IMARS:マルチメディア・コンテンツの価値を生かす先進的テクノロジー

Web 上の情報検索において、検索エンジンはテキストにのみ索引を付け、画像は関連テキストに基づいて検索しています。企業が使用しているデータベースでも、構造化されたデータは索引付けや照会が可能ですが、画像やビデオは BLOB（Binary Large Object）型データとして処理され検索の対象外でした。

今日、高機能カメラを搭載したモバイル・デバイスが世の中に普及したことに伴い、画像や映像コンテンツが急増しています。しかし、画像や映像を検索するために、テキスト・ベースのメタ・データを手動で付与していては、リッチ・メディア・コンテンツの爆発的な増加に対応できません。手動によるメタ・データのタグ付けは、非常にコストと時間がかかる上に主観的で一貫性のないタグが付けられがちで、システムのパフォーマンス低下を招いてしまいます。マルチメディア・コンテンツの価値を十分に生かすには、新しいテクノロジーが必要です。

画像の視覚的特徴を理解することができるテクノロジー、IBM Multimedia Analysis and Retrieval System（以下、IMARS）はまさにそれを実現しています。

Matthew Hill

Senior Software Engineer
Multimedia Research,
Intelligent Information Management,
IBM T.J. Watson Research Center

【プロフィール】
T.J.ワトソン研究所のシニア・ソフトウェア・エンジニアとしてIMARSの開発の従事。現在は、IBM InfoSphere Streamsを活用したIMARSのスケールアップに注力している。IMARSおよび関連技術にかかわる複数の特許取得・学術論文発表などの成果も挙げている。前職では産業用マシン・ビジョンのソフトウェア開発を担当。コンピューター・サイエンス修士（コロンビア大学）。

IMARS について

IBM Research は、長年にわたり IMARS の研究開発に取り組んできました。この間、標準的なプロセッサー、メモリーおよびネットワーキングの進歩により、IMARS を実現する技術も進化してきました。1990 年代、IBM Research は QBIC（Query By Image Content）システムにより、画像検索テクノロジーをリードしていました。QBIC によって、画像の視覚的特徴を検索キーとして照会を行い、類似する画像を検索できるようになったのです。現在もこのような類似性検索のためのさまざまなアプリケーションがありますが、IMARS は視覚情報の処理における人間とコンピューターとの認識能力の隔たり（セマンティック・ギャップ）を埋めるために新しい機械学習技術を採用しており、見た目が類似している画像を探すのではなく、意味的に類似する画像を検索するという非常に際立った特徴があります。

IMARS は、数百から数千にも上る大量の学習用画像データを参照し、分析対象の画像を視覚的に定義している概念が何かを分析することによって、学習用画像データと共通する視覚的概念（例えば「爆発」「ビル」「食物」「結婚」など）にモデル化します。その後、学習によって作成されたラベルを基に、画像や映像に自動的にメタ・データをタグ付けすることで、分析対象のコンテンツの索引付けや検索を実現しています。これらのラベルを組み合わせて、より複雑なモデルを作成することもできるので、例えば「Web から取得したビデオ 10 万本の中から車のタイヤ交換をしている人が映っているビデオを検索」というような照会も可能になるのです。

IMARSとは

IMARS は、画像および映像をその視覚的特徴に基づいて分類するための学習可能システムです。機械学習の新技術を採用しており、学習用データを基に画像を分類するための識別機能を提供するとともに、優れた画像特徴抽出技術を利用して、学習によりさらに識別機能を高めることができます（**図 1**）。IMARS が実現する機能は以

下の通りになります。

- 大規模なマルチメディア・リポジトリー内の画像や映像に対する自動ラベル付け。
- リアルタイムの映像ストリームに対する視覚的特徴に基づく分類、フィルタリング、ルーティング。
- ほかのマルチメディア・データ分析（音声、言語、テキスト、OCR）と組み合わせたコンテンツ・ベースの効果的なマルチメディア検索。

