修 士 論 文

マルチメディアコンテンツの高度利用のためのアノテーション技術の基礎と応用

350303258 山本 大介

名古屋大学大学院 情報科学研究科

メディア科学専攻

2004年度

目 次

概要

本論文では、オンラインビデオアノテーションのための技術および、それを実現したシステム iVAS (intelligent Video Annotation Server) について述べる。このシステムでは、閲覧者がインターネット上に存在する任意のビデオコンテンツに対して掲示板風インタフェースにより、アノテーション（コンテンツに対するメタ情報）を関連付けることができる。また、このシステムはカット検出情報や色ヒストグラムを得るために動画像解析を行い、自動的にアノテーション編集 Web ページを生成する。閲覧者は次の三つの方式でアノテーションを作成することができる。一つ目は閲覧者によって、人やオブジェクトの名前や状況説明や感想を対話的にアノテーションする方式である。二つ目は、マウスのボタンをクリックするという簡便なインタフェースによって、閲覧者の主観的な印象を任意のビデオのシーンに対して関連付ける方式である。三つ目は、他者の入力したテキストに対して○×評価を行う評価アノテーションである。生成されたアノテーション情報は iVAS に接続された XML データによって蓄積・管理される。また、アノテーションに基づく応用として、ビデオ検索、ビデオ簡約、ビデオコンテンツコミュニティ支援に関するシステムを構築した。我々のアプローチの主要な利点の一つは、人手によるアノテーションと自動的に生成されるアノテーション（色ヒストグラムやカット情報など）とを容易に統合できる点にある。また、不特定多数の閲覧者からアノテーションされることが期待されるために、アノテーションデータの信頼性や正確さを考慮する必要がある。そこで、閲覧者のフィードバックを元にしたアノテーションの信頼性の機械的な評価手法を提案する。将来、これらの基盤技術がビデオコンテンツの流通によって促進される新しいコミュニティの形成に貢献できるものと考えている。

Abstract

In this paper, we present a new Web-based video annotation system, named iVAS (intelligent Video Annotation Server). Web audiences can associate any video content on the Internet with their own annotations. The system analyzes video content in order to acquire cut/shot information and color histograms. And it also automatically generates a Web content for generating and editing annotations. Then, audiences can create annotation data by three methods. The first one helps the users to create text data such as person/object names, scene descriptions, and comments interactively. The second method facilitates the users associating any video fragments with their subjective impression by just clicking a mouse button. And the third method is evaluation annotation by clicking O-X buttons for each text comments and messages written by other users. The generated annotation data are accumulated and managed by an XML database connected with iVAS. We also developed some application systems based on annotations such as video retrieval, video simplification, and video-content-based community support. One of the major advantages of our approach is easy integration of hand-coded and automatically-generated (such as color histograms and cut/shot information) annotations. Additionally, since our annotation system is open for public, we must consider some reliability or correctness of annotation data. We also present an automatic evaluation method of annotation reliability using the users' feedbacks. In the future, these fundamental technologies will contribute to the formation of new communities centered around online video content.

第1章　はじめに

コンテンツ産業は、日本にとって重要な産業の一つであるのと同時に、一般の人々が手軽に享受することができる文化的あるいは芸術的作品を公開し、その活動を促進する営みの一つである。かつて活版印刷の発明によって人々が手軽に文学に親しむことができたのと同様に、CD-ROM・DVD-ROM といったメディアが安価に製造できるようになったことで、人々は手軽に映像や音楽などのコンテンツを楽しむことが可能になった。近年、インターネットの発達に伴いこれらのメディアは Web 上のダウンロードサービスに取って代わろうとしている。音楽コンテンツのダウンロードサービスは、iTune Music Store[1]や着うたフル[2] に代表されるように成功を収めている。これらの仕組みの特徴は、メディアが必要ないため安価に複製が可能であるという点にある。そのため、コンテンツの価格を下げることが可能になり、コンテンツ産業全体の発展に寄与すると考えられている。

そこで筆者は、誰もが自由にマルチメディアコンテンツにアクセスでき、そのコンテンツを通じてコンテンツ提供者にとってもコンテンツ閲覧者にとっても利益がある社会を形成できれば、きわめて有意義であると考える。近年のブロードバンドネットワークの発達により、Web を通したマルチメディアコンテンツの配信が実現可能になった。しかしながら、誰もが自由にテキストと写真から成り立つ Web ページにアクセスできるように、マルチメディアコンテンツにアクセスできるようにはなっていない。その原因の一つは、コンピュータにとってマルチメディアコンテンツはあまりにも情報量が多過ぎて、コンテンツをうまく処理できないからである。たとえば、ユーザは検索エンジンを用いて自分の望む Web ページに到達することが可能であるが、マルチメディアコンテンツの場合、コンテンツの意味内容に基づく検索は困難である。しかしながら、マルチメディアコンテ

[1]米国 Apple 社の音楽コンテンツ配信サービス。http://www.apple.com/itunes/store/
[2]au の携帯電話向け音楽コンテンツ配信サービス。http://www.au.kddi.com/ezweb/au_dakara/chaku_uta_full/

