Multiple Evidence Combination in Image Retrieval: Diogenes Searches for People on the Web*

Y. Alp Aslandogan, Clement T. Yu
aslan,yu@eecs.uic.edu
Department of EECS, University of Illinois at Chicago

Abstract

In this work, we examine evidence combination mechanisms for classifying multimedia information. In particular, we examine linear and Dempster-Shafer methods of evidence combination in the context of identifying personal images on the World Wide Web. An automatic web search engine named **Diogenes**[1] searches the web for personal images and combines different pieces of evidence for identification. The sources of evidence consist of input from face detection/recognition and text/HTML analysis modules. A degree of uncertainty is involved with both of these sources. Diogenes automatically determines the uncertainty locally for each retrieval and uses this information to set a relative significance for each evidence. To our knowledge, Diogenes is the first image search engine using Dempster-Shafer evidence combination based on automatic object recognition and dynamic local uncertainty assessment. In our experiments Diogenes comfortably outperformed some well known commercial and research prototype image search engines for celebrity image queries.

1 Introduction

Dealing with imprecise and uncertain information is a challenging yet very common task. Evidence combination is a powerful tool used for managing uncertainty in fields such as robot vision, remote surveillance, automated equipment monitoring and medical diagnosis. In information retrieval, evidence combination has been used successfully in integrating the evidence of different query representations or different retrieval and ranking strategies [3, 4, 9]. Bayesian statistical models and Fuzzy sets are among the means used by researchers to integrate different pieces of evidence [7]. The ultimate goal of evidence combination is to improve the accuracy of a classifier. Depending on the context, the objects to be classified may be documents, images, landscape or physical objects.

In this work we describe a system called **Diogenes** that classifies facial images from the World Wide Web (the *web* in the sequel) using different methods of evidence combination. To accomplish its goal, Diogenes relies on two pieces of evidence: Visual and textual. Web pages that contain facial images and an accompanying body of text are retrieved and analyzed in parallel by two modules: A face detection/recognition module and an text/HTML analysis module. The face detection/recognition module examines the images on a web page for a facial image and when it finds one attempts to identify the person in the image. This module uses a database of known personal images. The text/HTML analysis module analyzes the body of the text with the aim of finding clues about who appears in each image. The output of the text analysis and face detection/recognition modules are merged using different evidence combination mechanisms to classify each image.

Although search engines for image retrieval on the web exist [14, 15], few of them them take advantage of the full text of the web pages and none performs object recognition. Diogenes analyzes the full text of web pages and integrates this information with the output of a face recognition module using a formal model. The approach is applicable to many other image retrieval and classification tasks where multiple bodies of evidence are available. Key contributions of this paper are the following:

*Research supported in part by NSF grant under ISR-95 08953

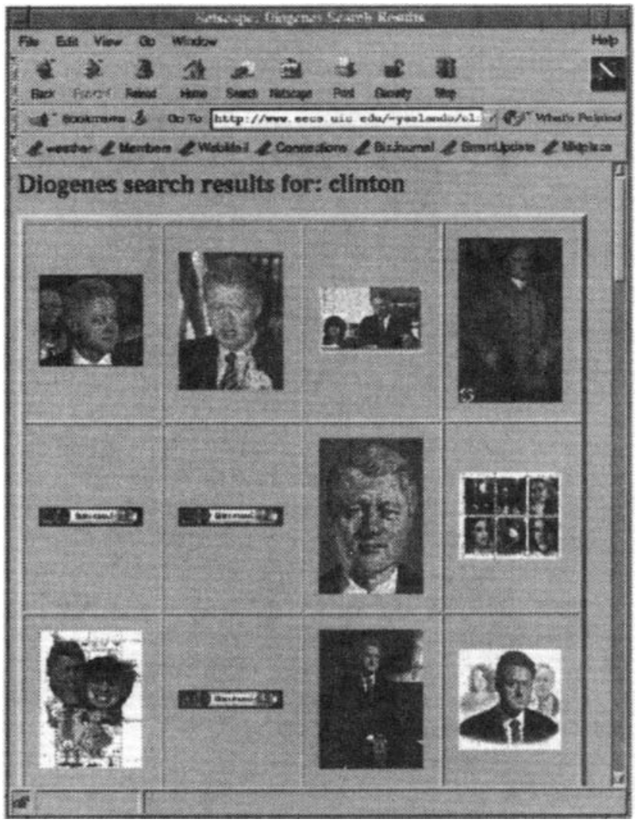
[1] After philosopher Diogenes of Sinope, d.c. 320 B.C. who is said to have gone about Athens with a lantern in day time looking for an honest man.

