2004－AVM－45　(5)
2004／6／18

高度映像検索のためのメタデータ記述とシステム開発

伊藤学[1]　小池真由美[1,2]　池田佳代[3]　日高宗一郎[4]　青木輝勝[1]

[1]東京大学　先端科学技術研究センター　[2]（有）エスパリエ
[3]（有）エクセリードテクノロジー　[4]国立情報学研究所

E-mail : ito@mpeg.rcast.u-tokyo.ac.jp

あらまし　ネットワークがすみずみまで行き渡った IT 社会が実現する今日，“デジタルコンテンツ流通”が加速することは間違いない．このとき重要なことは，誰でも簡単に映像コンテンツを作成，発信をできる環境を作り上げることである．このような背景のもと画像検索に関するニーズが急速に高まってきており，これに歩調を合わせるかのように MPEG-7 に代表される画像検索のためのメタデータの国際標準もほぼ固まりつつある．そこで本研究開発では，誰もが簡単に映像コンテンツを創生・発信できる環境を実現することを大目標として，高度なデジタルコンテンツ検索のためのメタデータ記述とシステム開発について述べる．

Meta-data Description and System Development for Advanced Video Retrieval

Manabu ITO[1], Mayumi KOIKE[2], Kayo IKEDA[3],
Soichiro HIDAKA[4] and Terumasa AOKI[1]

[1] The University of Tokyo RECAST　[2] Espalier Inc.
[3] Excellead Technology Inc.　[4] National Institute of Informatics

Abstract *There is no doubt that the distribution of digital contents is accelerating on the network highway in Information Technology today. It's very important on this occasion, making the video contents available to everyone else for producing and distributing. Under this situation, there are growing needs for image/video retrieval. The International Standards of content metadata, such as MPEG-7, are almost finalized as if they keep line with such growing needs. This paper describes meta-data description and system development for advanced video retrieval of digital contents*

１．　はじめに

ネットワークがすみずみまで行き渡った IT 社会が実現する今日, “ デジタルコンテンツ流通 ”が加速することは間違いない. このとき重要なことは, 誰でも簡単に映像コンテンツを作成, 発信をできる環境を作り上げることである.

従来画像検索の研究においては, 動画像中のどのような特徴量を用いて画像検索を行うか, その特徴量を動画像中からどのようにして抽出するか, あるいはあらかじめ検索目標となる画像が用意されており, どのようにしてその目標画像が動画像中のいずれにあるかを特定するか, などの観点から研究が進められてきた. しかしながら, 画像検索の研究が数多くなされているにも関わらず, 十分実用に耐えうる検索技術が出現していないのは, 画像が本質的に持つ意味の多義性のために, キーワード付けや内容把握, 内容検索の実現が簡単ではない, ということが最大の要因であると考えられる. 人間は画像に対してキーワードのような言語的な認識だけでなく, 意味的な認識と感覚的な認識とを組み合わせて認識しており, また文字検索と異なり, 検索の前に検索対象がはっきりしていないことも少なくない. つまり, 画像検索の難しさは, 検索目標であるにもかかわらずその目標画像に対する記憶があいまいであり, 目標画像のイメージを正確に描けないことに起因しているとも言える. したがって, 高精度な画像検索の実現のためには, ユーザの検索要求をどのように入力し, システムがその入力をどのように解釈・処理し, 結果をユーザに返すかというユーザインターフェースの観点からの積極的な検討が必要不可欠であり, 特にユーザのクエリー生成をどのように支援するかは画像検索技術における最大の課題であると言える.

このような背景を受け, 本研究開発では, 高度な映像検索を行うための, 検索クエリー入力インターフェースとして, ユーザーにとって直感的で使いやすい入力方式と考える “ 3D 入力インターフェース ”と, 会話から抽出される検索クエリーに音声ノンバーバル情報を加えクエリーに重み付けなどを行う“ 音声ノンバーバル情報抽出 ”, さらにはサーバ内において高速検索を可能とする“ 超高速 XML メタデータ検索 ”の概要と, これらを搭載した, テストベッド構築の概要と, コンテンツに付与されている MPEG-7 準拠のメタデータ構造について報告する.