ンツの意味内容を何らかの手段でコンピュータにとって容易に理解可能な状態にできれば、マルチメディアコンテンツをより容易に扱うことが可能になる。

それを解決する手段の一つとして、ビデオアノテーションという仕組みが近年考えられている。つまり、ビデオコンテンツに何らかの方法を用いてコンテンツのメタ情報（本論文ではこれらのメタ情報を一般にアノテーションと呼ぶ）を記述する仕組みである。ただ単純にコンテンツのタイトルを記述するだけでは不十分で、コンテンツの中身をメタ情報として検索しやすい形で保存する必要がある。ビデオコンテンツのメタ情報を記述する形式として、MPEG-7[8] という形式が策定されている。ビデオコンテンツの情報を XML ベースのテキスト形式で記述する仕組みである。つまり、ビデオコンテンツのアノテーション情報を記述するための枠組みは提供されている。しかしながら、通常ビデオコンテンツのアノテーションを詳細に施すには非常に多くの人的コストが必要である。アノテーションを効率よく作成するツールや手法は様々な方式が考えられているが、まだまだ不十分である。それはビデオコンテンツの解析を自動化するのには限界があり、人手に担う部分が必要であるからである。そのために、十分にコストが低いアノテーション作成手法が求められている。

このようにアノテーションには非常に大きなコストがかかるが、それと同じくらいの利益もある。例えば、放送局には、今なおビデオテープに収録された膨大なコンテンツが倉庫に保管されている。これらのコンテンツは番組タイトルや日付などの表層的なめたデータによってテープ単位であっても必要なシーンを探し出すためには人間の目で確認する必要がある。アノテーションの仕組みをこれらのコンテンツに適用すれば、ユーザが膨大なコンテンツの中から自分の好きな俳優の登場するコンテンツのシーンを閲覧したい、あるいはこの野球選手の高校時代の活躍を見たい、今放送されている音楽のアーティストに関する情報を知りたい、この人は誰なのか知りたい、などといったように様々な要求がある。検索以外にも、この作品を 5 分に要約して欲しい、自分の好きなシーンだけが登場するように再構成して閲覧したい、などといった要約・再構成などの要求もある。さらには、この作品のこのシーンが好きなので類似するコンテンツはないかなどといった要求もあるかもしれない。これらの応用は、一般に自動解析だけでは不可能であり、アノテーションが必須である。逆に言えば、アノテーションさえあれ

ばこのようなサービスは比較的容易に実現できる。この仕組みを Web 上に存在するすべてのマルチメディアコンテンツに適用できたとするならば、ユーザが自由にマルチメディアコンテンツにアクセスできるようになる手助けとなるであろう。

そこで、筆者はコンテンツ産業を活性化するためには、以下の二つの仕組みを備えている必要があると考えた。一つ目は、より多くの閲覧者を獲得できる仕組みであり、二つ目は、ビデオコンテンツ対して低コストで詳細なアノテーションを行うことを可能にする仕組みである。アノテーション技術は最終的にコンテンツの価値を高めることにつながり、コンテンツホルダーにとっても閲覧者にとっても利益となると考えるためである。ここで、筆者はインターネットの双方向性という特性に着目する。インターネットは TV 放送などと違ってビデオコンテンツを閲覧するだけでなく、ユーザからなんらかのフィードバックを得ることが可能である。そこで、もしも閲覧者自身に何らかの方法でアノテーションに参加してもらうことができるとしたら、コストがほとんどかからずにアノテーションを行うことができると考えられる。しかしながら、従来型のアノテーションツールを利用したアノテーションを促しても、非常に限られたコンテンツとユーザに対してしかアノテーションはされないと考えられる。では、どのような形ならば閲覧者にアノテーションを促すことができるのであろうか。

筆者は次に 2 ちゃんねる[3]の実況板に注目した。2 ちゃんねるの実況板とは、現在放映中の TV 放送に対して、議論を行うための掲示板である。人気のあるコンテンツならば、1 分間に 100 以上もの投稿がある場合もある。これらの情報を解析すれば、アノテーションとしての情報を得ることが可能ではないかと考えた。しかしながら、これらの情報は一切構造化されておらず、また信頼性の低い情報が多く含まれている。もちろん、ビデオコンテンツを解析するよりは簡単であろうと考えられるが、新たにビデオコンテンツに適した形の電子掲示板を作ることができれば、より効率良くアノテーションを作成する仕組みが実現できるだろう。また、電子掲示板をコンテンツと密に連携させることによって、そのコンテンツを中心としたコミュニティを形成でき、コンテンツとそのコミュニティを結びつけ発展させることも可能になる。ひいては、多くの閲覧者を獲得する手段となり、前述した二つの問題を解決する手段となり得ると考えられる。

[3]http://www.2ch.net/

そこで、どのような形式のインタフェースがふさわしいのかという問題になる。筆者は2ちゃんねるの実況板のような掲示板タイプのインタフェースが馴染み易いのではないかと考えた。閲覧者が、一般的なWebブラウザを用いてコンテンツを閲覧しつつついでに任意のシーンに対して、コメントを書き込むことができればよいと考えた。また、単純にそのシーンについて「おもしろい」「おいしそう」など、わざわざテキストで書き込む必要がないような閲覧者の主観的な印象を容易に投稿できるインタフェースとして、印象アノテーションの仕組みを構築した。つまり、コミュニケーションやコメントを書き込むといったユーザが日常的に行っている、あるいは行ってくれることが期待できるようなインタフェースを用意することによって、アノテーションを促進するための技術である。

