J-070

映像検索のためのクエリー生成とインターフェース構築

A Study on Query Generation and Interface Architecture for Video Retrieval

伊藤 学[1)] 小池 真由美[1),2)] 池田 佳代[1),3)] 日高 宗一郎[4)] 青木 輝勝[1)]
Manabu ITO Mayumi KOIKE Kayo IKEDA Soichiro HIDAKA Terumasa AOKI

1. はじめに

高速回線の整備やデジタル家電の普及，さらに家庭用PCへは映像編集ソフトの標準搭載など，コンテンツの取得・閲覧のみならず，制作・配信までもが容易に行える時代となった。このような社会が実現する今日，“デジタルコンテンツ流通”がますます加速することは間違いないと考える．このことは，コンテンツが爆発的に増加することも意味している。このとき重要なことの１つは，目的のコンテンツを効率よく検索する技術であり，特にユーザのクエリー生成と検索インターフェースをどのように支援するかは画像検索技術における課題と言える．本報告では，これらの課題を解決する手段として，手書き略画による検索をサポートする 3D 入力インターフェースと，会話などから得られるキーワード検索に対して，検索効率を向上させるため，それらに重み付けを行うための音声ノンバーバル情報抽出，この 2 つのインターフェースと，これらを搭載した AVR（Advanced Video Retrieval）テストベッドの概要について述べる。

2. AVR テストベッドの概要

AVR テストベッドは前項で述べられた２つのインターフェースを有し，これらの有効性を実証するためのシステムである。図１にシステム全体構成を示す．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．

3. ２つの検索インターフェースとメタデータ

3.1 3D 画像入力インターフェース

従来の画像検索研究においては，動画像中に含まれるオブジェクトなどの特徴量をあらかじめ抽出しておき，ユーザが，自分の意図するコンテンツのイメージをペイントなどの手書きツールを用いて描き，それをクエリーとして検索する手法がある。しかしながら，自分のイメージを手書きで生成することは，誰もが得意とするものではない。これは，目で見える世の中の物体はほとんど 3D であるにもかかわらず，略画にする場合 2D として描かなくてはいけないことが要因と考える。しかも，手書き略画による検索の場合，略画の精度が検索精度に大きな影響を与えている。ここでは，手書き略画による検索をサポートする目的で，人間の直感に近い検索インターフェース提供を目的とした 3D 入力インターフェースを考案している。

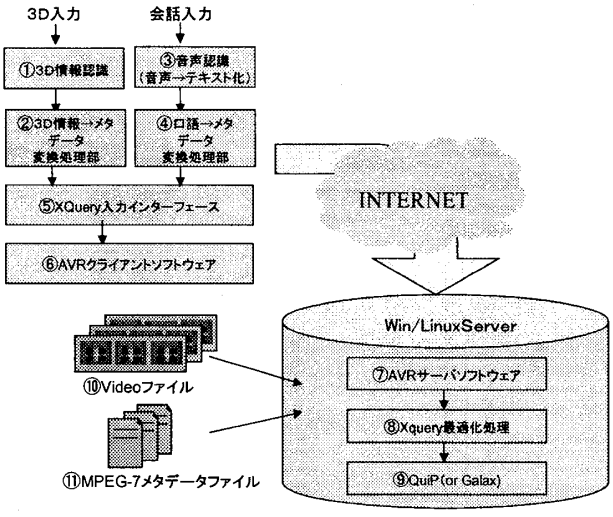

図１　AVR システム全体構成

3D 画像入力インターフェースとは，3 次元のバーチャル空間に，あらかじめ用意されたオブジェクトや，抽象的な直方体，立方体，球などを配置し，極めて容易にイメージ画像を作成できるインターフェースである．これにより生成された状態を，3D→2D 自動変換モジュールに送り，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）を並行的に作成した 3D モデルを”撮影”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である．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[1]に入力させ，最終的な演算結果（検索結果画像）を得る．

一般的に，人間が目で見て立体と感じる被写体までの距離は，おおよそ 10 メートル以内と言われている。しかしながら，通常のコンテンツは背景なども写るため，10 メートル以上の被写体も考慮する必要がある。そこで，本インターフェースの研究要素として，１）近距離における 3D インターフェースと，２）遠近の相関を考慮した 3D インターフェースの 2 つを検討している。図２にインターフェースの概念を示す。

[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

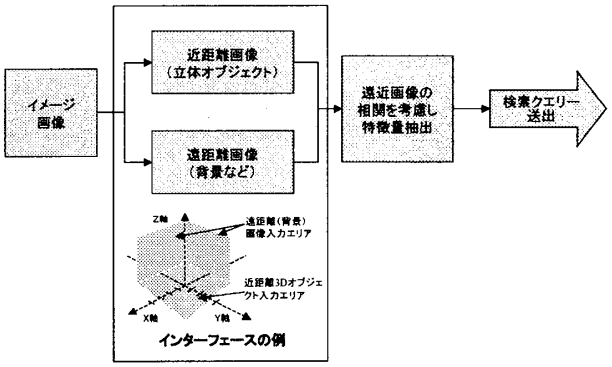

図２　3D 入力検索インターフェースの概念

3.2　音声ノンバーバル抽出技術（会話入力）

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．

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．

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．

3.3　コンテンツメタデータ

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．

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 3 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．

図３　MPEG-7 準拠コンテンツメタデータ

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]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