—蓄える・探す—

大規模に蓄積された画像および音声を対象とする情報検索

Information Retrieval for Large-scale Image and Audio Archive

廣池 敦　Hiroike Atsushi
大淵 康成　Obuchi Yasunari

日立製作所は，大規模に蓄積された画像および音声情報を対象とする検索技術の研究開発を，長年にわたり進めている。画像検索に関しては，類似画像検索システム「EnraEnra（エンラエンラ）」，およびそれを構成する要素技術である高速ベクトル検索，画像中からの対象物の検出，画像からの言語情報の推定などがある。また，音声検索に関しては，音声中の検索語の検出技術，話者識別，感情認識などがある。

1. はじめに

現代社会では，映像や画像コンテンツ，音声，文書などのさまざまな情報が大量に生み出されている。大規模な蓄積データを扱うためには，データ中から効率よく必要な情報を抽出するための検索技術が必要である。

ここでは，非構造データに対応した検索技術として，画像検索技術と音声検索技術における日立の取り組みについて述べる。

2. 画像検索技術

2.1 類似画像検索

類似画像検索とは，ある与えられた画像と「見た目が似ている画像」を，対象となる画像集合中から探し出す技術である。基本的な方式は，まず，画像に含まれる情報を数値ベクトルとして表現した「画像特徴量」を画像から抽出する。これによって，各画像は，特徴量が張るベクトル空間中の一点として表現される。ある問い合わせ画像が与えられると，このベクトル空間中で近傍に存在するデータを探索し，検索結果として示すという仕組みである。

画像特徴量としては種々のものが考えられるが，日立独自のシステムでは，色分布特徴量，エッジパターン特徴量（局所的な輝度変化パターンの分布を算出したもの）などを用いている。検索時には，数千次元程度に及ぶこれらの特徴量を，統計的次元圧縮によって数百次元程度に変換して用いている。

一般論から言えば，画像特徴量とは，画像の情報すべてを表現することを目的とするものではない。例えば「色」は，画像が持つ重要な情報ではあるが，検索目的によっては色彩の差異を無視したほうが適切となる場合も多い。したがって，画像特徴量は，着目する性質に関してはデータ間の差異を強調し，そうでない性質に関しては同一視するように構成する必要がある。

日立製作所では，1997年ごろ，データベース製品の新規機能搭載を目的として，類似画像検索の技術開発に着手し，当初開発された技術は，「HiRDB Image Search Plug-in（特徴量検索プラグイン）」として製品化された。

当時，インターネットの急速な普及などを背景に，日立製作所を含め，幾つかの企業が「データベースシステムのマルチメディア化」という観点からこの種の技術の製品化に取り組んでいたが，あまり一般的な技術としては広がらなかった。1990年代後半，画像・映像データの標準化などは進んでいたが，通信やストレージの制約もあり，必ずしも大規模な電子データが流通・蓄積されていたわけではなかった。そもそも検索技術とは，データ量が人間に扱えない規模になってこそ意義を持つものであり，実際に大規模なデータが存在している今だからこそ，このような検索技術へのニーズが現実化していると言える。

現状の情報検索は，「言葉」を用いたものが中心であり，おそらく，将来もそうであろう。ただし，われわれを取り巻く情報は，例えば他人に情報を伝える際に，グラフを用いたり，図を描いて説明したりするように言語情報だけではない。

このことからも明らかなように，実際には，非常に多くの言語化が困難な情報，あるいは非言語情報としてしか存

在しえない情報が存在する。このような従来の検索技術では取りこぼされてきた情報に基づく検索機能を提供するのが，類似画像検索である。

2.2　類似画像検索技術「EnraEnra」

日立製作所中央研究所では，1997年から2001年まで，経済産業省（旧通商産業省）の国家プロジェクト「技術研究組合新情報処理開発機構（RWCP：Real World Computing Partnership）」に参画し，主として，検索結果のユーザーへの提示方式に関する研究開発を行った。当時の代表的な成果は，類似画像検索の結果を，画像特徴量空間から構成された三次元可視化空間中を「動き回る画像の群れ」として表現するモデルである[1), 2)]（図1参照）。

