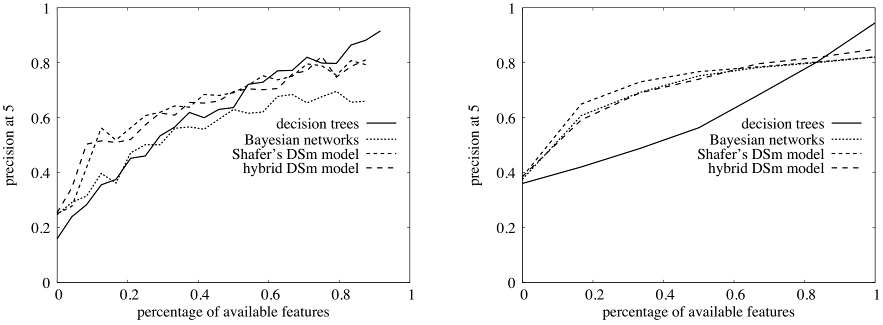
The mean computation time required to retrieve the five most similar cases, using each method, is given in table 18.4. All experiments were conducted using an AMD Athlon 64-bit based computer running at 2 GHz. Most of the time is spent during the computation of the image signatures. However, note that, if the wavelet coefficient distributions are simply modeled by histograms, the time required to compute the signatures can be greatly reduced (0.25 s instead of 4.57 s for DRD, 2.21 s instead of 35.89 s for DDSM).

To study the robustness of these methods, with respect to missing values, the following procedure has been carried out:

- for each case x_j in the database, $j = 1..N$, 100 new cases are generated as follows. Let n_j be the number of features available for x_j, each new case is obtained by removing a number of features randomly selected in $\{0, 1, ..., n_j\}$.
- we plot the precision at five according to the number of available features for each generated case (see figure 18.6).

Database	DRD	DDSM
computing the signatures (for 1 image)	4.57 s	35.89 s
computing the distance with each image signature in the database (for 1 image)	0.033 s	1.14 s
mean retrieval time (decision trees [15])	*17.24 s*	*99.50 s*
mean retrieval time (Bayesian networks [16])	*40.12 s*	*148.23 s*
mean retrieval time (Shafer's model)	32.21 s	148.13 s
mean retrieval time (Shafer's model + Bayesian networks)	40.58 s	148.27 s
mean retrieval time (Hybrid DSm model)	33.02 s	149.94 s
mean retrieval time (Hybrid DSm model + Bayesian networks)	40.77 s	150.01 s

Table 18.4: Computation times.



(a) DRD

(b) DDSM

Figure 18.6: Robustness with respect to missing values.

We can see from these plots that, using the two DSmT based methods, a satisfying precision at five can be obtained faster, as new features are available, than if using the decision tree based method. However, when the patient records are complete, the decision tree based method is more efficient. It is the case for DDSM, in which there are no missing information (see table 18.3). With the proposed methods, a sufficient precision can be reached before all the features are inputted by the user. As a consequence, the user can stop formulating his query when the returned results are satisfactory. On DRD for instance, a precision at five of 60% can be reached after inputting less than 30% of the features (see figure 18.6): with this precision, the majority of the retrieved cases (3 out of 5) belong to the right class.

18.6 Discussion and conclusion

In this chapter, we introduced two methods to include image sequences, with contextual information, in a CBR system. The first method is based on a Shafer's model consisting of two hypotheses. It is used to assess the relevance of each case in the database, independently, for the query. The second model is based on a hybrid DSm model consisting of N hypotheses, one for each case in the database. This model takes into account the similarities between cases in the database, to better assess their relevance for the query. Whatever the model used, the same similarity measure, between any case in the database and the query, is defined for each feature. Then, based on these similarity measures, a belief mass function, modeling our confidence in each feature, is designed. Finally, these belief mass functions are fused, in order to estimate the relevance of a case in the database for the query, as a consensus between all the available case features. For both models, a PCR rule of combination was used to fuse the belief mass functions. For Shafer's model, a new rule of combination, with a time complexity evolving polynomially with the number of sources is introduced in appendix 18.8. For the hybrid DSm model, a new encoding of the elements in the Dedekind lattice is proposed to allow the computation of the PCR5 rule of combination. The use of a Bayesian network is proposed to improve the management of unavailable features. These methods are generic: they can be extended to databases containing sound, video, etc: the wavelet transform based signature, presented in section 18.3.1, can be applied to any n-dimensional digital signal, using its n-dimensional wavelet transform ($n = 1$ for sound, $n = 3$ for video, etc). The methods are also convenient for they do not require being trained each time a new case is added to the database.

These methods have been successfully applied to two medical image data-bases, for computer aided diagnosis. For this application, the goal of the retrieval system is to select the five patient records, in a database, maximizing the similarity with the record of a new patient, examined by a physician. On both databases, a higher mean precision at five is obtained with the hybrid DSm model than with Shafer's model. The mean precision at five obtained for DRD (81.8%) is particularly interesting, considering the few examples available, the large number of unavailable features and the large number of classes taken into account. On this database, the methods outperform usual methods [14, 24] by almost a factor of 2 in precision. The improvement is also noticeable on DDSM (84.8% compared to 71.4%). On this database, these DSmT based methods are less precise than a previously proposed decision tree based method [15]. However, we showed that a satisfying precision at five can be obtained faster, as new features are available, using the DSmT based methods: this is interesting in a medical application, where patient records are sometimes incomplete. As a conclusion, the results obtained on both medical databases show that the system is precise enough to be used in a diagnosis aid system.