SIGIR 2000 7/00 Athens, Greece
© 2000 ACM 1-58113-226-3/00/0007...$5.00

- The use of Dempster-Shafer evidence combination method with object recognition and automatic, local uncertainty assessment. Diogenes employs a face recognition module and a text/HTML analysis module. Both of these modules produce numeric values indicating their degrees of uncertainty. These values are obtained *automatically* (without user interaction) and *locally* (separately for each retrieval/classification).

- Experimental, quantitative comparison of Linear and Dempster-Shafer combination methods in the context of personal image retrieval from the web as well as comparisons with the existing commercial and research prototype web image search engines. In comparison to search engines with reasonable recall, Diogenes had significantly better average precision.

In the rest of the paper we first give an overview of the system architecture of Diogenes in section 2. This section also introduces the visual and textual features used by the image classifier. In sections 3 and 4 we describe the Linear and Dempster-Shafer methods for evidence combination respectively and how each is applied to the image retrieval domain. The results of preliminary experiments are given in section 5. We review related work and compare Diogenes to existing commercial and research prototype image search engines in section 6. Section 7 summarizes key points of the paper and explores future directions.



Diogenes search results for: clinton

Figure 1: Diogenes search results for Clinton.

2 Overview

Diogenes is a web-based, automated image search engine designed specifically for personal facial images. It travels the web off-line and builds an index. In response to a query like "retrieve pictures of Bill Clinton" it searches its index and prepares a page containing Bill Clinton images. Figure 1 shows a snapshot of Diogenes search results for "Bill Clinton" query. Diogenes works with the web pages that contain a facial image accompanied by a body of text. The approach is to take advantage of the *full text* and HTML structure of web pages in addition to the visual analysis of the images themselves and to combine the two pieces of information in a formal framework. The system architecture and indexing process of Diogenes are depicted in Figure 2. We describe the visual and text/HTML features used by the classifier further below.

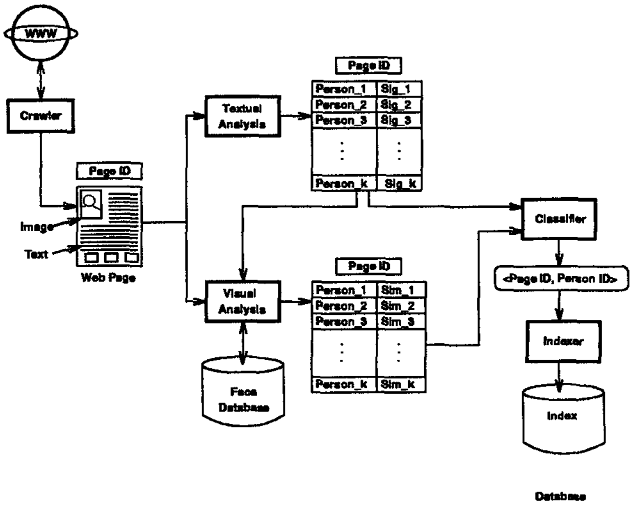


Figure 2: Diogenes system architecture.

2.1 Visual Feature

The visual feature used by the classifier of Diogenes consists of the output of a combined face detection/recognition module. The neural network-based face detection module [11] examines an image to find a human face. If a face is found, the location is indicated to an intermediate module which crops (cuts out) the facial portion and submits it to the face recognition module. Diogenes uses a face recognition module which implements the eigen-face method [18]. This module uses a set of known facial images for training. Each of these training images has an associated personal name with it. At recognition time, a set of distance values between the input image and those of the training set are reported. These distances indicate how dissimilar the input image is to the training images. In addition, a global distance value called "Distance From Face Space" or DFFS is also reported. This is the global distance of the input image from the facial image space spanned by the training images. Diogenes uses this latter value as an indicator of the accuracy of the recognition.