一方，検索エンジン側に関しては，類似画像検索という特殊な機能を効率的に実現するために，独自のアーキテクチャおよびAPI（Application Programming Interface）を開発した。まず，画像特徴量の定義に必要となる，多様なパラメータ設定，アルゴリズムの選択・組み合わせ，新規アルゴリズムの実装などを柔軟に実現するためのスクリプト言語型APIを実装した。この機能の重要な点は，特徴量定義の柔軟性と，大量画像を想定した処理の高速性を両立させたところにある。

画像特徴量を抽出した後は，これを管理し検索するシステムが必要となる。われわれは，効率的な類似検索を実現するために，独自アーキテクチャのデータベースシステムを開発した。

当初のシステムは，類似検索機能だけがサポートされた単純なものであったが，実際の運用では，画像特徴量以外のメタ情報を扱う必要性が発生する。現在のシステムでは，テキスト情報，特徴量以外の各種数値情報などの管理が可能であり，それらを用いた検索機能もサポートされている。

このように，日立の類似画像検索システムは，収集系，配信系などの周辺技術も取り込みながら進化してきた。これらの技術の総体が「EnraEnra」である。

ソフトウェアライブラリとしてのEnraEnraは，Java※）層とネイティブ層の2層構造から構成される。特徴量抽出，単一テーブルを対象とした管理・検索などの基本機能は，ネイティブライブラリとして実装され，Javaのクラスにラップされた形でJava側に提供される。一方，通信関係，データベース上の複数テーブル間の連携，並列分散処理などの機能は，Java層で実装されている。

このような構造をとることにより，アプリケーション開発の効率性と処理の高速性の両立を実現している。

2.3　高速類似ベクトル検索

厳密に類似ベクトルを検索するためには，与えられた問い合わせベクトルと，検索対象となるデータベース中の全ベクトルとの距離計算を行い，その大小を比較する必要がある。これには，データ件数に比例した計算量を要する。そこで，検索処理の高速化のための近似アルゴリズムとして，クラスタリングを用いた方式を開発した[3)]。

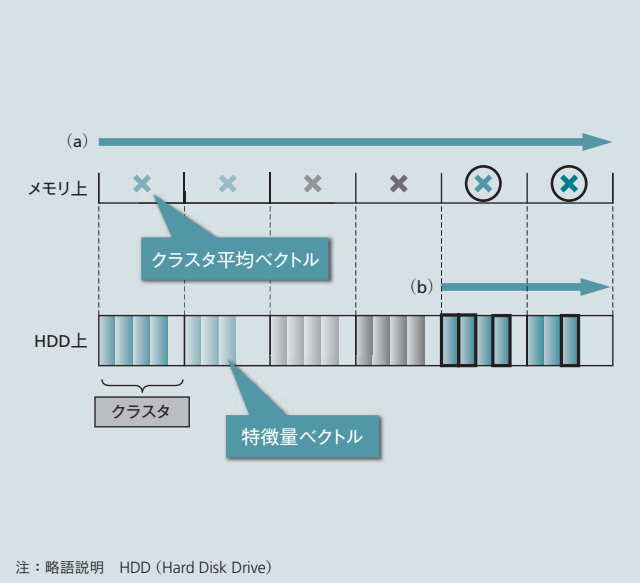
一般に，クラスタリングとは，類似したベクトルをクラスタという単位にまとめることを言う。各クラスタの代表として，各クラスタに含まれるメンバの平均ベクトルを保持し，それとの比較を行うことによって，類似したベクトルが含まれる可能性が高いクラスタを選別することができる。

通常のクラスタリングでは，あらかじめ分類するべきクラスタの個数を設定し，全データを振り分けるような処理を行う。これに対して，われわれの手法では，クラスタに入るメンバ数のほうに上限値を設定する。データ登録時には，登録データと最も類似したクラスタを見つけ，そこに登録しようとするが，その際にそのクラスタのメンバ数があふれた場合は新たにクラスタを追加していく。