18.7 References

[1] A. Aamodt, *Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches*, AI Communications, Vol. 7, No. 1, pp. 39-59, 1994.

[2] S. Antani, L.R. Long, G.R. Thoma, *A biomedical information system for combined content-based retrieval of spine x-ray images and associated text information*, Proc. of the Indian Conference on Computer Vision, Graphics, and Image Processing, pp. 242-247, Ahmadabad, India, December 16-18, 2002.

[3] J.C. Bezdek, *Fuzzy Mathemathics in Pattern Classification*, Ph.D. thesis, Applied Math. Center, Cornell University, Ithaca, NY, U.S.A., 1973.

[4] I. Bichindaritz, C. Marling, *Case-based reasoning in the health sciences: What's next?*, Artificial Intelligence in Medicine, Vol. 36, No. 2, pp. 127-135, 2006.

[5] E. Bruno, N. Moenne-Loccoz, S. Marchand-Maillet, *Design of multimodal dissimilarity spaces for retrieval of video documents*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 30, No. 9, pp. 1520-1533, 2008.

[6] J.-M. Cauvin, C. Le Guillou, B. Solaiman, M. Robaszkiewicz, P. Le Beux, C. Roux, *Computer-assisted diagnosis system in digestive endoscopy*, IEEE Transactions on Information Technology in Biomedicine, Vol. 7, No. 4, pp. 256-262, 2003.

[7] J. Dezert, F. Smarandache, M. Daniel, *A generalized pignistic transformation*, in F. Smarandache and J. Dezert (editors): Advances and Applications of DSmT

for Information Fusion, chapter 7, pp. 143-153, American Research Press, 2004. http://www.gallup.unm.edu/˜smarandache/DSmT-book1.pdf.

[8] M.N. Do, M. Vetterli, *Wavelet-based texture retrieval using generalized Gaussian density and Kullback-Leibler Distance*, IEEE Transactions on Image Processing, Vol. 11, No. 2, pp. 146-158, 2002.

[9] M. Heath, K. Bowyer, D. Kopans, R. Moore, W.P. Kegelmeyer, *The Digital Database for Screening Mammography*, Proc. of the Fifth International Workshop on Digital Mammography (IWDM 2000), M.J. Yaffe (editor), pp. 212-218, Medical Physics Publishing, 2001.

[10] M. Lamard, G. Cazuguel, G. Quellec, L. Bekri, C. Roux, B. Cochener, *Content Based image retrieval based on Wavelet Transform Coefficients distribution*, Proc. of the 29th annual international conference of IEEE Engineering in Medecine and Biology Society (EMBS 2007), pp. 4532-4535, Lyon, France, August 23-26, 2007.

[11] C. Le Bozec, E. Zapletal, M.C. Jaulent, D. Heudes, P. Degoulet, *Towards content-based image retrieval in a HIS-integrated PACS*, Proc. of the Annual Symposium of the American Society for Medical Informatics (AMIA 2000), pp. 477-481, Los Angeles, CA, U.S.A., November 4-8, 2000.

[12] A. Martin, C. Osswald, *A new generalization of the proportional conflict redistribution rule stable in terms of decision*, in F. Smarandache and J. Dezert (editors): Advances and Applications of DSmT for Information Fusion II, chapter 2, pp. 69-88, American Research Press, 2006. http://www.gallup.unm.edu/˜smarandache/DSmT-book2.pdf.

[13] H. Müller, N. Michoux, D. Bandon, A. Geissbuhler, *A review of content-based image retrieval systems in medical applications - clinical benefits and future directions*, International Journal of Medical Informatics, Vol. 73, No. 1, pp. 1-23, 2004.

[14] R. Nuray, F. Can, *Automatic ranking of information retrieval systems using data fusion*, Information Processing and Management, Vol. 42, No. 3, pp. 595-614, 2006.

[15] G. Quellec, M. Lamard, L. Bekri, G. Cazuguel, B. Cochener, C. Roux, *Recherche de cas médicaux multimodaux à l'aide d'arbres de décision*, Ingénierie et Recherche BioMédicale (IRBM), Vol. 29, No. 1, pp. 35-43, 2008.

[16] G. Quellec, M. Lamard, L. Bekri, G. Cazuguel, C. Roux, B. Cochener, *Multimodal Medical Case Retrieval using Bayesian Networks and the Dezert-Smarandache Theory*, Proc. of the Fifth IEEE International Symposium on Biomedical Imaging (ISBI 2008), pp. 245-248, Paris, France, May 14-17, 2008.

[17] G. Quellec, M. Lamard, P.M. Josselin, G. Cazuguel, B. Cochener, C. Roux, *Optimal wavelet transform for the detection of microaneurysms in retina photographs*, IEEE Transactions on Medical Imaging, Vol. 27, No. 9, pp. 1230-1241, 2008.

[18] Hong Shao, Wen-Cheng Cui, Hong Zhao, *Medical Image Retrieval Based on Visual Contents and Text Information*, Proc. of the IEEE International Conference on Systems, Man and Cybernetics (SMC 2004), pp. 1098-1103, The Hague, The Netherlands, October 10-13, 2004.

[19] F. Smarandache, J. Dezert, *Proportional Conflict Redistribution Rules for Information Fusion*, in F. Smarandache and J. Dezert (editors): Advances and Applications of DSmT for Information Fusion II, chapter 1, pp. 3-68, American Research Press, 2006. http://www.gallup.unm.edu/˜smarandache/DSmT-book2.pdf.