2.2 Text/HTML Feature

The text/HTML analysis module of Diogenes determines a degree of association between each personal name on a web page and each facial image on that page. This degree of association is based on two factors: Page-level statistics and local (or structural) statistics. Page-level statistics such as frequency of occurrence [12, 19] and location-within-the-page (title, keyword, body text etc.) are independent of any particular image. Local/structural statistics are those factors that relate a name to an image. In an earlier work [2], we have shown that structured queries and image descriptions provide a better framework for matching different descriptions of the same phenomenon as opposed to free text descriptions. Diogenes takes advantage of the HTML structure of a web page in determining the degree of association between a personal name and an image. The factors of interest include whether the name occurs in proximity of the image, whether they are enclosed in the same HTML tag, whether the name is part of the image URL/path.

Once the visual and text/HTML features are computed for a particular web page, it is the task of the classifier to combine them. In the following we examine two approaches for evidence combination: Linear and Dempster-Shafer. These approaches have been implemented and evaluated experimentally.

3 Linear Evidence Combination

The first method of evidence combination we study is the linear combination. A simple linear combination is a weighted sum of normalized individual features. Either the feature values (similarities) themselves or the ranks can be combined[9]:

$$Rank_{final} = \omega_1 * Rank_{FR} + \omega_2 * Rank_{TA}$$

where ω_1 and ω_2 are weights, $Rank_{FR}$ is the rank of a person as determined by the face recognition module and $Rank_{TA}$ is the rank of a person as determined by the text/HTML analysis module. Diogenes implements a feature-value combination scheme:

$$Score_{combined} = \omega_1 * Score_{FR} + \omega_2 * Score_{TA}$$

Where $Score_{FR}$ and $Score_{TA}$ are the numeric "degree of association" scores assigned to each pair of personal name and facial image on a web page.

The simplest approach to assigning weights to classifier inputs is to use constant weights. A more sophisticated approach might improve these weights by various learning algorithms. A third approach is to use the confidence of the values supplied to the classifier by the textual and visual analysis modules. If facial recognition is not confident of its similarity values (relatively low similarities) then relying heavily on face recognition is not a good strategy. If on the other hand face recognition is confident, meaning very high similarity for at least one image, than it can be assigned a higher weight. This approach results in the following combination formula

$$Score_{combined} = conf_{FR} * Score_{FR} + conf_{TA} * Score_{TA}$$

where $conf_{FR}$ and $conf_{TA}$ are the *confidence* values for face recognition and text/HTML analysis respectively.

4 Dempster-Shafer Evidence Combination

Dempster-Shafer Theory of Evidence (a.k.a. Mathematical Theory Of Evidence) is intended to be a generalization of Bayesian theory of subjective probability [13]. It introduces concepts such as *frame of discernment, basic probability assignments, plausibility,* and *confidence interval* to incorporate uncertainty into probability calculations. It also provides a method for combining independent bodies of evidence using Dempster's rule.

A key concept in Dempster-Shafer theory is the *frame of discernment*. A frame of discernment Θ is an exhaustive set of mutually exclusive hypotheses or propositions. The set of all subsets of Θ (the power-set of Θ) is denoted by 2^{Θ}. All of the elements in this power-set, including the elements of Θ, are propositions.

Each hypothesis or proposition is assigned a degree of belief supported by a piece of evidence. This degree of belief, also named a ***basic probability assignment*** or ***mass function*** is denoted by m and has the following property:

$$m(\emptyset) = 0 \textit{ and } \sum_{A \subseteq \Theta} m(A) = 1.$$

In this definition, A is any element of 2^{Θ}. The quantity $m(A)$ is a measure of that portion of the total belief committed exactly to A. $m(A)$ is atomic in the sense that it can not be further subdivided among subsets of A and does not include portions of belief committed to subsets of A. It represents the likelihood information we have for A and A alone. This information is not obtained by combining other beliefs or by inference. The quantity $m(\Theta)$ is a measure of that portion of the total that remains uncommitted after commitment of belief to various proper subsets of Θ. It accounts for our "ignorance" about those subsets of Θ for which we have no specific belief. Thus, the basic probability assignment to Θ represents the *uncertainty* of the evidence:

$$m(\Theta) = 1 - \sum_{A \subset \Theta} m(A)$$

If in a body of evidence the basic probabilities assigned to proper subsets of Θ add up to 1, then this would make $m(\Theta) = 0$ meaning we have very high confidence in this body of evidence and no uncertainty.