2. 画像検索に関する従来研究

画像検索に用いる入力インターフェースを大別すると, (A)テキスト入力型, (B)略画入力型, (C)オブジェクト選択型, (D)画像探索型に分けることができる.

上記分類(A)は, [1]〜[4]などのようにテキスト語句によるもので, 任意名詞句や制限された名詞句, また印象語や感性語, 動作語などをユーザが入力し検索を行うものである.

一方, 上記分類(B)は[5]〜[11]などのようにスケッチを描くことでクエリーとするものである. これらのスケッチによる検索は, ユーザの略画の描き方にも精度が依存するが, 一般的に厳密な検索に用いることは困難である.

上記分類(C)オブジェクト選択型とは, 例えば簡単な図形を矩形領域中に配置し, 検索目標のイメージを構成したり, また検索目標のアイコンを配置, 変形しユーザの目標画像を示したりするものである.

上記分類(D)画像探索型とは, あらかじめ検索したい画像を所有している場合に限って利用できる技術である. このような方式は検索目標の画像が厳密に用意できる点では, 文章中からの単語の検索と同様高い検索精度が期待できるクエリー入力方式である.

しかしながら, 一般的な画像検索としてユーザが検索したいと考えている画像を参照画像として用意しなければならないことや, 画像構成が限られてしまうなど, ユーザの直観的な入力を阻害する可能性もある.

３．本研究で取組む課題

今までにない, 高度な映像検索を可能とするため, 本研究において取り組む要素として以下

のテーマが挙げられる.

・　３Ｄ入力検索ンターフェース
・　音声ノンバーバル情報抽出技術
・　超高速 XML メタデータ検索技術

本項では, これら３つの技術の説明と, これらについての有効性を検証するためのテストベッド構築, さらにはテストベッドに蓄積されるデジタルコンテンツに付与される MPEG-7 準拠検索メタデータ構造について述べる.

３．１　3D 入力インターフェース

人間が「絵を描く」という行為はたとえ簡易略画とは言え一般的には一部の人間を除くと非常に苦手としているのが現状であろう. この要因としては,

・　正確な形状・色の再現が困難であること.
・　この世の物は元来すべて 3D であるのに 2D で表現しなければならないこと.

の２点が挙げられる. 本節ではこれらのうち特に(2)の負担を軽減することで, より直感的なインターフェースが実現できることを目指す.

写真を見たときに一般的には人間は２D 写真でありながら３D 的な情報, 例えば「手前のりんごと奥のりんごは同じくらいの大きさで奥行き方向に 50cm くらい離れている」などの見方で脳に記憶させている.

このため, より直感的な入力インターフェースを設計するにあたっては 3D モデリング的手法を取り入れることが望ましいと考え, 図１に示す３D 入力インターフェースを提案する.

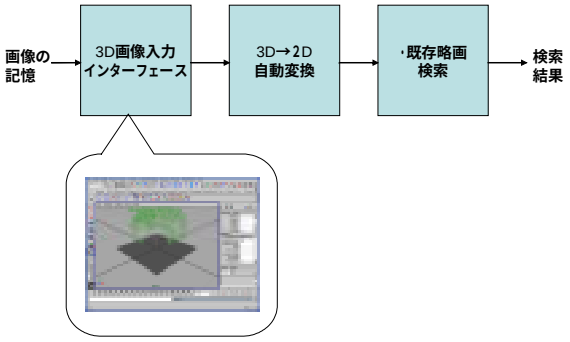


図１　3D 入力インターフェース

図１において,「3D 画像入力インターフェース」とは, 一般的な市販 3DCG 制作ソフトウェアのようなインターフェースを有し, 直方体, 立方体, 球, 回転体などが極めて容易に作成できるモジュールである. 続く「3D→2D 自動変換モジュール」は本提案方式のキーとなる部分であるが, 一般的に使われているカメラワークの各種技法（クローズアップショット, ウェストショット, ミディアムショット, ニーショットなど）を 15 種類程度を用いてそれぞれ並行的に作成した 3D モデルを”撮影”をする. これはまさに 3D モデリングを 2D 略画に変換する処理に他ならない. 最後にこのように撮影された 15 枚程度の略画を従来同様の略画検索ツール[12]に入力させ, 最終的な演算結果（検索結果画像）を得る.