このような仕組みを用意することによって、だれもが手軽に優良なコンテンツに親しむ環境ができ、コンテンツとそのコミュニティの発展に寄与できるものと考えている。

しかしながら、現状ではWeb上でコンテンツを自由に扱うために解決しなければならない問題は多い。とりわけ、インターネット上でコンテンツをやり取りする場合には著作権的な問題が大きな壁となっている。近年、音楽コンテンツのインターネット上での違法コピーの増加によって、コンテンツホルダーはインターネット上でのコンテンツ配信に極めて警戒を強めている。そのため、現在のコンテンツ技術はコンテンツの流通をより制限する方向へと向かっており、誰もがコンテンツを自由に閲覧できる環境を構築することが困難になっている。例えば、コンテンツホルダーはDRM (Digital Right Management)[4]といった違法コピーを防止するための技術の導入を積極的に進めている。これらの技術は時としてユーザの利便さを損なうことも多い。たとえば、日本のデジタル放送では映像の自由なコピーができない。許されるのは、一回限りのコピーと、ムーブと呼ばれる録画してある元データを消去した後に新しいメディアへ移動させることのみである。これでは、コンテンツのバックアップをとることができず、技術の発達とともにユーザの利便性が低下することになりかねない。もちろん、Webの発達とともに第三者によって不特定多数のユーザにコンテンツの違法コピーが配布される可能性が高まり当然の措置であるともいえる。しかしながら、ユーザは自由にコンテンツ

[4]デジタルデータの著作権を保護する技術。音声・映像ファイルにかけられる複製の制限技術や、画像ファイルの電子透かしなどもDRMに含まれる。

を扱えず不便を感じ、コンテンツに関して興味を削がれる可能性もある。これではコンテンツ産業全体の衰退を後押しすることになるかもしれない。

インターネットを通じて誰もがアノテーションを行うことが可能な環境を構築するためには、コンテンツを誰もが手軽に視聴できる環境を用意する必要がある。しかしながらそのような仕組みはコンテンツホルダーにとっての利益はなく、コンテンツ産業は衰退する。しかしながら、現状の NHK 以外の放送局はユーザから受信料等の料金を課すことなく成立するビジネスモデルを構築している。これは、大多数の人々に見られるということを利用し、広告収入から利益を得る仕組みである。つまり、映像コンテンツの配信は広告ビジネスとして成立するという側面も持ち合わせている。同じようなビジネスモデルを、Web 上でも実現できたとするならば無料でコンテンツを閲覧するといったことが可能になるかもしれない。さらに、関連グッズや映像特典付き DVD などといったパッケージメディアコンテンツを販売する仕組みと組み合わせれば、さらなる収入拡大も見込める。

それでも、誰もがビデオコンテンツを Web 上から自由に閲覧できるような社会ができることに否定的な意見もあるかもしれない。しかしながら、現状でも、一部の映像コンテンツは広告ビジネスと組み合わせられ Web 上で配信されている例はいくつかある。例えば、テレビコマーシャルと連動した映像配信である。最近のテレビコマーシャルの一部には、人気俳優が出演するストーリ性を持ったものもある。テレビコマーシャルでは 15 秒しか放映されないが、Web 上から数分の短編映像コンテンツが無料配信されるという事例も一般化してきている。ユーザにとっては人気俳優の演じるストーリ性の高いコンテンツを楽しむことが可能であるし、コンテンツ提供者にとっては商品の宣伝になる。さらに livedoor[5]など、映画コンテンツを無料配信している例もある。比較的制作費の安い若手の映像作家が、広告提供者の指定する商品を映画のシーンに多く登場させることによって製作資金を提供してもらい、映像全体を無料配信する例である。映像全部を配信しても成り立つビジネスモデルである。さらには、TV 放送している連続もののアニメコンテンツを Web 上からも見られるようにしている例もある。これらの例からも分かるように、将来多くの映像コンテンツが Web を通して無料で閲覧できるようになる可能性は十分に高い。そのために、本研究のような手法は将来役に立つ

[5]http://www.livedoor.ne.jp

であろうと考えている。

本論文では、以上の考察に基づいてビデオアノテーションの新たな手法を提案し、その実現のための基礎技術とその手法によってもたらされる応用技術の両方に関する筆者の研究成果を報告する。

本論文の構成は次の通りである。まず2章において、オンラインビデオコンテンツとその関連技術の解説を行う。特に、ビデオコンテンツに対するアノテーションには、従来からあるオフラインビデオアノテーションと筆者らが提案するオンラインビデオアノテーションの2種類についての解説を行う。次に、3章において、筆者らが作成したオンラインビデオアノテーションシステムiVASの仕組みについて解説する。さらに、4章において、オンラインビデオアノテーションによって得られた情報を用いた応用例をいくつか紹介し、5章において、それらの実験と評価を行う。6章において、今後の研究の方向性を示し、7章において本研究を総括する。

第2章　ビデオコンテンツへのアノテーション

本論文では、主にオンラインビデオコンテンツに対するアノテーションについて扱う。オンラインビデオコンテンツとはインターネットを通じてアクセスできるビデオコンテンツである。もちろん、ビデオテープやDVD等に収録された映画など、世の中には非常に多くのオフラインビデオコンテンツが存在しており、それらの存在も無視できない。しかしながら、これらに対するアノテーションの研究はいろいろなされてきているので、サーベイだけに留めておき、これらの研究と本研究の違いに関する考察は行わない。