[20] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, *Content-based image retrieval at the end of the early years*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 12, pp. 1349-1380, 2000.

[21] R.K. Srihari, A. Rao, B. Han, S. Munirathnam, X. Wu, *A model for multimodal information retrieval*, Proc. of the IEEE International Conference on Multimedia and Expo (ICME 2000), pp. 701-704, New York City, NY, U.S.A., 30 July - 2 August, 2000.

[22] J. Venn, *On the Diagrammatic and Mechanical Representation of Propositions and Reasonings*, Dublin Philosophical Magazine and Journal of Science, Vol. 9, No. 59, pp. 1-18, 1880.

[23] C. Wilkinson, F. Ferris, R. Klein et al., *Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales*, Ophthalmology, Vol. 110, No. 9, pp. 1677-1682, 2003.

[24] D.R. Wilson, T.R. Martinez, *Improved Heterogeneous Distance Functions*, Journal of Artificial Intelligence Research, Vol. 6, pp. 1-34, 1997.

[25] G.V. Wouwer, P. Scheunders, D.V. Dyck, *Statistical texture characterization from discrete wavelet representations*, IEEE Transactions on Image Processing, Vol. 8, No. 4, pp. 592-598, 1999.

[26] S. Wu, S. McClean, *Performance prediction of data fusion for information retrieval*, Information Processing and Management, Vol. 42, No. 4, pp. 899-915, 2006.

18.8 Appendix: PCR rule with polynomial complexity

In this appendix, we focus on frames of discernment consisting of two hypotheses $\Theta = \{\theta_1, \theta_2\}$. We make no assumptions on the model used for the fusion problem: it can either be Shafer's model, the free DSm model or a hybrid DSm model. We propose, in section 18.8.1, an algorithm to compute the conjunctive rule $m_\cap(X), \forall X \in D^\Theta$ (see equation 18.11), whose complexity evolves polynomially with the number of sources s.

$$m_\cap(X) = \sum_{(X_1,\dots,X_s)\in(D^\Theta)^s, X_1\cap\dots\cap X_s = X} \prod_{i=1}^{s} m_i(X_i) \tag{18.11}$$

Then we propose a new PCR rule, based on the same principle, in section 18.8.2. Let $k_{12\dots s}$ be the total conflicting mass:

$$k_{12\dots s} = \sum_{(X_1,\dots,X_s)\in(D^\Theta)^s, X_1\cap\dots\cap X_s \equiv \emptyset} \prod_{i=1}^{s} m_i(X_i) \tag{18.12}$$

Each term in this sum is called a partial conflicting mass. The principle of the PCR rules is to redistribute the total conflicting mass $k_{12\dots s}$ (PCR1, PCR2) or the partial conflicting masses (PCR3, ..., PCR6) between the sets $X_c \in D^\Theta$ involved in the conflict [12, 19]. The conflict is redistributed to each set X_c proportionally to their belief mass. We illustrate the PCR5 rule on the following problem with two hypotheses and two sources. Suppose for instance that θ_1 and θ_2 are exclusive, as a consequence $m_{PCR5}(\theta_1 \cap \theta_2) = 0$ and $k_{12} = m_1(\theta_1)m_2(\theta_2) + m_1(\theta_2)m_2(\theta_1)$. So $m_1(\theta_1)m_2(\theta_2)$ is redistributed between $m_{PCR5}(\theta_1)$ and $m_{PCR5}(\theta_2)$ proportionally to $m_1(\theta_1)$ and $m_2(\theta_2)$, respectively. Similarly, $m_1(\theta_2)m_2(\theta_1)$ is redistributed between $m_{PCR5}(\theta_1)$ and $m_{PCR5}(\theta_2)$ proportionally to $m_2(\theta_1)$ and $m_1(\theta_2)$, respectively. Indeed, $\theta_1 \cup \theta_2$ is not involved in the conflict. As a consequence, we obtain the following fused mass function:

$$\begin{cases} m_{PCR5}(\emptyset) = m_{PCR5}(\theta_1 \cap \theta_2) = 0 \\ m_{PCR5}(\theta_1) = m_\cap(\theta_1) + \frac{m_1(\theta_1)}{m_1(\theta_1)+m_2(\theta_2)} m_1(\theta_1)m_2(\theta_2) \\ \qquad\qquad + \frac{m_2(\theta_1)}{m_2(\theta_1)+m_1(\theta_2)} m_2(\theta_1)m_1(\theta_2) \\ m_{PCR5}(\theta_2) = m_\cap(\theta_2) + \frac{m_2(\theta_2)}{m_1(\theta_1)+m_2(\theta_2)} m_1(\theta_1)m_2(\theta_2) \\ \qquad\qquad + \frac{m_1(\theta_2)}{m_2(\theta_1)+m_1(\theta_2)} m_2(\theta_1)m_1(\theta_2) \\ m_{PCR5}(\theta_1 \cup \theta_2) = m_\cap(\theta_1 \cup \theta_2) \end{cases} \tag{18.13}$$

The algorithms we propose impose a constraint on the belief mass function m_i defined for each source $i = 1..s$ to fuse: only elements $X \in \{\theta_1, \theta_2, \theta_1 \cup \theta_2\}$ can have a non-zero mass. Nevertheless, the set $\theta_1 \cap \theta_2$ is taken into account within the rules of combination. The generalization to problems involving more hypotheses is discussed in section 18.8.3.