EnraEnraにおけるクラスタリング処理で重要な点の一つに，データの管理方式がある。通常，この種の処理では，クラスタとそのメンバとを対応づける索引情報だけを管理する。これに対して，われわれのシステムでは，クラスタのメンバである各データの特徴量ベクトルをクラスタ単位でHDD（Hard Disk Drive）上に格納する。これによって，

図1｜類似画像検索結果の可視化表現
類似画像検索の結果を，画像特徴量から構成された可視化空間中を「動き回る画像の群れ」として表現する。

※）Javaは，Oracle Corporation およびその子会社，関連会社の米国およびその他の国における登録商標である。



注：略語説明　HDD（Hard Disk Drive）

図2｜クラスタリングされた特徴量ベクトルの管理方式と検索

各データの特徴量ベクトルはクラスタの単位でHDD上に格納される。検索時には，メモリ上に展開されたクラスタ平均ベクトルを走査することによって，複数個の類似クラスタを取得し（a），類似クラスタ内の特徴量ベクトルを走査することによって，最終的な検索結果を得る（b）。

検索時には，効率的に「類似したデータをまとめて」アクセスすることが可能となる（**図2**参照）。

この高速検索技術の実現により，1万時間の放送映像の検索や，100台のカメラを接続した映像監視，1億件のWeb画像の検索といった大規模データを対象としたプロトタイプの構築が可能となった。また，例えば映像監視向けの応用では，検索よりも，むしろ更新系の高速性が重要となる。EnraEnraの検索技術は，単に検索が速いというだけではなく，リアルタイム監視などのアプリケーションにも対応可能な，高速のデータベース更新が可能なことも特長である。

2.4　類似顔検索と一般オブジェクト検出

漠然と画像全体の類似性を評価しただけでは，実際のユーザーのニーズに対応できない場合が多い。そこで必要となるのが，画像中から着目すべき事物が含まれる部分領域を取り出す「検出技術」である。

検出技術として一般的なのは，デジタルカメラなどにも搭載されている正面顔の検出である。われわれは，放送映像を対象とした検索で，顔検出によって取り出された顔画像に対して類似検索を行うシステムを構築した。デモンストレーションとしては，例えば，インターネット上で見つけた俳優や政治家などの著名人の写真を手がかりに，その人物が登場するシーンを検索し，そこから映像を再生する，といったことを実現した。

画像に含まれる事物は，もちろん顔以外にも多様なものが存在する。ただし，一般事物の検出は，画像認識の分野では，いまだに難易度が高い課題と考えられている。われわれは，類似検索技術を用いた「一般オブジェクト検出」技術を提案している[4)]。

アプローチは，非常に単純である。検出対象の事例となる画像の集合を用意し，いずれかの事例画像と類似性が高い画像中の部分矩形（くけい）領域を探し出す。実際には，事例画像の集合を検索対象，検出領域の候補となる部分画像を問い合わせ画像とする類似検索を行うことになる。当然ながら，一つの画像の中に，位置，大きさ，縦横比が異なる膨大な数の検出領域候補が想定される。したがって，単に検索が速いだけでは不十分である。そこで，領域候補の選択と絞り込みなどを効率的に行うアルゴリズムを開発することによって，実用的な検出技術としてこの手法を実装した。

2.5　Web画像を用いた画像アノテーション

アノテーションとは，あるデータに対して関連する情報（メタデータ）を注釈として付与することを意味する。日立製作所は，類似画像検索の研究開発の着手時から，画像とテキスト情報との統計的な関連性に着目した研究開発を行っていた。ただし，当時の検討対象は，検索用キーワードが付与された数万件程度の画像であった。2008年に，Web画像検索サービス「GazoPa」を立ち上げるため，改めてWeb画像のクローリングを実施した。現状の収集画像数は約1億件である。

クローリング時には，画像のタグを取り囲む前後のテキスト情報も収集しており，画像のメタ情報の一つとして管理されている。ただし，これらのテキスト情報を構成する単語のうち，いずれが画像と関連するかは不明である。場合によっては，テキスト全体が画像とは無関係かもしれない。われわれが構築した画像アノテーションシステムでは，類似画像検索結果中の単語の分布を統計的に評価することによって，その類似画像検索結果を特徴づける単語の集合を抽出する[5)]。これによって，ノイズが多い言語情報の中から，問い合わせ画像と関連すると考えられる単語を推定する。この推定処理は，当然ながら言語依存性はない（**図3**参照）。