IMARSの価値

マルチメディア・コンテンツの手動タグ付けは、膨大な人的資源とコストを要し、一般的にその映像コンテンツの再生時間の 10 倍を超える時間が必要といわれています。例えば 1 時間の映像にタグを付けるのに 10 時間かかるということです。それにもかかわらず、手動で付けたラベルは大抵一貫性のない不完全なものになりがちです。画像や映像のデータ量が急増する今日、従来 100%手入力で行っていたタグ付け作業に代って IMARS を活用することにより、学習用データとして 1~5%の映像コンテンツに手作業でタグ付けするだけで、残りの 95 ～ 99%に対するタグ付けを自動化することができます。そして学習用データから作成されたラベルとその信頼度をコンテンツに対して自動的に付与することで、従来ラベル付けに掛かっていたコストの大幅な削減を実現するのです。

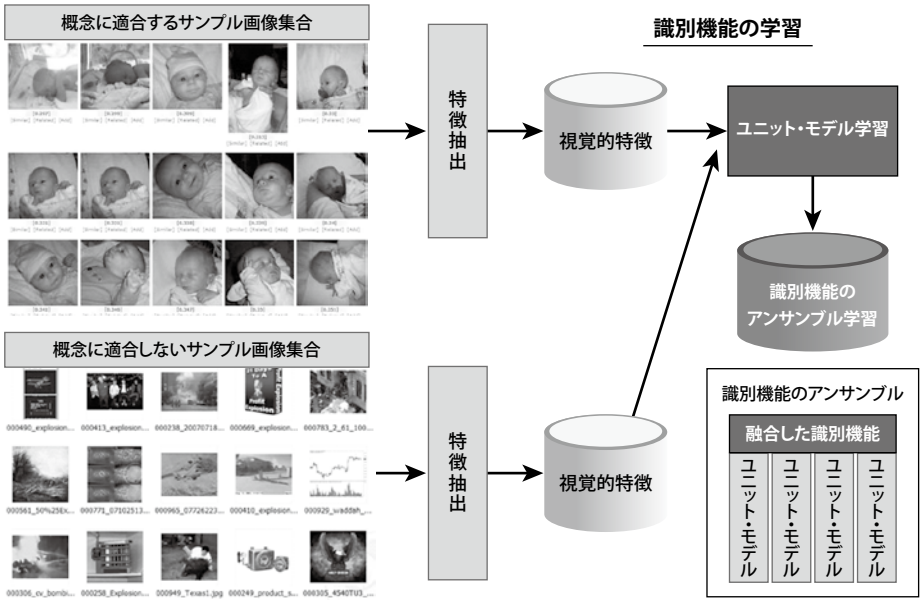
技術的アプローチ

これまでのマルチメディア・コンテンツに対する分析・特徴抽出・分類技術の進歩により、その検索機能やフィルタリング機能も向上してきました。しかし、マルチメディア・コンテンツから自動抽出できるのは、低レベルの特性表現（色、テクスチャー、形状、モーションなど）であり、マルチメディア・システムのユーザーが識別できる意味のある記述（物体、イベント、シーン、人、概念など）との間には、依然としてセマンティック・ギャップ（人間とコンピューターとの識別能力の隔たり）が残っています。それに対して IMARS テクノロジーは、視覚的な特徴を分析して画像や映像コンテンツに自動的にタグ付けをするために、以下のような独自のアプローチを採用しています。

- コンテンツに付与されたラベル間の関係や相関を活用して、処理のパフォーマンスを向上させる。
- コンテンツに対して、ラベルとそのラベルに対する信頼度を自動的に割り当て、手動タグ付けの労力を減らし検索処理を改善させる。
- 概念的な関係性を表すオントロジーを利用して、そのコンテンツが意味する概念を構造化し、検出パフォーマンスを向上させる。

画像や映像が表現する概念の検出

IMARS において、画像や映像が表現する概念モデルのライブラリーを作成するための鍵となるのが機械学習です。その学習過程においては、人との対話が必要になります。ただし、手入力が必要なのは小さなデータ・セット（学習用データ・セット）についてのみで、作成されたモデルは学習・検証され、まだラベル付けされていない大規模な画像・映像のリポジトリーに対して適用できるようになるのです。