本章では、まずビデオアノテーションとは何かを説明した後に、オフラインビデオアノテーションとオンラインビデオアノテーションについて関連研究を交えつつ解説する。オフラインビデオアノテーションでは、筆者らの以前の研究も含まれている。オンラインビデオアノテーションとオフラインビデオアノテーションについてそれぞれの利点と問題点を指摘することによって、オンラインビデオアノテーションを行う理由を明確にする。

2.1　アノテーションとは

アノテーションとは、手元の辞書を引いてみると「注釈 」という意味であり、文章や画像・映像等につけた注釈を意味する。辞書的な意味としては、人間が人間のためにつけた注釈を意味することが多いが、本論文では、機械が人間のために、あるいは人間が機械のために、さらには機械が機械のためにつける注釈も含む。本論文でのアノテーションの意味は、コンテンツ（テキスト・文章・映像・音声・音楽・Webページなどのマルチメディアコンテンツ）に対して、その意味内容

や意味構造といった情報を記述し、コンテンツと関連付けることを意味する。コンテンツに対してアノテーションを付与すると、そのコンテンツの意味内容や構造を解析するための手助けとなり、機械的な処理だけでは十分な精度が得られないような処理が高精度で可能になることが期待できる。アノテーションの概念図を図 2.1 に示す。

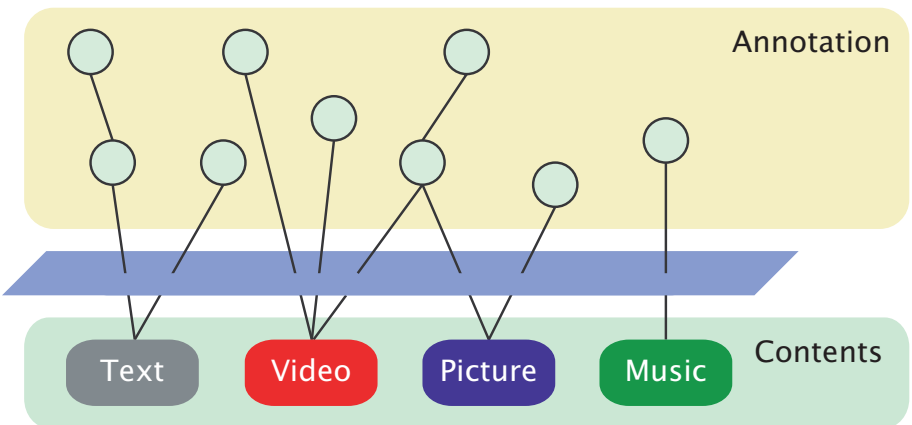

図 2.1: アノテーションの概念図

テキストに対して構文情報等のアノテーションを行う枠組みとしては、GDA (Global Document Annotation)[22] がある。GDA は、テキストに詳細な構文や意味などの情報を XML で記述を用いて記述する国際標準規格である。例えば、GDA において”Time flies like an arrow.”という文は以下のように記述される。

```
<su>
  <np sem="time0">time </np>
  <v sem="fly1">flies </v>
  <adp>like <np>an arrow</np></adp>.
</su>
```

同様に、マルチメディアコンテンツに対して詳細な意味内容情報を記述するアノテーションの国際規格として、MPEG-7[8] がある。カット情報・音声情報・著作権情報・各種認識結果などを XML で記述することが可能である。MPEG-7 の例を図 2.2 に示す。

```
<?xml version="1.0" encoding="Shift_JIS" ?>
- <Mpeg7 type="complete" xmlns="http://www.mpeg7.org./2001/MPEG-7_Schema"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xml:lang="ja">
  - <Description xsi:type="ContentEntityType">
    - <DescriptionMetadata>
        <Version>1.0</Version>
        <LastUpdate>2002-05-30T13:00:00+00:00</LastUpdate>
      - <Comment>
          <FreeTextAnnotation>シナリオから作成</FreeTextAnnotation>
        </Comment>
        <PrivateIdentifier>documentary-metadata</PrivateIdentifier>
      - <Creator>
        - <Role href="urn:mpeg:mpeg7:cs:RoleCS:2001:AUTHOR">
            <Name>作成者</Name>
          </Role>
        - <Agent xsi:type="OrganizationType">
            <Name>PRMU VDB-WG</Name>
          - <Contact xsi:type="PersonType">
            - <Name>
                <GivenName>智幸</GivenName>
                <FamilyName>八尾</FamilyName>
              </Name>
            - <ElectronicAddress>
                <Email>tomo-yao@ist.osaka-u.ac.jp</Email>
              </ElectronicAddress>
            </Contact>
          - <ElectronicAddress>
              <Uri>http://www.ieice.or.jp/PRMU</Uri>
            </ElectronicAddress>
          </Agent>
        </Creator>
        <CreationTime>2002-05-30T13:00:00+00:00</CreationTime>
      </DescriptionMetadata>
    - <MultimediaContent xsi:type="AudioVisualType" id="multimediaContent-1">
      - <AudioVisual id="audiovisual-1">
        - <MediaInformation id="mediaInformation-1">
          - <MediaProfile id="mediaProfile-1">
```

図 2.2: MPEG-7 の記述例 (一部抜粋)