一般的なカメラワークに関してはすでにある程度技法が確立しており, 15 種類程度の技法に基づき撮影をしておけば通常の写真, 映像にはほとんど対応可能であり, これこそが本方式において最も重要な技術根拠となっている.

３．２　音声ノンバーバル情報抽出技術

これまで情報検索技術については, 多くの手法が提案され, 実際現在の WWW サーチエンジン等でもそれらの技術は使用されているが, ほとんどの場合, その基礎としてキーワード入力に基づきワードマッチングする手法が使用されている. すなわち, 文書の内容を形態素解析に基づき単語に分解し, これらの単語の情報（出現の有無, 出現頻度, 出現位置等）を統計処理することによって検索結果を返すシステムである.

一方, 本提案のように会議中の会話内容（音声情報）を入力とする場合には, 上述した手法の他にも非常に多くの情報が含まれている. 具体的には,

・誰がしゃべった言葉か？
・何人がしゃべった言葉か？
・声の大きさ, トーンはどうか？

等である. これらのノンバーバル情報を既存検索技術と組み合わせることにより, 検索クエリ

ーの重み付けを行い, 非常に効率的な検索が可能となる. 図２に音声ノンバーバル情報を用いた検索の概念を示す.

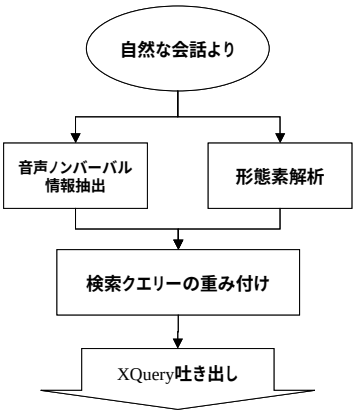


図２　音声ノンバーバル情報検索の概念

まず初めに検討している音声ノンバーバル情報は, 会話中に発言された単語(名詞句など)の回数とそれらを発した際の声のパワーについてである. 複数人で会話中に何度も発せられた単語は, そこにいる人が共通的にイメージしている内容であることは容易に判断できる. これらの方式を用いている例は研究がなされているが, これにその単語が発せられた際の声のパワーを組み込むことによって, より厳格に重み付けを行う. 図３に音声ノンバーバル情報を用いた重み付けの概念を示す.

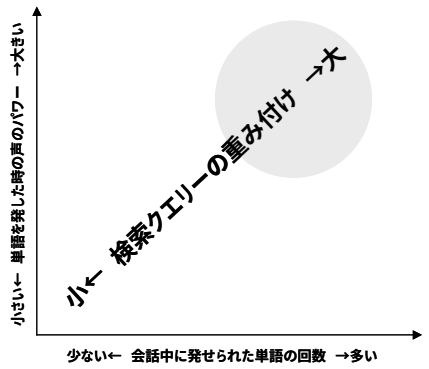
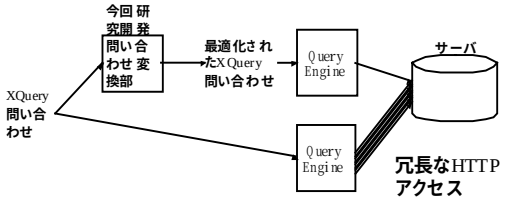


図３　検索クエリーの重み付け概念

３．３　超高速 XML メタデータ検索技術

本技術では, 映像検索実験を行うテストベッド内において格納されているコンテンツに付与され, 検索対象となる XML メタデータに対し, 前記 3.1 および 3.2 に説明した検索ツールより吐き出される検索クエリーを, XQuery に変換されて実行される最適化についての研究である. ここでの高速化は, この XQuery のソース to ソースの最適化を行うことにより実現する. 具体的には, 冗長な通信や計算を生じるような XQuery の問い合わせ式を, 意味的に等価で冗長性を軽減するような式に変換する. 図４に XQuery 最適化の概念を示す.

この技術により最適化されたアルゴリズムをテストベッド内にて実装することにより, 意図するコンテンツの高速検索に寄与する.