Example 4.1 Consider a set of three persons for whom we have a database of images: *Clinton, Gingrich,* and *Dole*. For a query image I_Q, we are interested in knowing which person it belongs to. So, we may form the following propositions which correspond to proper subsets of Θ:

P_C : *I_Q belongs to Clinton.*
P_G : *I_Q belongs to Gingrich.*
P_D : *I_Q belongs to Dole.*
P_C, P_G : *I_Q belongs to either Clinton or Gingrich.*
P_C, P_D : *I_Q belongs to either Clinton or Dole.*
P_G, P_G : *I_Q belongs to either Gingrich or Dole.*

With these definitions, the 2^{Θ} would consist of the following:

$$2^{\Theta} \;=\; \{\{P_C\},\{P_G\},\{P_D\},\{P_C,P_G\},\{P_G,P_D\}, \{P_C,P_D\},\{P_C,P_G,P_D\},\emptyset\}$$

In many applications basic probabilities for every proper subset of Θ may not be available. In these cases a non-zero $m(\Theta)$ accounts for all those subsets for which we have no specific belief. In a typical image classification setup, assuming one person per image, we have positive evidence for individual persons only

$$m(A) > 0 : A \in \{\{P_C\},\{P_G\},\{P_D\}\}$$

and possibly a degree of uncertainty is associated with the body of evidence, $m(\Theta)$. This means, for the remaining propositions other than Θ itself, we do not have a specific belief, hence zero basic probability:

$$m(A) = 0 : A \in \{\{P_C,P_G\},\{P_G,P_D\},\{P_C,P_D\},\emptyset\}$$

The belief assigned to the whole body of evidence, $m(\Theta)$, accounts for these subsets. Hence

$$m(\Theta) = 1 - \sum_{A \in \{P_C,P_G,P_D\}} m(A)$$

It should be noted that, unlike probability theory, the ***basic probability assignments*** to sets in a frame of discernment is not monotonically increasing with respect to the subset relationship. In other words, a set does not necessarily have a larger basic probability than its subsets. This is due to the definition of the Dempster-Shafer ***basic probability*** given above.

4.1 Dempster's Rule for Evidence Combination

Suppose we are interested in finding the combined evidence for a hypothesis C. We may think of C as a class assignment in pattern recognition. C is a member of 2^{Θ}, where Θ is our *frame of discernment.* Given two independent sources of evidence m_1 and m_2, Dempster's rule for their combination is as follows:

$$m_{1,2}(C) = \frac{\sum_{A,B\subseteq\Theta, A\cap B=C} m_1(A)m_2(B)}{\sum_{A,B\subseteq\Theta, A\cap B\neq\emptyset} m_1(A)m_2(B)}$$

Here $m_{1,2}(C)$ is the combined Dempster-Shafer probability for C. m_1 and m_2 are the basic probabilities assigned to sets A and B respectively by two independent sources of evidence. A and B are supersets of C. A and B are not necessarily proper supersets and they may as well be equal to C or to the frame of discernment Θ.

The numerator accumulates the evidence which supports a particular hypothesis and the denominator conditions it on the total evidence for those hypotheses supported by both sources.

4.2 Using Dempster-Shafer Theory in Image Retrieval

In this study, we are interested in classifying personal images obtained from the web. We have two sources of evidence: The output of a face recognition module (FR) which classifies the image and the output of a text/HTML analysis module (TA) which analyzes the text that accompanies the image. Both modules attempt to identify the person in the image based on different media. We assume that if more than one person appears in an image, identifying one of them is sufficient. We designate the two pieces of evidence as m_{FR} and m_{TA} respectively. By default, these two modules operate independently: The results of face recognition module does not affect the text/HTML score and vice versa. Hence the independence assumption of the theory holds. The text/HTML analysis module determines a degree of association between every personal name-facial image pair on the web page. It assigns numerical values to different personal names for each image indicating this degree of association. The set of personal names for which text/HTML analysis module have degrees of association forms the frame of discernment Θ_{TA} for text/HTML analysis. The face recognition module assigns a distance value to each person in our database of known personal images. We convert these values to similarity values. In order to have the same frame of discernment for the two modules, we limit the face recognition database to those persons whose names appear on the web page. Furthermore, we assume that for any person for which we have no stored image, the face recognition similarity value is zero. If we use Dempster's Rule for combination of evidence we get the following:

$$m_{FR,TA}(C) = \frac{\sum_{A,B\subseteq\Theta,A\cap B=C} m_{FR}(A)m_{TA}(B)}{\sum_{A,B\subseteq\Theta,A\cap B\neq\emptyset} m_{FR}(A)m_{TA}(B)}$$

Again, C designates a hypothesis which is an element of 2^{Θ}. In the case of classification of personal images, it is possible to simplify this formulation. Our face recognition and text/HTML analysis modules give us information about individual images and the uncertainty of the recognition/analysis. This means we have only beliefs for singleton classes (persons) and the body of evidence itself ($m(\Theta)$). The latter is also the uncertainty in the body of the evidence as described in section 4. This means for any proper subset A of Θ for which we have no specific belief, $m(A) = 0$.

To illustrate this simplification on Example 4.1, consider first the set of all propositions in 2^{Θ}:

$$\begin{aligned} 2^{\Theta} &= \{\{P_C\},\{P_G\},\{P_D\},\{P_C,P_G\},\{P_G,P_D\},\\ &\quad \{P_C,P_D\},\{P_C,P_G,P_D\},\emptyset\} \end{aligned}$$

Corresponding to combinations of pairs of evidence for the elements of 2^{Θ} we would have the following terms in the numerator of Dempster's combination:

$$\begin{gathered} m_1(P_C)m_2(P_C), m_1(P_C)m_2(\{P_C,P_G\}),\\ m_1(P_C)m_2(\{P_C,P_D\}), m_1(P_C)m_2(\Theta),\\ m_1(\{P_C,P_G\})m_2(P_C), m_1(\{P_C,P_D\})m_2(P_C),\\ m_1(\Theta)m_2(P_C). \end{gathered}$$

We have non-zero basic probability assignments for only the singleton subsets of Θ and the Θ itself. So terms like $m_1(P_C)m_2(\{P_C,P_G\})$ evaluate to zero because of one of the components of the term, $m_2(\{P_C,P_G\})$, is zero.

Since we are interested in the ranking of the hypotheses and the denominator is independent of any particular hypothesis (i.e. same for all) we can further simply as follows:

$$\begin{aligned} rank(P_C) \quad \propto \quad & m_{FR}(P_C)m_{TA}(P_C) + m_{FR}(\Theta)m_{TA}(P_C)\\ & + m_{FR}(P_C)m_{TA}(\Theta) \end{aligned}$$

Here $\propto$ represents ‘is proportional to" relationship, $m_{FR}(\Theta)$ and $m_{TA}(\Theta)$ represent the uncertainty in the bodies of evidence m_{FR} and m_{TA} respectively. These are obtained as follows: For face recognition, we have a "distance from face space" (DFFS) value for each recognition. This value is the distance of the query image to the space of eigen-faces formed from the training images [18]. Diogenes uses the DFFS value to estimate the uncertainty associated with face recognition. If the DFFS value is small, the recognition is good (uncertainty is low) and vice versa. The following is Diogenes' formula for the uncertainty in face recognition:

$$m_{FR}(\Theta) = 1 - (\frac{1}{ln(e + DFFS)})$$

For text analysis, uncertainty is inversely proportional to the maximum value among the set of degree of association values assigned to name-image combinations.

$$m_{TA}(\Theta) = \frac{1}{ln(e + MDA)}$$

Where MDA is the maximum numeric "degree of association" value assigned to a personal name with respect to a facial image among other names. As described in section 2.2, each degree of association for an image name pair is a function of the frequency of occurrence of that name, location relative to the image, HTML tags shared with the image, etc.

Both face recognition and text analysis uncertainties are obtained locally, i.e. for each retrieval and automatically without user interaction. This feature distinguishes Diogenes from other applications where the users provide the uncertainties [8].