18.8.1 Algorithm for the conjunctive rule

Let m_1, m_2, ..., m_s be the belief mass functions defined for each source of evidence. From the constraint above, a belief mass $m_i(X)$ is assigned to each element $X \in D^\Theta$ for each source $i = 1..s$ according to:

$$\begin{cases} m_i(\theta_1) + m_i(\theta_2) + m_i(\theta_1 \cup \theta_2) = 1 \\ m_i(\theta_1 \cap \theta_2) = 0 \end{cases} \tag{18.14}$$

Consequently, the conjunctive rule is simplified as follows:

$$m_\cap(X) = \sum_{(X_1,...,X_s)\in\{\theta_1,\theta_2,\theta_1\cup\theta_2\}^s, X_1\cap...\cap X_s = X} \prod_{i=1}^{s} m_i(X_i) \tag{18.15}$$

For a two-source problem, we obtain:

$$\begin{cases} m_\cap(\theta_1 \cup \theta_2) = m_1(\theta_1 \cup \theta_2)m_2(\theta_1 \cup \theta_2) \\ m_\cap(\theta_1) = m_1(\theta_1)m_2(\theta_1) + m_1(\theta_1)m_2(\theta_1 \cup \theta_2) + m_1(\theta_1 \cup \theta_2)m_2(\theta_1) \\ m_\cap(\theta_2) = m_1(\theta_2)m_2(\theta_2) + m_1(\theta_2)m_2(\theta_1 \cup \theta_2) + m_1(\theta_1 \cup \theta_2)m_2(\theta_2) \\ m_\cap(\theta_1 \cap \theta_2) = m_1(\theta_1)m_2(\theta_2) + m_1(\theta_2)m_2(\theta_1) \end{cases} \tag{18.16}$$

Let us interpret the computation of $m_\cap$ graphically. For this purpose, we cluster the different products $p = \prod_{i=1}^{s} m_i(X_i)$, $(X_1, ..., X_s) \in \{\theta_1,\ \theta_2,\ \theta_1 \cup \theta_2\}^s$ according to:

- the number $n_1(p)$ of terms $m_i(\theta_1)$, $i = 1..s$, in p,
- the number $n_2(p)$ of terms $m_i(\theta_2)$, $i = 1..s$, in p.

Precisely, we create a matrix T_s in which each cell $T_s(u, v)$ contains the sum of the products $p = \prod_{i=1}^{s} m_i(X_i)$, $(X_1, ..., X_s) \in \{\theta_1,\ \theta_2,\ \theta_1 \cup \theta_2\}^s$ such that $n_1(p) = u$ and $n_2(p) = v$. In the case $s = 1$ and $s = 2$, we obtain the matrices T_1 and T_2, respectively, given in figure 18.7.

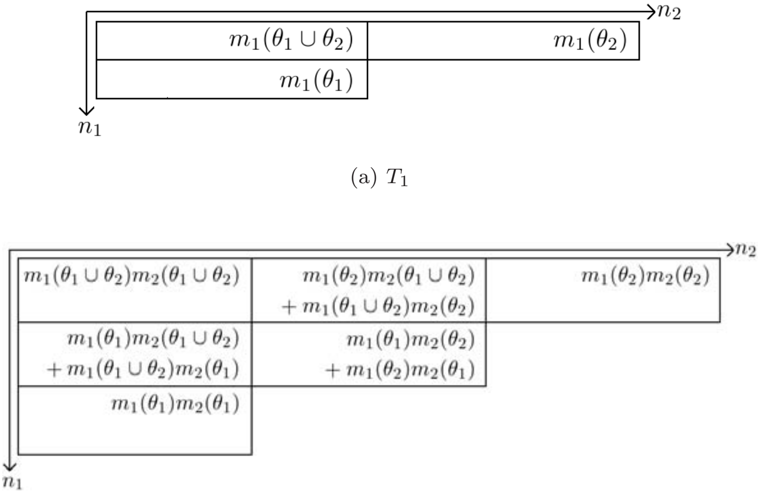
From figure 18.7 and equation 18.16, we can see that $m_\cap$ can be computed from T_s:

$$\begin{cases} m_\cap(\theta_1 \cup \theta_2) = T_s(0, 0) \\ m_\cap(\theta_1) = \sum_{u=1}^{s} T_s(u, 0) \\ m_\cap(\theta_2) = \sum_{v=1}^{s} T_s(0, v) \\ m_\cap(\theta_1 \cap \theta_2) = \sum_{u=1}^{s} \sum_{v=1}^{s} T_s(u, v) \end{cases} \tag{18.17}$$

The structure of matrix T_s is illustrated on figure 18.8.