学習用データが提供されると、IMARSでは新しい識別機能が作成される。

図 1. 学習用データの視覚的特徴を分析して識別機能を組み合わせ、画像を分類する IMARS

技術的ブレークスルー

IMARS は、数千にも及ぶ概念を学習し分類するという技術的ブレークスルーを達成しました。この技術によって、画像や映像の高度な分析・検索を行う

IMARS はさまざまな映像のカテゴリーに対して識別機能の学習が可能。例えば、スポーツの場合、150 種類以上のスポーツの画像を自動的に識別する。

図 2. IMARS の活用例

ための強力な「概念を識別するベース」が確立されました。この非常に大規模な識別機能の集まりから生成される信頼度を分析することにより、「モデル・ベクトル」技法に基づいた高精度な検索と照会が可能になりました。ここで、ある識別機能に関する信頼度がほかの識別機能に関する信頼度に影響する確率を使用して処理を行うというもう 1 つのブレークスルーが生まれました。これは、例えばコンピューターがビデオ・クリップに「空」「水」「砂」「人々」の存在を示すラベルを割り当てた場合、そのビデオ・クリップが「海辺」である信頼度が上昇するというもので、さらに学習用データ・セット内の相関と統計情報を抽出することによって、このような関係をコンピューターが自動的に学習することができます。

IMARS の事例

IMARS は、Web 写真、ユーザーが作成したビデオや写真、ソーシャル・メディア、モバイル・メディア、ニュース映像のアーカイブ、ロゴなどのブランド・コンテンツ、医用画像など、さまざまな領域の画像および映像の分析や検索に適しています。典型的な事例を**図 2** に示します。

この事例では、IMARS は 150 種類以上のスポーツを識別できるように、高品質な学習用データを用いて機械学習を行いました。事前学習が完了すると、IMARS は新しい画像や映像に対して自動タグ付け、フィルタリング、検索ができるようになり、例えばアイス・ホッケーとカーリングといった視覚的にある程度似ているスポーツに関しても区別することが可能になります。これは、最も際立った画像特徴の違いを自動的に発見してカテゴリー分けをするというユニット・モデル・アプローチによって実現されています。

また IMARS は、津波被害調査で撮影された画像の分類においても活用されています（本誌 20 ページ以下：インタビュー③参照）。調査者によって撮影された数百枚の写真から「崩壊した建物」や「がれき」などの概念モデルを作成しました。**図 3** は、自動学習された「がれき」画像を識別する機能によって高順位を付けられた画像を示しています。

図 3 を見ると、信頼度の高い画像は確かにがれきを示していることが分かります。しかし、どのようにすれば、大量の画像に対する識別機能の性能を定量化できるのでしょうか。わたしたちは、IMARS の機械学習のために「テスト・セット」を用意し、適合率・再現率という測

定基準を使用して、識別能力に関する統計量を計算しました。**図 4** は「崩壊した建物」を分類する識別能力に関する適合率・再現率のグラフです。IMARS によってテスト・セットの各画像に対する各ラベルの信頼度リストが生成されると、この信頼度リストをテスト・セットの「正解データ（Ground Truth）」と比較して、信頼度の閾値（しきいち）を推定し設定することができます。

適合率を算出するには、設定した閾値より高い信頼度を持つと判断された画像のうちの正解画像数を、閾値より高い信頼度を持つと判断された画像の総数で割ります。再現率は、閾値より高いスコアを持つと判断された画像のうちの正解画像数を、正解画像の総数で割ります。例えば、**図 4** の右下がりの線は、100% の適合率では実際に崩壊した建物の写真の約 50%を再現でき、閾値を下げることにより約 65% の適合率となった場合は崩壊した建物の約95%を再現できたことを示しています。