ビデオコンテンツに対するアノテーションを一般にビデオアノテーションと呼ぶ。筆者は、ビデオアノテーションには、オンラインビデオアノテーションとオフラインビデオアノテーションの 2 種類のアプローチがあると考えている。そこで、それぞれについて以下の節で説明する。

2.2 オフラインビデオアノテーション

本節では、まずビデオアノテーションについて解説する。オフラインビデオアノテーションとは、ローカルにあるビデオコンテンツに対して、特定のアノテータがアノテーションを行うビデオアノテーションである。例えば、放送局が所有する膨大の数のビデオコンテンツを効率よく検索するためには、それぞれのコンテンツに対してなんらかのアノテーションが必須であると考えられている。そのため、従来から広く研究されてきた分野であり様々な研究成果が報告されている。まず始めに、筆者らが作成したビデオアノテーションエディタについて解説を行った後に、関連する研究について報告する。

2.2.1 ビデオアノテーションエディタ

オフラインビデオアノテーションを行う例として、筆者らが卒業論文で作成したビデオアノテーションエディタ [24] がある。ビデオアノテーションエディタは、オフラインビデオコンテンツに対してのアノテーションを効率よく行うためのツールである。MPEG ビデオなどの動画コンテンツに対し、いかにして簡単かつ十分なアノテーションを行うかについて議論し、さらにその支援ツールおよび応用システムの実装を目的とする。具体的には、本研究で開発されたツールをもとにして得られたアノテーションデータを用いた、自然言語での意味的ビデオ検索も実現している。

本ツールによって作成される XML アノテーションデータには、コンピュータによって自動解析されるデータと、ユーザがツールを使って付与されるデータが含まれている。コンピュータによって自動解析されるデータには、動画像に含まれるカットの時間位置や、カラーヒストグラム情報、オブジェクトトラッキング情報などがある。ユーザーがツールを使って付与するデータには、シーンやオブジェクトに対する意味情報をあらわす意味属性情報やコメント情報などが含まれる。さらに、それぞれの意味属性は検索や要約に便利な意味内容も定義されている。つまり、解析されたデータとともに、そのデータの意味する内容も記述されており、アノテーションデータ単独で、意味内容が理解できるような仕組みになっている。

そのために特定のアプリケーションに依存しにくく、永続性・独立性の高いアノテーションデータが得られるように工夫している。

ユーザが検索サーバに検索要求を出すと、検索サーバがアノテーションデータをもとにして検索を行い、検索結果をWebブラウザに表示する仕組みである。検索結果をユーザに適した形で表示するために、長尾らが開発したSemantic Transcoding[10]も利用することを検討している。

また、このアノテーションデータを元にXMLデータベースであるXindice[1]を用いてデータベースを構築し、Java Servletを用いて、従来は困難だと考えられていた自然言語による動画像検索システムをも試作し、アノテーションとその利用方法についても言及することにより、より実用的なビデオアノテーションについての提案をすることができた。

我々はテレビやインターネットのストリーミング配信・ビデオ・DVD等様々な形で動画コンテンツを見て楽しんでいる。しかしながら、膨大な動画コンテンツが存在しながらそれらを検索したり要約したりすることはできない。それは、動画コンテンツ自体は単なるバイナリデータであり、動画の意味する内容が記述されていないためである。その意味内容を記述する規格としてMPEG-7[8]が規格化されている。MPEG-7はMPEGによって規格化された動画コンテンツの意味内容を記述するための規格であり、動画圧縮を扱った既存のMPEG(MPEG-1,MPEG-2やMPEG-4)などとは違って、それ自身は動画のバイナリ情報を含んでいない。つまり、動画コンテンツに対するメタデータ（本研究ではアノテーションデータと呼んでいる）を記述する規格である。

そこで問題となるのは、いかに効率のよいアノテーション作成ツールを作るかということである。ビデオにアノテーションやインデックス情報等を付与するツールは様々なものが存在するが、Webに最適化されたアノテーションツールは少ない。MPEG-7記述ツール[6, 14]のように、XMLベースのツールは存在するが、高度に自動化されたツールはなく、大部分が手入力によるものであり、使い勝手が悪く試作段階のものでしかないという事実は否めない。そこで、本研究で、効率よくアノテーションをするツールとしてビデオアノテーションエディタ(図2.3)を試作した。このツール自体は長尾らが開発したビデオアノテーションエディタ[10]を元にしているが、新たに作り直したツールである。

図 2.3: ビデオアノテーションエディタ

ビデオアノテーションエディタの機能としては以下のものを備える。

1. カットの自動検出と修正機能
2. カット・シーンの重要度の推定と修正機能
3. カットの階層構造の推定と修正機能
4. カットに対する選択式手動アノテーション
5. オブジェクトトラッキング
6. オブジェクトに対する選択式手動アノテーション
7. アノテーター情報の付与

8. コンテンツ情報の付与

9. 色ヒストグラムの自動抽出

10. XML 形式での保存・出力

オブジェクトトラッキングやカット検出、ヒストグラムの自動抽出等はほぼ完全に自動化されている。もちろん、ヒストグラム以外は誤検出も考えられるので、直感にあった修正機能もつけている。本来ならば、音声認識部分にも対応している必要があるが、現状ではサポートされていない。読み込める動画コンテンツの種類は、Microsoft の DirectX8 に含まれる DirectShow[1] でサポートされる形式であり、MPEG-1, MPEG-2, MPEG-4 などに対応している。