Equation 18.17 can be explained from equation 18.15 as follows:

- in cell $T_s(0, 0)$, the intersection of the propositions $X_1 \cap ... \cap X_s$ is $\theta_1 \cup \theta_2$ because the product assigned to this cell does not contain any terms $m_i(\theta_1)$ or $m_i(\theta_2)$,
- in cells $T_s(u, 0)$, $u \geq 1$, the intersection of the propositions $X_1 \cap ... \cap X_s$ is θ_1 because each product assigned to these cells contains at least one term $m_i(\theta_1)$ (u terms, precisely) and does not contain any term $m_i(\theta_2)$,



$m_1(\theta_1 \cup \theta_2)$	$m_1(\theta_2)$
$m_1(\theta_1)$	

(a) T_1

$m_1(\theta_1 \cup \theta_2)m_2(\theta_1 \cup \theta_2)$	$m_1(\theta_2)m_2(\theta_1 \cup \theta_2)$ $+ m_1(\theta_1 \cup \theta_2)m_2(\theta_2)$	$m_1(\theta_2)m_2(\theta_2)$
$m_1(\theta_1)m_2(\theta_1 \cup \theta_2)$ $+ m_1(\theta_1 \cup \theta_2)m_2(\theta_1)$	$m_1(\theta_1)m_2(\theta_2)$ $+ m_1(\theta_2)m_2(\theta_1)$	
$m_1(\theta_1)m_2(\theta_1)$		

(b) T_2

Figure 18.7: Matrices T_1 and T_2

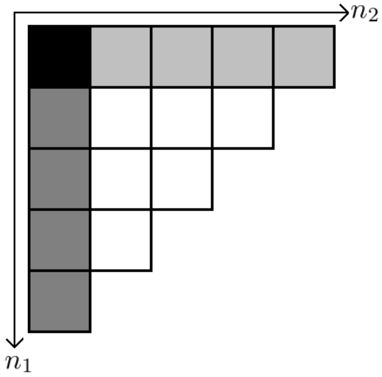


Figure 18.8: Structure of matrix T_s. According to equation 18.17, the black cells, the dark gray cells, the light gray cells and the white cells contain the belief mass assigned to $\theta_1 \cup \theta_2$, θ_1, θ_2 and $\theta_1 \cap \theta_2$, respectively.

- in cells $T_s(0,v)$, $v \geq 1$, the intersection of the propositions $X_1 \cap ... \cap X_s$ is θ_2 because each product assigned to these cells contains at least one term $m_i(\theta_2)$ (v terms, precisely) and does not contain any term $m_i(\theta_1)$,

- in cells $T_s(u,v)$, $u, v \geq 1$, the intersection of the propositions $X_1 \cap ... \cap X_s$ is $\theta_1 \cap \theta_2$ because each product assigned to these cells contains at least one term $m_i(\theta_1)$ (u terms, precisely) and at least one term $m_i(\theta_2)$ (v terms, precisely).

From equation 18.17, we see that if T_s can be built in a time polynomial in s, then $m_\cap$ can also be computed in a time polynomial in s.
We describe below an algorithm to build T_j from T_{j-1} in a time polynomial in j, $j = 2..s$. Its principle is illustrated on figure 18.9, in the case $j = 2$. We first compute three intermediate matrices $T_j^{\theta_1}$, $T_j^{\theta_2}$ and $T_j^{\theta_1 \cup \theta_2}$:

$$T_j^{\theta_1}(u,v) = \begin{cases} T_{j-1}(u-1,v) \times m_j(\theta_1), & u = 1..j, v = 0..j \\ 0, \quad u = 0, v = 0..j \end{cases} \tag{18.18}$$

$$T_j^{\theta_2}(u,v) = \begin{cases} T_{j-1}(u,v-1) \times m_j(\theta_2), & u = 0..j, v = 1..j \\ 0, \quad u = 0..j, v = 0 \end{cases} \tag{18.19}$$

$$T_j^{\theta_1 \cup \theta_2}(u,v) = T_{j-1}(u,v) \times m_j(\theta_1 \cup \theta_2), \quad u = 0..j, v = 0..j \tag{18.20}$$

T_j is then obtained as the sum of the three matrices:

$$T_j = T_j^{\theta_1} + T_j^{\theta_2} + T_j^{\theta_1 \cup \theta_2} \tag{18.21}$$

We first check that all the products $\prod_{i=1}^{j} m_i(X_i)$, $X_1, ..., X_j \in \{\theta_1, \theta_2, \theta_1 \cup \theta_2\}$ are generated by this procedure (hypothesis $\mathcal{H}_1(j)$). Then we check that these products appear in the correct cell of T_j (hypothesis $\mathcal{H}_2(j)$). Both hypothesis are checked by induction.

1. *Basis*: hypotheses $\mathcal{H}_1(1)$ and $\mathcal{H}_2(1)$ can be easily checked on figure 18.7 (a).

2. Suppose hypothesis $\mathcal{H}_1(j-1)$ is true. Each term $p = \prod_{i=1}^{j} m_i(X_i)$, $(X_1, ..., X_j) \in \{\theta_1,\ \theta_2,\ \theta_1 \cup \theta_2\}^j$, can be written as the product of a term $q = \prod_{i=1}^{j-1} m_i(X_i)$, $(X_1, ..., X_{j-1}) \in \{\theta_1,\ \theta_2,\ \theta_1 \cup \theta_2\}^{j-1}$, which appears in T_{j-1}, according to hypothesis $\mathcal{H}_1(j-1)$, and a belief mass m: m is either $m_j(\theta_1)$, $m_j(\theta_2)$ or $m_j(\theta_1 \cup \theta_2)$. According to m, p appears either in $T_j^{\theta_1}$, or in $T_j^{\theta_2}$, or in $T_j^{\theta_1 \cup \theta_2}$ (see equations 18.18, 18.19 and 18.20). As a consequence, p appears in T_j (see equation 18.21): hypothesis $\mathcal{H}_1(j)$ is true.