・最適な問い合わせに書き換えることにより冗長な通信が削除される

図４　XQuery 最適化の概念

３．４　テストベッド構築とコンテンツメタデータ

前項 3.1～3.3 で述べられた各技術（3D 入力インターフェース, 音声ノンバーバル情報抽出, 超高速 XML メタデータ検索）について，これらの有効性を実証するためのテストベッドを構築した. 図５にその全体概要を示す. クライアント PC 側には，3D 入力インターフェース（図中①, ②）及び音声ノンバーバル情報抽出システム（図中③, ④）を搭載し，これらにより吐き出されるメタデータ（検索クエリー）を⑤XQuery 入力インターフェースに送り，⑥AVR クライアントソフトウェアーを介して, サーバに検索を行う. サーバ側ではクライアントより送られてきた XQuery に対し, 冗長な通信や計算を生じるような XQuery の問い合わ

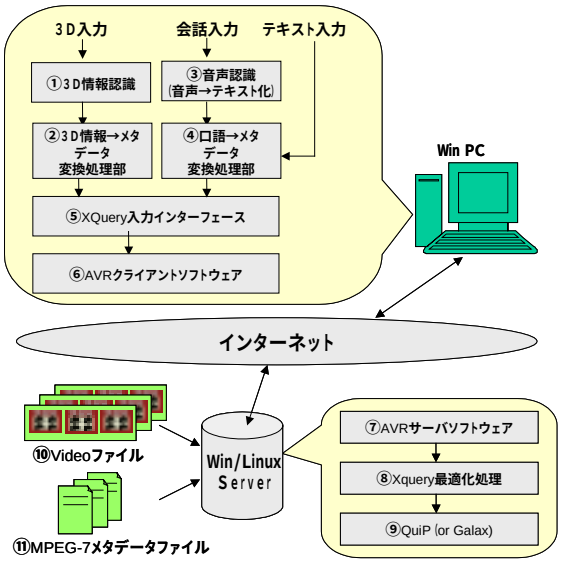


図５　システム全体図

せ式を, 意味的に等価で冗長性を軽減するような式に変換し, サーバ内に蓄積されているコンテンツ（⑩）に対し付与されている XML メタデータ（⑪）に高速に検索を実行するものである.

本テストベッド内でコンテンツの検索対象として用いているメタデータには MPEG-7 を採用している. MPEG-7 を用いる事は, 近年注目を集めているアーカイブや映像ライブラリーが分散設置された場合, 機器・データ間で記述フォーマットの共通化・互換性確保をする事で, ユーザにとって検索しやすい環境を提供できると考えたためである.

MPEG-7 において記述するメタデータとしては, 大きく分けて 2 種類ある. 一つはローレベルなメタデータ（Visual, Audio）, もう一つはハイレベルなメタデータ（MDS: Multimedia Description Schemes）である. 前者は, 画像（色, 形, 動きなど）や, 音声（音色, 効果音, メロディなど）に関する特徴量を PC などを用い自動的に抽出するもので, 後者は, コンテンツの内容（タイトル, 内容説明, キーワード, 制作日など）を手入力により記述するものである. [13]〜[15]

3D 情報による検索では, あらゆる特徴量抽出が考えられる. また, 抽出された情報が現在の MPEG-7 スキームにおいて, どのパートに属するのかなど, 今後検討を重ねていく必要がある. また, 新たな記述スキームの提案にいたる可能性も十分に秘めている. よって, まずコンテンツ検索に最低限必要と思われる記述項目を用意した.