データ全体の規模と個々のデータの信頼性は，しばしば相反する関係となる。画像アノテーションでは，データ数の増大に伴い推定精度が向上することが確認されている。このように，信頼性の低い情報を大規模に収集することによって，信頼性が高い情報を抽出するような発想は，ビッグデータの利活用では非常に重要と言える。

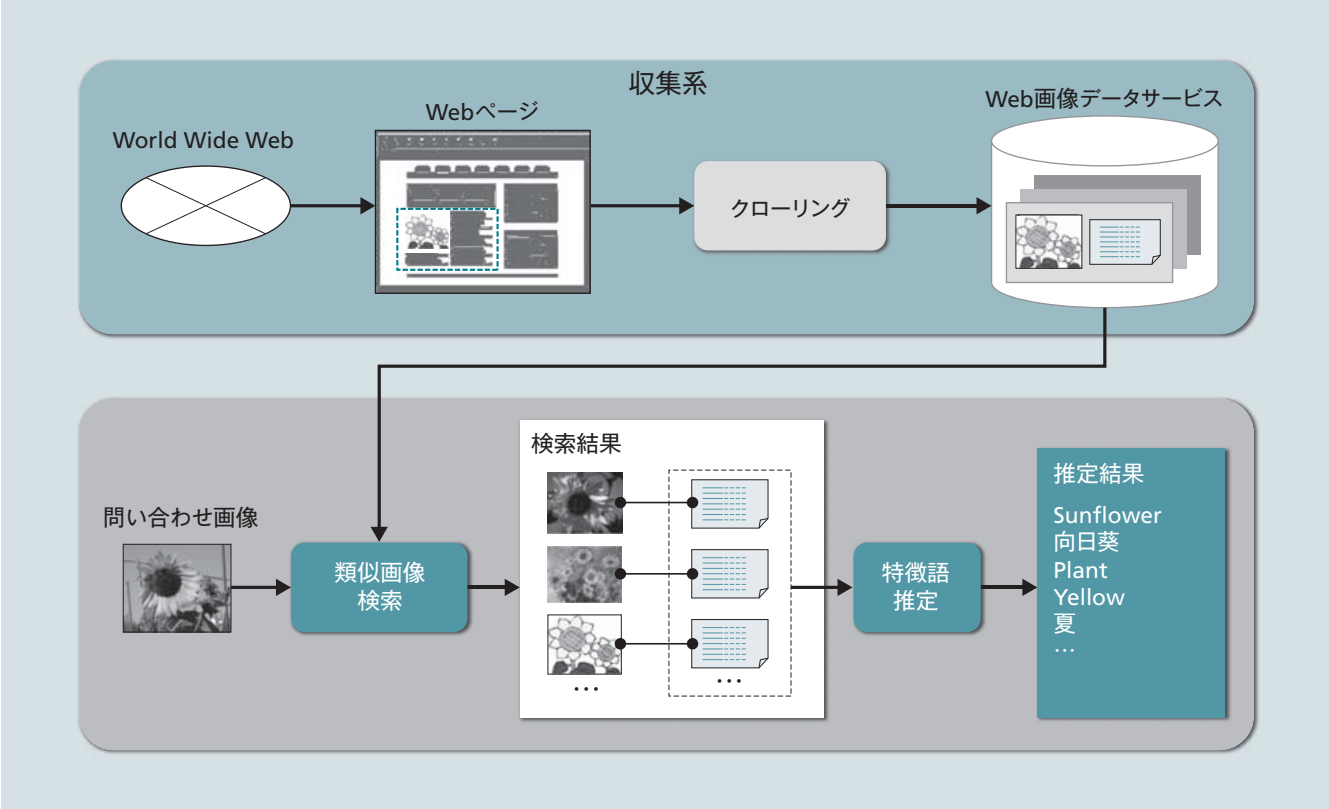


図3｜Web画像を用いた画像アノテーション
クローリングによってインターネット上から画像を収集し，データベース化する。その際に，画像周辺のテキストなどのメタ情報も関連情報として格納する。アノテーション処理では，類似画像検索の結果と関連づけられた単語の集合から，検索結果を特徴づける単語を統計的に推定する。