3. Suppose hypothesis $\mathcal{H}_2(j-1)$ is true. Let $p = \prod_{i=1}^{j} m_i(X_i)$, $(X_1, ..., X_j) \in \{\theta_1,$ $\theta_2, \theta_1 \cup \theta_2\}^j$. If X_j is θ_1, then by definition of n_1 and n_2, $n_1(p) = n_1(\frac{p}{m_j(\theta_1)})+1$ and $n_2(p) = n_2(\frac{p}{m_j(\theta_1)})$. According to equation 18.18, p appears in $T_j^{\theta_1}$ (and in T_j, consequently) one row below $\frac{p}{m_j(\theta_1)}$ in T_{j-1}. Since $\frac{p}{m_j(\theta_1)}$ appears in the correct cell of T_{j-1} (hypotheses $\mathcal{H}_2(j-1)$), p appears in the correct cell of T_j. A similar reasoning is applied if X_j is θ_2 (using equation 18.19) or $\theta_1 \cup \theta_2$ (using equation 18.20). As a consequence, hypothesis $\mathcal{H}_2(j)$ is true.

To compute T_j from T_{j-1}, $3(\frac{j(j+1)}{2})$ multiplications are required. Therefore, $\frac{3}{2}\sum_{j=1}^{s} j(j+1) = O(s^3)$ multiplications are required to compute T_s: the complexity of the proposed algorithm is polynomial in s.

18.8.2 Proposed PCR rule of combination

If hypotheses θ_1 and θ_2 are exclusive, then the belief mass assigned to $\theta_1 \cap \theta_2$ by the conjunctive rule is conflicting: $k_{12\ldots s} = m_{\cap}(\theta_1 \cap \theta_2)$. $\theta_1 \cup \theta_2$ is not involved in the conflict, as a consequence $k_{12\ldots s}$ should be redistributed between θ_1 and θ_2. In view of the number of partial conflicting masses $\prod_{i=1}^{s} m_i(X_i)$, $(X_1, ..., X_s) \in \{\theta_1, \theta_2, \theta_1 \cup \theta_2\}^s$, which is exponential in s, it is impossible to redistribute them individually (according to the PCR5 rule, for instance), if s is large. On the other hand, one could redistribute the total conflicting mass $k_{12\ldots s}$ (according to the PCR2 rule, for instance). Anyway a better solution is possible, taking advantage of the algorithm above: the conflicting mass can be redistributed more finely using matrix T_s.

As we build matrix T_s with the algorithm above, we compute in each cell c the percentages $p_1(c)$ and $p_2(c)$ of the belief mass in c that should be assigned to θ_1 and θ_2, respectively, in case of conflict. These percentages are initialized according to equation 18.22.

$$\begin{cases} p_1(T_1(0,0)) = p_1(T_1(0,1)) = 0 \\ p_1(T_1(1,0)) = 1 \\ p_2(T_1(0,0)) = p_2(T_1(1,0)) = 0 \\ p_2(T_1(0,1)) = 1 \end{cases} \tag{18.22}$$

At iteration j, after computing $T_j^{\theta_1}(u,v) = T_{j-1}(u-1,v) \times m_j(\theta_1)$, we compute for $u+v > 1$ and $u+v \leq j$:

$$\begin{cases} p_1(T_j^{\theta_1}(u,v)) = \frac{p_1(T_{j-1}(u-1,v))T_{j-1}(u-1,v)+m_j(\theta_1)}{T_{j-1}(u-1,v)+m_j(\theta_1)} \\ p_2(T_j^{\theta_1}(u,v)) = \frac{p_2(T_{j-1}(u-1,v))T_{j-1}(u-1,v)}{T_{j-1}(u-1,v)+m_j(\theta_1)} \end{cases} \tag{18.23}$$

Similarly, after computing $T_j^{\theta_2}(u,v) = T_{j-1}(u,v-1) \times m_j(\theta_2)$, we compute for $u+v > 1$ and $u+v \leq j$:

$$\begin{cases} p_1(T_j^{\theta_2}(u,v)) = \frac{p_1(T_{j-1}(u,v-1))T_{j-1}(u,v-1)}{T_{j-1}(u,v-1)+m_j(\theta_2)} \\ p_2(T_j^{\theta_2}(u,v)) = \frac{p_2(T_{j-1}(u,v-1))T_{j-1}(u,v-1)+m_j(\theta_2)}{T_{j-1}(u,v-1)+m_j(\theta_2)} \end{cases} \tag{18.24}$$

Then, after computing $T_j(u,v) = T_j^{\theta_1}(u,v) + T_j^{\theta_2}(u,v) + T_j^{\theta_1 \cup \theta_2}(u,v)$, we compute for $u+v > 1$ and $u+v \leq j$:

$$\begin{cases} p_1(T_j(u,v)) = \beta(u,v)(p_1(T_{j-1}(u,v)) + \\ \qquad \frac{p_1(T_j^{\theta_1}(u,v))T_j^{\theta_1}(u,v)+p_1(T_j^{\theta_2}(u,v))T_j^{\theta_2}(u,v)}{T_j^{\theta_1}(u,v)+T_j^{\theta_2}(u,v)}) \\ p_2(T_j(u,v)) = \beta(u,v)(p_2(T_{j-1}(u,v)) + \\ \qquad \frac{p_2(T_j^{\theta_1}(u,v))T_j^{\theta_1}(u,v)+p_2(T_j^{\theta_2}(u,v))T_j^{\theta_2}(u,v)}{T_j^{\theta_1}(u,v)+T_j^{\theta_2}(u,v)}) \end{cases} \tag{18.25}$$

$\beta(u,v)$ is a normalization factor, chosen so that $p_1(T_j(u,v)) + p_2(T_j(u,v)) = 1$ $\forall\ u, v$. Finally, the belief mass in each cell $(T_s(u,v))_{u>0,v>0}$ is redistributed between θ_1 and θ_2 proportionally to $p_1(T_s(u,v))$ and $p_2(T_s(u,v))$, respectively.