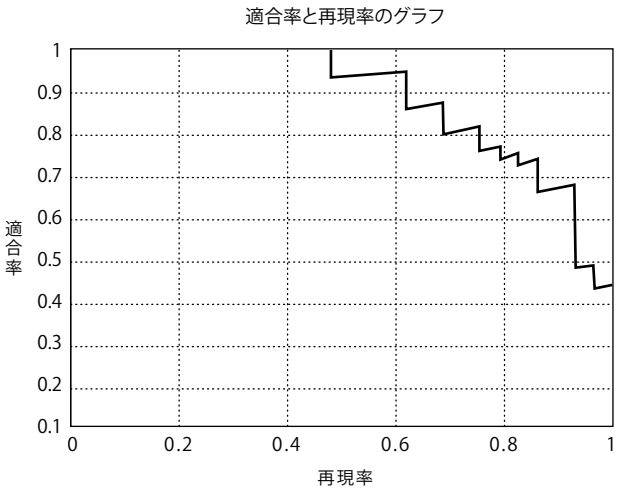


図 4.「崩壊した建物」の識別性能測定

大規模環境での高い識別性能の達成

この記事では、マルチメディアのデータ・サイズが爆発的に増加している中、IMARS がどのように役立つかを概観してきました。そこで IMARS は大規模な環境でも効果を発揮できるかという疑問が浮かぶのですが、もちろん発揮できます。

IBM のビッグデータ・プラットフォームには、IMARS が利用できる 2 つのコンポーネントである IBM InfoSphere BigInsights（以下、BigInsights）と IBM InfoSphere Streams（以下、InfoSphere Streams）があります。どちらもクラスター・コンピューティング・プラットフォームであり、ユーザーはクラウド内で非常に多くの CPU コアを簡単に利用することができます。BigInsights は、Apache Hadoop を基盤にしており、一度に大量のデータをバッチ処理する能力に優れています。InfoSphere Streams は、ブロードキャストからのビデオ・フレームのストリームなど、「移動中のデータ」をターゲットにしていて、低レイテンシー、高帯域幅のアプリケーションに適しています。BigInsights には IMARS の機械学習と識別機能を、InfoSphere Streams には特徴抽出と識別機能を実装することができます。また、IMARS は設計上、並列実行が非常に容易で、各識別機能はほかの識別機能から独立しているため、クラスター内での負荷分散が可能です。また、画像もそれぞれ独立して処理できるので、ワークロードの分散方法もさまざまに選択することが可能です。

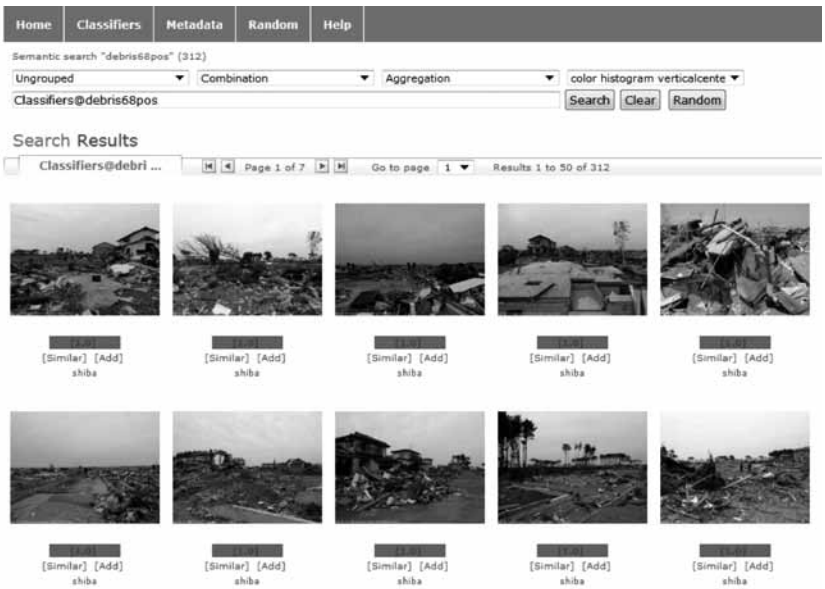


図 3.「がれき」ラベルとして高い信頼度が付けられた画像

IBM では、IMARS を「ビジュアル・ワールドを理解する先駆的なテクノロジー」と位置付け、さらなる識別機能の作成と、さらに多くのトレーニング・データの収集に努めています。少し前までは不可能と考えられていたアプリケーションが今では実現可能になっていることから、コンピューターが画像や映像が表現する概念を理解するという新しいマイルストーンを達成できる日も近いと考えています。