5 Experimental Results

For a quantitative evaluation ten celebrity image queries were submitted to Diogenes and three other web image search engines. Diogenes retrieved and analyzed an average of 1500 web pages per query. The following table compares the search results of Diogenes with those of three other image search engines.[2] For each search engine, there are two numbers. The first number shows the precision: the number of relevant images for the top 20 retrieved images. The second is the total number of images returned for the query. The number 20 was chosen based on the observation that the users typically do not browse beyond the top two pages of results and a typical results page contains 10 images. If the total number is less than 20, then the precision is computed over the retrieved total.

Query	WS	AV	LY	DG
B.Clinton	.90/10	.45/7596	.75/320	.90/98
H.Clinton	.55/9	.95/2423	.75/80	.95/24
K.Starr	.67/3	.75/1301	0.0/11	.70/34
M.Lewinsky	n/a	.50/1168	.95/47	1.0/84
M.Albright	1.0/5	.60/194	.40/5	.70/48
B.Gates	n/a	.60/9956	.60/245	.65/35
B.Netanyahu	1.0/2	.80/1022	.70/9	1.0/17
B.Yeltsin	.55/13	.55/1026	.85/12	.61/18
R.Limbaugh	n/a	.45/1577	.70/70	.80/52
OJ.Simpson	n/a	.30/94	.60/21	.65/57
Average	**.78**	**.60**	**.63**	**.80**

Table 5.1 WebSEEK (WS), AltaVista (AV), Lycos (LY), and Diogenes (DG) search results for celebrity queries.

Although Diogenes had only a fraction of the time available to the other search engines to produce the results reported here, it was able to outperform some of those in terms of total recall as well as precision. In particular, total recall for Diogenes was better than Lycos in queries 3,4,7,8, and 10. Diogenes had the best average precision of the four. Although the average precision for WebSeek was close to that of Diogenes, its poor total recall diminishes its value for practical retrieval purposes. The average precision for WebSeek was computed over 6 queries as it did not return any results for queries 4 and 10 and the web interface was not available for queries 6 and 9 at the time of this writing. The average precision of Diogenes was significantly higher than those of AltaVista and Lycos[3].

The following table shows how Dempster-Shafer evidence combination approach fares with other retrieval methods. The number in each cell shows the average precision over the top 20 retrieved images for a particular query. The first row shows the retrieval results based on Face Detection and text/HTML analysis. The second row shows the results when the search engine relied on face detection and recognition. The third row is for linear evidence combination as described in section 3 and the fourth row is for Dempster-Shafer evidence combination approach.

Table 5.2 below shows how the different combinations implemented by Diogenes fare against each other.

Query	FD/TH	FR	Lin.	DS
B. Clinton	.75	.25	.55	.90
H. Clinton	.95	.40	.95	.95
K. Starr	.20	.10	.20	.70
M. Lewinsky	1.0	.60	1.0	1.0
M. Albright	.85	.33	.60	.70
B. Gates	.50	1.0	.45	.65
B. Netanyahu	1.0	1.0	1.0	1.0
B. Yeltsin	.52	.50	.55	.61
R. Limbaugh	.90	.55	.85	.80
OJ Simpson	.50	.40	.55	.55
Average	**.72**	**.51**	**.67**	**.80**

Table 5.2 Performance comparison of search strategies implemented by Diogenes. Legend: FD/TH: Face Detection followed by Text/HTML analysis; FR: Face Recognition; Lin: Linear combination; DS: Dempster-Shafer combination.

As can be seen from the table the Dempster-Shafer combination produced better results than any individual method or the linear combination method described in section 3. The total recall of Dempster-Shafer combination was the highest or very close to the highest total for each retrieval. The images retrieved by Diogenes for some of the above queries can be viewed at http://www.eecs.uic.edu/~yaslando/diogenes/.

6 Related Work and Discussion

A number of web image search engines have been built in recent years including both research prototypes and commercial ones. Among the former category are WebSeer [15], WebSEEk [14],

[2] WebSeer project was discontinued and hence the web interface was not available for this evaluation.