Note that, for a two-source problem, the proposed rule of combination is equivalent to PCR5. The only cell of T_2 involved in the conflict is $T_2(1,1)$ (see figure 18.9 (e)), as a consequence, the mass redistributed to θ_1 and θ_2 is $m'_1 = p_1(T_2(1,1))(T_2^{\theta_1}(1,1)+T_2^{\theta_2}(1,1))$ and $m'_2 = p_2(T_2(1,1))(T_2^{\theta_1}(1,1) + T_2^{\theta_2}(1,1))$, respectively.

$$p_1(T_2^{\theta_1}(1,1)) = \frac{p_1(T_1(0,1))T_1(0,1)+m_2(\theta_1)}{T_1(0,1)+m_2(\theta_1)} = \frac{m_2(\theta_1)}{m_1(\theta_2)+m_2(\theta_1)} \tag{18.26}$$

$$p_2(T_2^{\theta_1}(1,1)) = \frac{p_2(T_1(0,1))T_1(0,1)}{T_1(0,1)+m_2(\theta_1)} = \frac{m_1(\theta_2)}{m_1(\theta_2)+m_2(\theta_1)} \tag{18.27}$$

$$p_1(T_2^{\theta_2}(1,1)) = \frac{p_1(T_1(1,0))T_1(1,0)}{T_1(1,0)+m_2(\theta_2)} = \frac{m_1(\theta_1)}{m_1(\theta_1)+m_2(\theta_2)} \tag{18.28}$$

$$p_2(T_2^{\theta_2}(1,1)) = \frac{p_2(T_1(1,0))T_1(1,0)+m_2(\theta_2)}{T_1(1,0)+m_2(\theta_2)} = \frac{m_2(\theta_2)}{m_1(\theta_1)+m_2(\theta_2)} \tag{18.29}$$

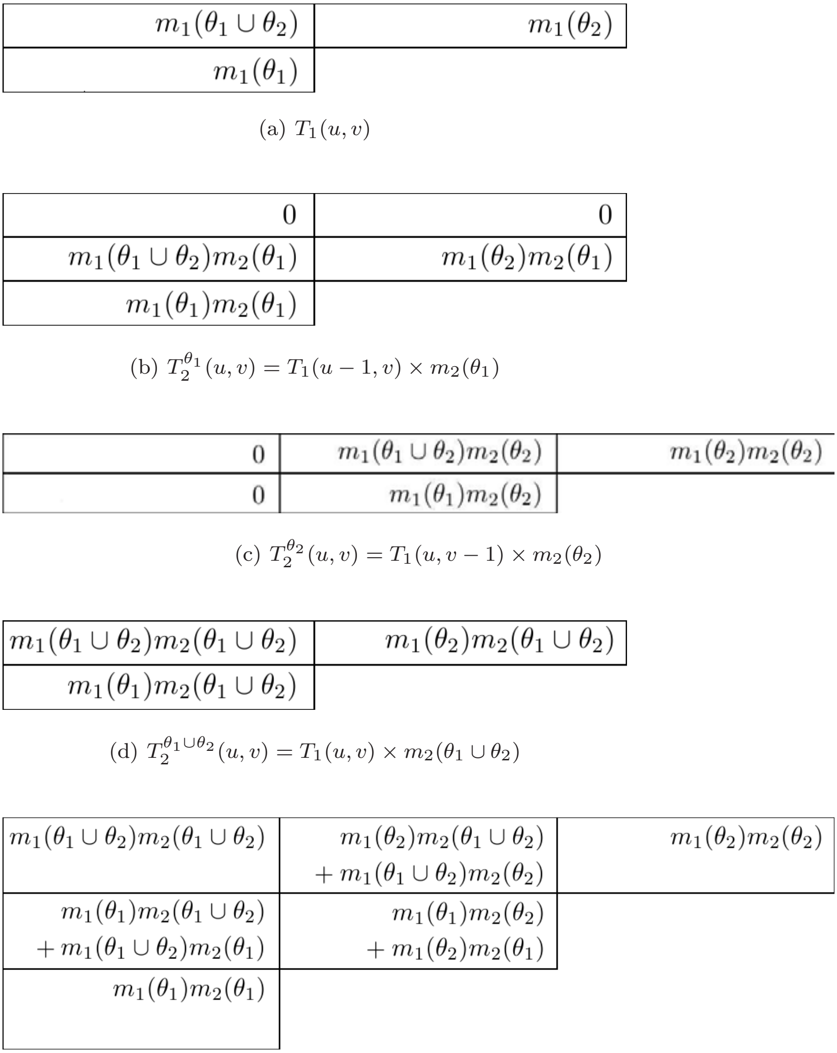
From equation 18.25 (with $p_1(T_1(1,1)) = p_2(T_1(1,1)) = 0$), we obtain the following expression for m'_1 and m'_2:

$$\begin{cases} m'_1 = p_1(T_2^{\theta_1}(1,1))T_2^{\theta_1}(1,1) + p_1(T_2^{\theta_2}(1,1))T_2^{\theta_2}(1,1) \\ m'_2 = p_2(T_2^{\theta_1}(1,1))T_2^{\theta_1}(1,1) + p_2(T_2^{\theta_2}(1,1))T_2^{\theta_2}(1,1) \end{cases} \tag{18.30}$$