3. 音声データを対象とする検索技術

3.1 大規模音声データの活用

テキストデータや画像データと同様に，大量の音声データが蓄積されるようになってきている。代表的な例はコールセンターであり，近年のコンプライアンス意識の高まりとも相まって，一定期間の通話データをすべて録音するという機関が増えている。また，インターネット上の動画投稿サイトが普及し，音声を含む動画データの蓄積量も増えてきている。そのほか，放送音声や会議音声，監視システムで収録した音声などからの情報抽出が可能になれば，さまざまに応用が広がることが期待される。

音声データは一覧性が低く，内容確認のためには人間が時間をかけて聴取しなければならないという課題があった。しかし，近年のさまざまな自動解析技術の進歩に伴い，こうしたデータを効率的に活用できるようになることが期待されている。以下では，そうした自動解析技術を代表するものとして，特定のキーワードの出現箇所を見つける音声検索語検出（STD：Spoken Term Detection），特定の話者が話している箇所を見つける話者識別，特定の感情を伴う音声を見つける感情認識について述べる。

3.2 音声検索語検出

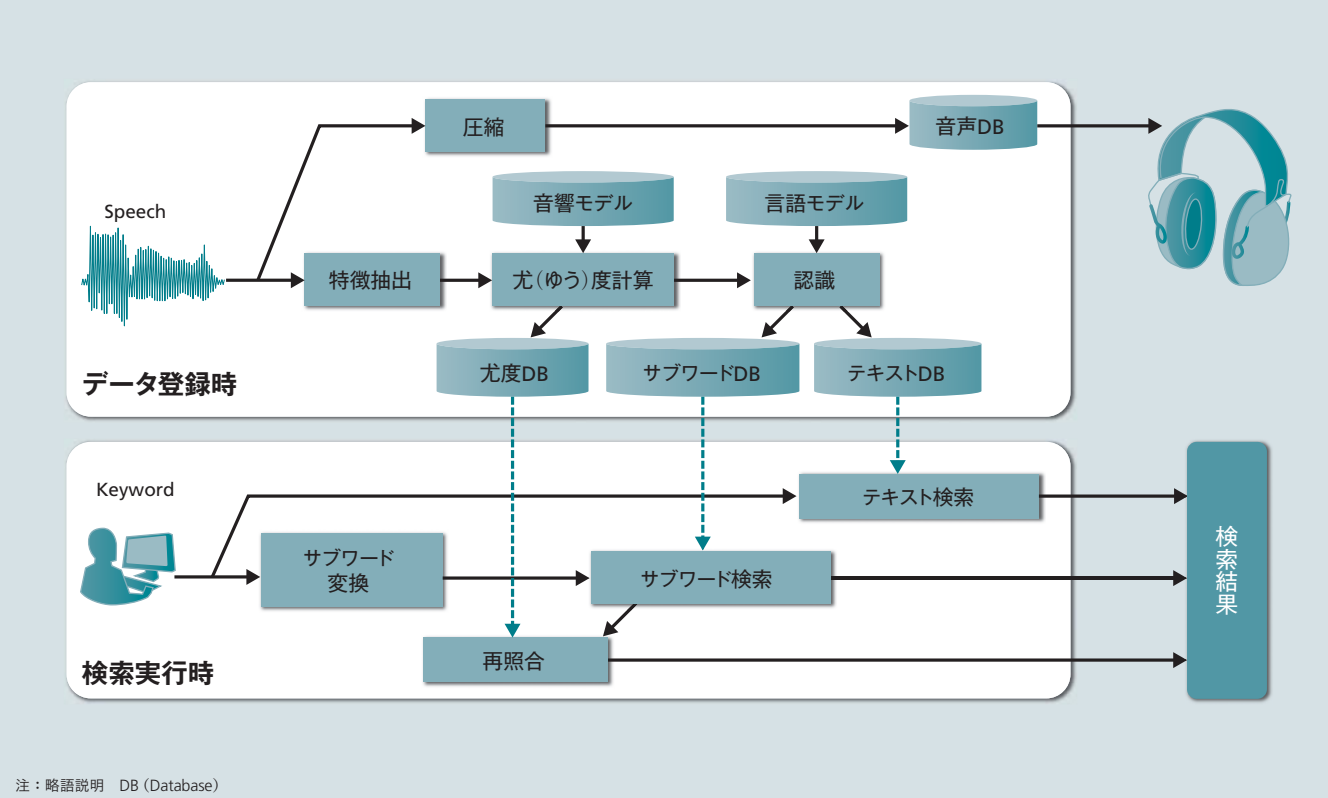
音声データ中に含まれるキーワードの位置を検出する音声検索語検出は，近年特に注目度が高まっている技術である。米国の標準化機関であるNIST（National Institute of Standards and Technology）主催のコンペティションが行われたり[6]，日本では学会主導で評価用データベースが配布されるなどしている[7]。