[3] For AltaVista image retrievals, the proprietary images collections such as Corbis^{TM} were excluded. Although using the proprietary image collections yields considerably better average precision than what is reported here, we restricted the AltaVista search domain to the web since our study is focused on retrieval from the web

ImageScape (http://ind134a.wi.leidenuniv.nl:2001/), Amore (http://www.ccrl.com/amore/), WebHunter [10], ImageRover [17], and PicToSeek [5]. Sometimes the techniques developed by earlier research were incorporated into web-based image search systems [1, 6, 16]. Commercial web text search engines such as Lycos and AltaVista also offer image search facilities.

We will review some of these systems that are most similar to Diogenes in terms of functionality and scope. We will consider four factors that are most relevant to our discussion: a) Whether they use *reference text* only or full text of web pages; b) whether they perform some form of visual analysis on the images; c) whether and how they integrate the textual and visual cues; d) their recall and precision.

The *reference text* is the piece of text enclosed in special HTML tags that contains the name or the URL (the web address) of the image, possibly a comment and an alternative text which appears in place of the image on a text-only web browser. In order to index images, both WebSEEk and to a large extent WebSeer rely on the words found in the image *reference* text. **Diogenes** on the other hand uses both the reference text and the *full text* of the web page.

Both WebSEEk and WebSeer analyze the images they retrieve from the web visually, but for different purposes. WebSEEk emphasizes the color information in images and allows for visual query by example. WebSeer on the other hand uses several visual criteria to classify images as graphs, cartoons, photographs, clip art, etc. ImageScape uses a text search engine to retrieve an initial set of pages and then uses visual similarity to retrieve further similar pages. A third, commercial image search engine, Lycos (http://www.lycos.com/picturethis/), does not perform any image analysis, but relies solely on the *reference* text.

WebSEEk uses reference text to classify images into categories such as "Animals", "Architecture", "Celebrities" etc. WebSeer goes one step further in image analysis by integrating a face detector. Hence, only images that contain a human face are returned in response to people queries. WebSEEk only allows single word queries and does not perform any face detection. The accuracy therefore is much lower than WebSeer in people queries. Lycos has the highest recall and lowest precision of the three. Since apparently it does not do any image analysis, the images returned may be totally unrelated to each other. Lycos does allow multiple word queries.

A portrait image search engine for the web, WebHunter, is developed by Munkelt et al. [10]. This system is similar to WebSeer in that both systems employ a module to test for the presence of a person in an image but they rely on the text to determine the owner of the image. Neither system uses face recognition. **Diogenes** on the other hand integrates the results of textual analysis and face recognition with a formal model to classify the images. The differences between **Diogenes** and some of the other search engines are summarized in the following table:

S. Engine	R.Text	F. Text	FD	FR	FI
Diogenes	Yes	Yes	Yes	Yes	Yes
WebSEEk	Yes	No	No	No	n/a
WebSeer	Yes	?	Yes	No	No
ImageScape	Yes	?	No	No	n/a
WebHunter	Yes	Yes	Yes	No	No
AltaVista	Yes	Yes	No	No	n/a

Table 5.3 Feature comparison of various search engines with Diogenes. Legend: R.Text: Reference Text; F. Text: Full Text; FD: Face Detection; FR: Face Recognition; FI: Formal Integration.

The simplified Dempster-Shafer evidence combination formula for image retrieval was originally described by Jose et al. in [8]. In this work, a (nonfacial) photograph retrieval system, called **Epic**, that uses Dempster-Shafer evidence combination is reported. The primary difference between Epic and Diogenes is how they obtain and use the uncertainty values: Epic lets its users input their confidence in the bodies of evidence prior to retrieval. Diogenes on the other hand uses a face recognition module and automatically determines uncertainty values for both text/HTML and face recognition based on system performance parameters. Although in certain cases the users may have some knowledge about the reliability of their information sources, in many cases including web image retrieval this is not a valid assumption. Furthermore, the information sources may not be equally reliable across retrieval sessions. In this case the user has to provide the confidence for each session or ignore the inaccuracy of having a single confidence value across sessions. Since Diogenes determines uncertainties for each retrieval session automatically, it is not prone to this problem.