$$\begin{cases} m'_1 = \frac{m_2(\theta_1)}{m_1(\theta_2)+m_2(\theta_1)} m_1(\theta_2)m_2(\theta_1) + \frac{m_1(\theta_1)}{m_1(\theta_1)+m_2(\theta_2)} m_1(\theta_1)m_2(\theta_2) \\ m'_2 = \frac{m_1(\theta_2)}{m_1(\theta_2)+m_2(\theta_1)} m_1(\theta_2)m_2(\theta_1) + \frac{m_2(\theta_2)}{m_1(\theta_1)+m_2(\theta_2)} m_1(\theta_1)m_2(\theta_2) \end{cases} \tag{18.31}$$

which is what we obtained for PCR5 (see equation 18.13).
The proposed PCR rule is compared qualitatively with other rules of combination, on a two-source problem supposing hypotheses θ_1 and θ_2 incompatible, in table 18.5.

The number of operations required to compute $p_1(c)$ and $p_2(c)$, for each cell c in T_s, is proportional to the number of operations required to compute T_s. Once p_1 and p_2 have been computed, the number of operations required to redistribute the conflicting mass is proportional to $\frac{s(s-1)}{2}$ (the number of white cells in figure 18.8). As a consequence, the complexity of this algorithm is also polynomial in s. It is thus applicable to a large class of fusion problems: for instance, it is applied to a problem involving 24 sources of evidence in section 18.5.1.



$m_1(\theta_1 \cup \theta_2)$	$m_1(\theta_2)$
$m_1(\theta_1)$	

(a) $T_1(u, v)$

0	0
$m_1(\theta_1 \cup \theta_2)m_2(\theta_1)$	$m_1(\theta_2)m_2(\theta_1)$
$m_1(\theta_1)m_2(\theta_1)$	

(b) $T_2^{\theta_1}(u, v) = T_1(u - 1, v) \times m_2(\theta_1)$

0	$m_1(\theta_1 \cup \theta_2)m_2(\theta_2)$	$m_1(\theta_2)m_2(\theta_2)$
0	$m_1(\theta_1)m_2(\theta_2)$	

(c) $T_2^{\theta_2}(u, v) = T_1(u, v - 1) \times m_2(\theta_2)$

$m_1(\theta_1 \cup \theta_2)m_2(\theta_1 \cup \theta_2)$	$m_1(\theta_2)m_2(\theta_1 \cup \theta_2)$
$m_1(\theta_1)m_2(\theta_1 \cup \theta_2)$	

(d) $T_2^{\theta_1 \cup \theta_2}(u, v) = T_1(u, v) \times m_2(\theta_1 \cup \theta_2)$

$m_1(\theta_1 \cup \theta_2)m_2(\theta_1 \cup \theta_2)$	$m_1(\theta_2)m_2(\theta_1 \cup \theta_2)$ $+ m_1(\theta_1 \cup \theta_2)m_2(\theta_2)$	$m_1(\theta_2)m_2(\theta_2)$
$m_1(\theta_1)m_2(\theta_1 \cup \theta_2)$ $+ m_1(\theta_1 \cup \theta_2)m_2(\theta_1)$	$m_1(\theta_1)m_2(\theta_2)$ $+ m_1(\theta_2)m_2(\theta_1)$	
$m_1(\theta_1)m_2(\theta_1)$		

(e) $T_2(u, v) = T_2^{\theta_1}(u, v) + T_2^{\theta_2}(u, v) + T_2^{\theta_1 \cup \theta_2}(u, v)$

Figure 18.9: Computation of T_2 from T_1.

set	$\theta_1 \cup \theta_2$	θ_1	θ_2	$\theta_1 \cap \theta_2$
m_1	0.7	0.1	0.2	0
m_2	0.3	0.4	0.3	0
conjunctive rule	0.21	0.35	0.33	0.11
Dempster's rule	0.24	0.39	0.37	0
PCR2	0.21	0.405	0.385	0
PCR5	0.21	0.411	0.379	0
proposed PCR rule	0.21	0.411	0.379	0

Table 18.5: Qualitative comparison with other rules of combination.

The memory requirements for the proposed rule of combination are also interesting compared to PCR5: $\left(7\frac{s(s+1)}{2} + 3\frac{(s-1)s}{2}\right) \times 8$ bytes for the proposed method (which corresponds to the cumulated size of matrices T_s, $p_1(T_s)$, $p_2(T_s)$, $p_1(T_s^{\theta_1})$, $p_2(T_s^{\theta_1})$, $p_1(T_s^{\theta_2})$, $p_2(T_s^{\theta_2})$, T_{s-1}, $p_1(T_{s-1})$ and $p_2(T_{s-1})$: the largest amount of memory needed at the same time), compared to $3^s \times 8$ for PCR5, if we use double precision real numbers.

18.8.3 Conclusion

In this appendix, we proposed an algorithm to compute the conjunctive rule in a time evolving polynomially with the number of sources. From this first algorithm, we derived a new Proportional Conflict Redistribution (PCR) rule of combination with a similar complexity. This rule is equivalent to the PCR5 rule for two-source problems (it is also equivalent to PCR6, in this case [12]). We restricted our algorithms to fusion problems consisting of two hypotheses: our goal was to reduce the complexity regarding the number of sources. However, the same principle could be applied to problems consisting of n hypotheses, using n-dimensional matrices.