音声検索語検出の代表的な方式として，音声認識技術を用いて入力音声をすべてテキスト化し，検索実行時にはテキスト検索技術を用いるというものがある。この方式は，文章中の単語のつながりを表す言語モデルがタスクにマッチしている場合にはよい性能を示すが，そうでない場合には精度が低くなるという問題がある。

極端な例として，言語モデルを構成する単語辞書に含まれない検索語に対しては，検出率がほぼ0になってしまう。ビジネス用途では，人名や製品名などの固有名詞で検索を行いたいことが多いが，これらの語が未知語である可能性は高い。

こうした問題に対し，われわれは，サブワード検索に基づく方式を採用している[8]。サブワード検索に基づく音声検索語検出の構成を**図4**に示す。この方式では，音声データを単語列として認識する代わりに，音素や音節などのサブワードの列として認識し，データベースとして登録しておく。

検索実行時には，入力キーワードをサブワード列として



注：略語説明　DB (Database)

図4｜サブワード検索に基づく音声検索語検出の構成

データ登録時には，人間が聴取するための圧縮音声データベースと，システムが検索を行う際に参照するための各種データベースとを作成する。検索実行時には，3種類の検索を並行して行う。テキスト検索は最も高速だが，未知語の検索ができないなど制約も多い。サブワード検索は，語彙の制約がなく，速度もテキスト検索にさほど劣らない。再照合は，上記二つに比べるとかなり低速だが，その分高い精度が得られる。なお，テキスト検索で高い精度を得ることが期待できない場合には，テキストデータベースの作成を省略することにより，データ登録の処理量を大幅に減らすことができる。

表現し，データ登録時に得られたサブワードデータベースとの間で検索を行う。対象言語に含まれる任意の語はサブワード列として表現可能なので，この方法では未知語の問題は生じない。

この方法の弱点は，サブワード列の誤認識によって検出漏れが生じることであるが，照合時にある程度の冗長性を持たせることで，漏れを減らすことができる。さらに，サブワード検索によって得られた候補に対し，より詳しいスペクトル情報を用いた再照合を行うことにより，誤検出を減らすことが可能になる。

これらの方式は，処理速度，検出精度，対応可能キーワード数などの点でそれぞれ長所と短所があり，実際のシステムでは，それぞれの特徴を生かしながら，直列あるいは並列に相互接続することにより，実用性の高い検索機能が提供されている。

3.3　話者識別

会議音声や監視音声などの解析では，「何を」話しているかの解析と同じくらい，「誰が」話しているかの情報が重要であるケースも多い。こうしたケースでは，連続した音声を発話者ごとに分割したうえで，それぞれの塊に対し，話者の特定を行う必要がある。

認証を目的とする話者識別では，登録時と認証時で同じ言葉を使ったり，システムが発話内容を指定したりといった方式が可能だが，大量音声データの解析では，常に発話内容未知の状態での識別が必要なため，より高度な識別技術が必要となる。話者の特定は，登録音声から作成した特定話者音響モデルと，識別対象音声とのマッチングによって行うが，近年，このマッチングに因子分析の手法を応用したiVectorと呼ばれる手法の導入により，識別精度が大きく向上するようになった[9]。