図６にサーバに格納される MPEG-7 メタデータの木構造を示す. コンテンツタイトル, ロケーション, 撮影時期, さらには誰が？何を？といったストラクチャーの他, コンテンツの ID, サムネイル及びコンテンツ実体の保管場所などの記述も対応している.

```
<MediaInformation>
        <MediaProfile>
                <MediaInstance>
                        <InstanceIdentifier>

                        <MediaLocator>
                                <MediaUri>

<CreationInformation>
        <Creation>
                <Title>

                <TitleMedia>
                        <TitleImage>

                <CreationCoordinates>
                        <Location>

                        <Date>
                                <TimePoint>

<TextAnnotation>
        <FreeTextAnnotation>

        <StructuredAnnotation>
                <Who>
                <WhatObject>
                <WhatAction>
                <Where>
                <When>
                <Why>
                <How>

        <KeywordAnnotation>
                <Keyword>
```

図６　MPEG-7 メタデータの木構造

４．まとめ

本稿では, 誰もが簡単に映像コンテンツを創生・発信できる環境を実現することを大目標として, 高度なデジタルコンテンツ検索を可能とするため, 今までにない入力インターフェースを搭載したコンテンツの高速検索システムについて述べた. 急速なデジタル化が進む今日, ネットワーク上に爆発するデジタルコンテン

ツをいかに効率よく検索するか, 現在取組んでいる課題は最重要項目であり, 急務となっている．今後は，システムの本格実装を目指し，それぞれのカテゴリにおいて実験･開発を行う予定である．

謝辞：本研究は総務省戦略的情報通信研究開発推進精度研究主体育成型研究開発平成 15 年度「簡単映像コンテンツ制作のための高度映像検索技術に関する研究（研究開発）」の一環として行われたものである．

文献

[1]芝田滝也,加藤俊一,"街路の景観画像データベースのイメージ語による検索",信学論 D-I Vol.J82-D-I No.1, pp.174-183, Jan.,1999

[2]宮森恒,粕谷英司,富永英義,"動作語を用いた問い合わせによる映像検索方式",信学論 D-II Vol.J80-D-II No.6, pp.1590-1599, June,1997

[3]原田将治,伊藤幸宏,中谷広正,"感性語句を含む自然言語文による画像検索のための形状特徴空間の構築", 情処学論 Vol.40, No.5, May,1999

[4]椋木雅之,田中大典,池田克夫,"対義語対からなる特徴空間を用いた感性語による画像検索システム",情処学論 Vol.42, No.7, July, 2001

[5]Shih-Fu Chang, William Chen, Horace J.Meng, Hari Sundaram, Di Zhong, VideoQ:An Automated Content Based Video Search System Using Visual Cues, Proc. of the 5th ACM international conference on Multimedia, 1997

[6]M.Flickner, H.Sawhney, W.Niblack, J.Ashley, Q.Huang, B.Dom, M.Gorkani, J.Hafner, D.Lee, D.Petkovic, D.Steele, and Peter Yanker, Query by Image and Video Content:The QBIC System, IEEE Computer Magazine, Vol.28,No.9,pp.23-32,Sep., 1995.

[7]金原史和,佐藤真一,濱田喬,"形状分解によるユーザの視点に基づいたシルエット画像検索,"情処学論 Vol.36, No.12, Dec.,1995

[8]黒川雅人,洪政国,"形状情報を用いた画像の類似検索システム", 情処学論 Vol.32 No.6,June 1991

[9]金原史和,佐藤真一,浜田喬,"プリミティブ分解による多様な検索条件を扱いうるカラー画像検索",情処学論 Vol.37, No.11, Nov.,1996.

[10]椋木雅之,美濃導彦,池田克夫,"対象物スケッチによる風景画像検索とインデックスの自動生成", 信学論 D-II Vol.J79-D-II No.6, pp.1025-1033, June,1996

[11]小早川倫広,星守,大森匡,照井武彦,"ウェーブレット変換を用いた対話的類似画像検索と民俗資料データベースへの適用",情処学論 Vol.40, No.3, Mar.1999

[12]青木秀一,青木輝勝,安田浩,"動画像からのシーン検索のための略画処理手法の提案",情報処理学会 CVIM 研究会,2002.1

[13]堀,“ MPEG-7 の概要と役割 ”, 情報処理学会シンポジウムシリーズ，2001，10，pp3-17（2001）

[14]柴田,“ MPEG-7 MDS チュートリアル ”, 情報処理学会シンポジウムシリーズ，2001，10，pp27-43（2001）

[15]ISO/IEC FDIS 15938-5：“ Multimedia Content Description Interface - Part 5: Multimedia Description Schemes ”, JTC1/SC29/WG11/ N4242（Oct. 2001）