これらの仕組みやツールを用いてローカルに存在するビデオコンテンツに対してアノテーションをある程度効率よく行うことが可能になる。

2.2.2 関連研究

オフラインビデオアノテーションに関連する研究についていくつか紹介をする。

まず始めに、本研究の元となった研究に長尾らが開発した VAE (Video Annotation Editor)[10] がある。これは、MPEG コンテンツに対し XML 形式のアノテーションデータを付与するツールであり、IBM の ViaVoice による音声認識や、カット検出・オブジェクトトラッキングなどを行うことができるアノテーションツールである。

同様なツールとして、リコーが開発した MovieTool[14] というツールが存在する。これは MPEG-7 記述用のビデオアノテーション用ツールである。 MPEG-7 を記述するために開発されたものであり、ビデオコンテンツに対し容易にアノテーションを行うことが可能である。特に、ビデオコンテンツをストーリ・シーン・カットといった階層構造表現も含めた MPEG-7 データを記述することが可能である。ま

[1] Microsoft 社の Windows 用マルチメディア拡張 API 群である DirectX に含まれる API の一つ。DirectShow を使うことで、動画や音声などの大容量のマルチメディアデータをストリーミング再生 (データを読み込みながら同時に再生すること) することができるようになる。DirectShow を利用すれば、ハードウェアの違いを気にすることなく、動画や音声などの再生制御をアプリケーションソフトから手軽に行えるようになる。

た、MPEG-7 スキーマを動的に読み取り解析する仕組みであるために、MPEG-7 記述の編集において利用可能なタグ候補を表示する機能や、MPEG-7 の文法を定義した MPEG-7 スキーマと MPEG-7 記述との整合性チェックを行える、MPEG-7 スキーマで定義されているすべてのメタデータを記述できる、MPEG-7 スキーマの動的読み込みにより今後のスキーマの変更や個別の拡張スキーマにも対応できるなど、MPEG-7 の手作業での記述の支援も行える。

また、IBM が開発した MPEG-7 対応アノテーションツールである VideoAnnEx Annotation Tool[6] がある。これは、動画コンテンツに対して、Static Scene, Key Objects, Event の情報を各シーンに対して意味属性を振ることができるツールである。Windows 標準のツリービューを利用して、あらかじめ登録されている ID から複数選択したり、 ID を増やしたりすることができるツールである。ツリービューを利用することにより意味属性をカテゴリ分けできる利点がある一方、ツリーが大きくなるにつれどこにどの意味属性があるのか分かりにくく、また、ツリーの一部分を閉じてしまうと、そのツリー以下に意味属性のチェックが入っていても分かりにくいという弱点がある。

さらに、ビデオアノテーションに対する取り組みに対して古典的な例として、MIT(マサチューセッツ工科大学) のメディアラボでの Daivs ら [4] によって開発された MediaStream は、アイコンによるビジュアル言語によって映像を構造化するようになっている。このツールの特徴は、3000 ものアイコンとその組み合わせによってビデオのオブジェクトにアノテーションすることができる点である。検索を考慮する場合、ビデオへの注釈を自然言語でつけるよりもアイコンを使って意味属性をふった方が分かりやすいためにこの方式を利用したと考えられる。このツールの難点は、映像をアイコンで表すために、非常に多くのアイコンが必要となる点である。開発の最終段階にはアイコンの数が 6000 個にもなったといい、アイコンは種類分けされてはいるもののアイコンの選択には慣れが必要であるし、そのアイコンの意味するものがアイコンだけでは分かりにくいという欠点があると考えられる。さらに、複雑なオブジェクトには複数のアイコンを組み合わせる必要があるので、複雑なオブジェクトを表現できる一方、煩雑な作業が必要になると考えられる。

以上の関連研究はいずれも手作業を伴うシステムであるが、カーネギーメロン大

学の Informedia プロジェクト [18] では完全自動でビデオアノテーションを行うシステムを開発している。画像認識、音声認識、自然言語処理技術を統合して 1000 時間ものビデオ映像の自動索引付けを行うシステムである。クローズドキャプションの利用、文字・音声の自動認識を駆使してキーワード索引の自動生成を行い、キーワードにもとづくビデオ映像検索が可能になっている。また、tf-idf 法 (term frequency/inverse document frequency)[16] を用いてビデオ映像の要約も実現している。実際に 1000 時間もの動画像に対して索引付けを行っており、非常に大規模なシステムでの自動認識と検索・要約を行っている点が興味深い。

また、アノテーションされたコンテンツの検索を効率よく行う仕組みも提案されている。是津ら [23] が開発した時点モデルによる映像区間の合成 : 時刻付きオーサリンググラフに関する研究では、各時点において、テキストにより映像のキーワードを記述する方式でアノテーションを行っている。テキストを記述するだけでは複雑な検索に不向きであるが、それぞれのキーワード間に互いに関係のあるものに対して、時刻印付きオーサリンググラフという無向グラフを記述することによりその問題を解決している。さらに、その関連付けはテキストに含まれるキーワードの類似性を用いて自動的に関連付けを行っている。 このグラフを用いて、自然言語検索文による検索が行えるようになっている。検索アルゴリズムは自然言語検索文に含まれるキーワードとノードに含まれるキーワードの類似計算を行って関連のあるノードを見つける。そして、それらのノードにおいて時間的つながりを考慮しつつ、極小部分グラフを求める。この極小部分グラフが求める検索結果である。