3.4　感情認識

業務音声の解析においては，トラブル検知を目的として，顧客が怒っている声を見つけたいといったニーズが存在する。あるいは，オペレータの評価やマーケティングなどの効率化を目的として，顧客のうれしそうな声を見つけたいというニーズもある。

このような，声からの感情認識では，そもそも正解が一意に定まらないうえに，感情のこもった実データの入手が難しいということもあり，専ら声優などによる模擬音声を使った研究が行われてきた。こうした条件下では，例えば基本4感情（平静・怒り・喜び・悲しみ）ならば70％近い認識率が得られるといった例もある[10]。しかし，日常の

音声に含まれる微妙な感情を識別することは難しく，実用のためには，音声検索語検出と組み合わせるなどの工夫が必要であろう。

4. おわりに

ここでは，非構造データに対応した検索技術として，画像検索技術と音声検索技術における日立の取り組みについて述べた。

ビッグデータという言葉によって期待される情報システムを実現するためには，収集，蓄積される大量かつ多様な情報の再活用を実現することが求められる。そのための要素技術として，画像検索技術と音声検索技術が重要になる。このような各メディア種別に応じた情報処理技術の高度化に加え，今後は，メディア種別間をまたぐような機能提供も必要である。日立製作所は，個々のメディアに応じた最適な情報処理技術とともに，多様な非構造化データを統一的に扱うことができるプラットフォームの開発を進める[11]。

参考文献など

1) 廣池，外：VR空間を用いた画像特徴量空間の可視化―画像データベースの検索・ブラウジングのためのユーザインターフェイス―，電子情報通信学会技術研究報告，PRMU98-86，17～24 (1998.9)
2) 廣池，外：大規模な画像集合のための表現モデル，日本写真学会誌，66 (1)，93～101 (2003.2)
3) D. Matsubara, et al. : High-Speed Similarity-Based Image Retrieval with Data-Alignment Optimization Using Self-Organization Algorithm, ISM2009, 312-317 (2009.12)
4) 渡邉，外：類似画像検索に基づく事例ベース一般オブジェクト検出，電子情報通信学会技術研究報告，PRMU2011-142，101～106 (2011.12)
5) 渡邉，外：大規模Web画像データベースを用いた画像アノテーションシステムの構築，情報処理学会研究報告，Vol. 2012-CVIM-181，No. 8，1-6 (2012.3)
6) NIST Information Access Division: Spoken Term Detection Portal, http://www.itl.nist.gov/iad/mig/tests/std/
7) 西崎，外：Spoken Term Detectionのためのテストコレクション構築とベースライン評価，情報処理学会研究報告，SLP-81，No.13 (2010.5)
8) 神田，外：多段リスコアリングに基づく大規模音声中の任意検索語検出，電子情報通信学会論文誌，Vol. J95-D，No. 4，969～981 (2012.4)
9) N. Dehak, et al. : Front-End Factor Analysis for Speaker Verification, IEEE Trans. ASLP, Vol. 19, No. 4, pp. 788-798 (2011.5)
10) N. Sato, et al. : Emotion Recognition using Mel-Frequency Cepstral Coefficients, Journal of Natural Language Processing, Vol.14, No.4, pp. 83-96 (2007. 7)
11) 池田，外：非構造化データ利活用のためのメディア処理技術，人工知能学会誌，Vol. 28，No. 1 (2013.1)

執筆者紹介

廣池 敦
1994年日立製作所入社，中央研究所 情報システム研究センタ 知能システム研究部 所属
現在，類似画像検索システムの研究開発に従事
博士 (工学)
日本心理学会会員

大淵 康成
1992年日立製作所入社，中央研究所 情報システム研究センタ 知能システム研究部 所属
現在，音声認識および音声情報処理の研究開発に従事
博士 (情報理工学)
IEEE会員，電子情報通信学会会員，情報処理学会会員，日本音響学会会員