7 Conclusion and Future Work

In this work we have discussed the design and implementation of Diogenes, a personal image search engine for the WWW. The novel feature of this search engine is its integration of text/HTML analysis with object recognition in a formal evidence combination framework. Combining evidence facilitates improvement over any individual evidence source as confirmed by our experimental evaluation. The main contribution of this work is to show the use of Dempster-Shafer evidence combination with object recognition and automatic local uncertainty assessment.

There are a number of other evidence combination strategies which we haven't experimented with in this work. These include Bayesian combination, neural networks and fuzzy sets. We plan to evaluate some of these approaches and compare with the Dempster-Shafer method experimentally.

The techniques discussed and illustrated in this work are applicable to a number of other problems where multiple sources of evidence are available with different degrees of uncertainty.

Acknowledgments

We would like to thank Matthew Turk (Microsoft), Alex Pentland (MIT), Takeo Kanade (CMU) and Henry A. Rowley(CMU) for making software available for this project.

References

[1] B. Agnew, C. Faloutsos, Z. Wang, D. Welch, and X. Xue. Multi-media Indexing Over the Web. In *Proceedings of SPIE Vol. 3022*, pages 72–83, 1997.

[2] Y. Alp Aslandogan, Charles Thier, Clement T. Yu, Jun Zou, and Naphtali Rishe. Using Semantic Contents and WordNet(TM) in Image Retrieval. In *Proceedings of ACM SIGIR Conference*, Philadelphia, PA, 1997.

[3] Nicholas J. Belkin, P. Kantor, Colleen Cool, and R. Quatrain. Combining Evidence for Information Retrieval. In *Proceedings of TREC 1993*, pages 35–44, 1993.

[4] Nicholas J. Belkin, Paul B. Kantor, Edward A. Fox, and J. A. Shaw. Combining the Evidence of Multiple Query Representations for Information Retrieval. *Information Processing and Management*, 31(3):431–448, 1995.

[5] Theo Gevers and Arnold W. M. Smeulders. PicToSeek: A Content-Based Image Search System for the World Wide Web. In *Proceedings of SPIE Visual 97*, 1997.

[6] Amarnath Gupta and Ramesh Jain. Visual Information Retrieval. *Communications of the ACM*, 40(5):69–79, 1997.

[7] David L. Hall. *Mathematical Techniques in Multisensor Data Fusion*. Artech House, 1992.

[8] Joemon M. Jose, Jonathan Furner, and David J. Harper. Spatial Querying for Image Retrieval: A User Oriented Evaluation. In *ACM SIGIR*, pages 232–240, 1998.

[9] Joon Ho Lee. Analyses of Multiple Evidence Combination. In *Proceedings of ACM SIGIR*, pages 267–275, 1997.

[10] Olaf Munkelt, Oliver Kaufmann, and Wolfgang Eckstein. Content-based Image Retrieval in the World Wide Web: A Web Agent for Fetching Portraits. In *Proceedings of SPIE Vol. 3022*, pages 408–416, 1997.

[11] Henry A. Rowley, Shumeet Baluja, and Takeo Kanade. Neural Network-Based Face Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38, Jan 1998.

[12] Salton, G. *Automatic Text Processing*. Addison Wesley, Mass., 1989.

[13] Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[14] J. R. Smith and S. F. Chang. a Content-based Image and Video Search Engine for the World-Wide Web. *IEEE Multimedia*, Summer 1997.

[15] Michael J. Swain, Charles Frankel, and Vassilis Athitsos. WebSeer: An Image Search Engine for the World Wide Web. Technical Report TR-96-14, University of Chicago, Department of Computer Science, July 1996.

[16] L. C. Tai and R. C. Jain. ImageGREP: Fast Visual Pattern Matching in Image Databases. In *Proceedings of SPIE Conference on Storage and Retrieval for Image and Video Databases*, San Diego, CA, February 1997.

[17] Leonid Taycher, Marco LaCascia, and Stan Sclaroff. Image Digestion and Relevance Feedback in the ImageRover WWW Search Engine. In *Proceedings of SPIE Visual 97*, 1997.

[18] M. Turk and A. Pentland. Eigenfaces for Recognition. *Cognitive Neuroscience*, 3(1):71–86, 1991.

[19] Clement T. Yu and Weiyi Meng. *Principles of Database Query Processing for Advanced Applications*. Data Management Systems. Morgan Kaufmann, 1998.