2.2.3 問題点

以上のようにオフラインコンテンツに対するアノテーションにはいくつかの例があることが分かった。筆者らが卒業研究で作成したビデオアノテーションエディタや MIT の MediaStream などの研究例においては、人間の手作業がある程度介在したアノテーションの手法である。詳細で確実なアノテーションができることが期待できる反面、非常に大きな人的コストがかかってしまうのが現状である。また、Informedia のようなビデオ映像に対して自動索引付けを行うようなシステム

については、音声認識や画像認識などの精度に強く依存してしまう。そのために、人的コストはかからないが、十分な精度が得られないといった問題がある。

ビデオコンテンツの解釈や認識を行う場合、機械では十分な精度が得られないという問題点がある一方、人手では作業できる量に限界があるという根本的な問題点を含んでいる。もちろん、貴重なコンテンツなどアノテーションコストを上回るほどの利益が得られる場合にはアノテーションをする価値があるため、これらの研究を続けていく意義はあると考えられる。

2.3　オンラインビデオアノテーション

前節までは主にオフラインでのマルチメディアアノテーションについて解説した。本節では、オンラインでのアノテーションについて解説する。オンラインアノテーションとは、ネットワーク特にインターネットを通じて行うアノテーション全般のことをいう。オフラインアノテーションの場合は、専属のアノテータが一人もしくは少人数で時間をかけて行う必要があるが、オンラインアノテーションではその制約がない。つまり、複数人のアノテータがネットワークを通して一つのコンテンツにアノテーションを行うことによって、一人当たりのアノテーションコストを極小化しようとする試みである。アノテーションには多大な人的コストがかかる場合が多く、また、コンテンツに対する詳細な知識を必要とする場合も多い。そのため、アノテータが一人で一つのコンテンツに対してアノテーションするよりも、複数人で協力して個々の知識を統合する方が的確で効率よくアノテーションを作成することが可能になる。さらに、コンテンツ閲覧者やユーザにアノテーションを促す仕組みを用意することができれば、より多くの知識や意見を集約でき、より詳細なアノテーションを行える可能性が高い。具体的には、コンテンツ閲覧画面とアノテーションのインタフェースを統合することによって気軽にアノテーションを行える環境を用意すればよい。このようにオンラインビデオアノテーションは、オフラインビデオアノテーションの欠点を補うことができると考えている。

ここでは、まず、オンラインビデオコンテンツとは何かという議論からはじめる。

2.3.1 オンラインビデオコンテンツとその関連技術

本論文では、オンラインビデオコンテンツとは、インターネットを通じて配信されるビデオコンテンツ一般を指す。

ハードディスクの大容量化や携帯電話等の簡便な動画記録デバイスの大衆化、さらにはブロードバンド環境の普及に伴い、インターネット上に多数の動画コンテンツが公開されるなど、ネットワークを通してデジタルビデオコンテンツに容易にアクセス可能な環境が整備されつつある。

それに加え、ビデオコンテンツを配信・再生する媒体・手段も多様化している。従来型のコンテンツ配信は、一部の限られたコンテンツプロバイダが、放送電波やDVDなどの流通に載せて行うものであった。現在は、インターネットの発達と共に誰もがコンテンツを容易に配信可能な環境を構築可能である。また、携帯電話やPDAなどの個人用小型端末やPCなどでネットワークからコンテンツを閲覧するという環境が整いつつある。これらの端末はインターネットに常時接続が可能であり、視聴者からのフィードバック情報を容易に取得が可能である。さらには、コンテンツプロバイダと視聴者間の1対1の通信だけでなく、視聴者対視聴者間通信も可能である。そのために、視聴者間だけでアノテーション情報やメタ情報を共有するという草の根的な活動も実現可能である。このことは、従来型のコンテンツ配信の形態が大きく変化する可能性を示唆していると言える。

また、コンピュータの計算能力の向上と共に、ビデオコンテンツのフォーマットも多様化している。ビデオCDやPCなどで採用されているMPEG-1、DVDやデジタル衛星放送などに利用されているMPEG-2、インターネット上で広く利用されているMPEG-4などがある。MPEG-4にもいくつかの派生があり、QuickTime形式、RealVideo形式、Windows Media Video形式、DivX形式などがある。さらに圧縮効率を高めた、MPEG-4の追加規格としてH.264[2]という形式も標準化されつつある。これらの形式にも、それぞれダウンロード型配信とストリーミング型配信の二種類が存在する。ダウンロード型配信では、コンテンツの全てもしくは一部分をダウンロードしてから配信する形式であり、ストリーミング型配信ではダウンロードしながらコンテンツの配信を行う形式である。また、デジタルデータの著作権を保護する技術として、Digital Rights Management (DRM)なども標

[2]ISO/IEC 14496-10 MPEG-4 Part10 Advanced Video Coding/ITU-T Rec. H.264

準化されつつある。これには、音声・映像ファイルにかけられる複製の制限技術などといったメータデータレベルのものから、画像や動画ファイルの電子透かしなどの技術も採用されている。

また、ネットワーク上で効率よくコンテンツ配信を行うための技術としてQuality of Service (QoS) などもある。これは、リアルタイム性を必要とするパケットを優先的に処理する技術である。さまざまな通信インフラが混在するインターネット上でQoSを実現するため、標準プロトコルRSVPの策定などの技術開発がすすんでいる。このように、着々とインターネットでコンテンツ配信を行うための環境が整いつつある。

また、ビデオコンテンツをそのまま知的処理するのは困難なため、計算機に分かりやすいメタ情報をつけるという考え方も一般化してきている。そのための規格としてMPEG-7などがある。MPEG-7は、映像の圧縮方式の規格を制定したMPEG-1/2/4などとは全く違いメタデータの記述形式である。MPEG-7はXMLで記述されており、インターネットやSemantic Web[17] などの技術と親和性が高い。

また、TV放送に関連するメタ情報として、iEPG (Internet Electronic ProgramGuide) などがある。これは、テレビ番組情報をインターネット上から取得可能でありTV放送コンテンツを一意に識別する手段となる。

近い将来、非常に多くのビデオコンテンツがネットワークを通して配信されるようになると考えられる。そのために、我々はオンラインビデオコンテンツの存在を無視することはできない。そこで、本研究では、このようなオンラインビデオコンテンツが日常的に利用可能である世界を想定している。

2.3.2 オンラインビデオアノテーションとは

オンラインビデオアノテーションとは、ネットワーク上に存在しているビデオコンテンツに対してアノテーションを行うことをいう。つまり、ネットワーク上に存在しているビデオコンテンツの一部分とアノテーション情報を関連付けることによって高度な知的処理を実現しようとするものである。オンラインビデオアノテーションの特徴としては、インターネットなどのネットワークに接続可能な人ならば誰でもアノテーションに参加することが可能な点が上げられる。例えば、

閲覧者自身にアノテーションを促す環境を構築すれば、多くの人からのアノテーションを期待できるなど利点が大きい。

一般に、ビデオコンテンツへのアノテーションには多大な人的コストがかかり、ビデオ解析処理の高速化や精度の向上を図ってもそれほど改善されない。それは、ビデオコンテンツの深い意味情報付与には人間の高度な判断や解釈を必要とするからである。そのため、オンラインビデオアノテーションでは、人的資源の不足を補うという観点からいって有用であると言える。

しかしながら、オンラインビデオアノテーションにもいくつかの解決しなければ問題が存在する。

まず、アノテーション情報の量を確保すると同時に情報の質を確保する必要があるという一見相反する問題があげられる。閲覧者による映像コンテンツへのリアルタイム掲示板書き込みの例として、大規模匿名掲示板である 2 ちゃんねるの実況板がある。人気のある TV コンテンツに対しては 1 分間に数十から百以上もの書き込みがあることは珍しくなく、このような匿名掲示板システムならば、アノテーションの量を確保できる可能性が高い。しかしながら、信頼性の高い良質な情報もあるが多くは信頼性の低い情報や日本語として正しくないものであり質は低い。そのために、いかにアノテーションの選別を行うかという問題がある。後に紹介する iVAS[26] では、個々のアノテーションに対してユーザからのフィードバックを元にしたアノテーション信頼度という概念を導入することによりその問題の解決を図っている。

さらに、複数のアノテータのアノテーション情報をいかに統合するかという問題もある。Pradhan らの研究 [13] では、複数のアノテータによるビデオアノテーション間に、矛盾や不完全性が認められた場合に、これらを抽象化することで解消を図る手法を提案している。iVAS では、アノテーション信頼度付きアノテーション情報を基にして確率的な処理を行うことができ、より現実的な解決法が得られるのではないかと考えている。

2.3.3 オンラインビデオアノテーションの仕組み

オンラインアノテーション、つまりネットワークからマルチメディアコンテンツに対してアノテーションを行う仕組みについて解説する。まず、コンテンツは、

ネットワークからアクセス可能な領域に置かれていると仮定する。そのコンテンツが置いてあるURL情報やコンテンツのタイトル情報、コンテンツIDなどをキーとして、コンテンツを一意に識別できるものとする。ユーザはそのコンテンツのキー及びコンテンツのメディア時間と、アノテーションの対象となる情報とを関連付けることによってアノテーションを行う。アノテーション情報には、誰が、いつ、どのコンテンツの、どの部分に対して、どのような方式でアノテーションしたかという情報も含まれており、他のユーザのアノテーション情報も含めて共有する。

Webコンテンツに対するアノテーション及びそれを共有する仕組みとして、広津[5]らは以下の三つの方式を提案している。

1. Client model
2. Proxy model
3. Server model

Client modelはクライアント側のWebブラウザにアノテーション情報を保持し、クライアント間でデータを共有する方式である。既存のインターネットを変更することなく実現できるという利点があるものの、クライアント側に専用のウェブブラウザを用意する必要があり、またアノテーションデータの共有が比較的困難であるという欠点がある。Server modelは、サーバ側にアノテーションのための仕組みを用意する方式で、クライント側に特殊な仕組みが必要ないという利点があるものの、コンテンツがあるサーバごとにアノテーションのための仕組みを用意する必要があり、また全てのユーザをサーバごとに認識する必要があるという欠点がある。そこで、広津らはProxy modelを推薦している。Proxy modelは、ウェブコンテンツとユーザグループの間にアノテーションのためのプロキシサーバを用意することにより、プロキシサーバでアノテーション情報を管理する方式である。クライアント側にも既存のサーバ側にもアノテーションのための仕組みを用意する必要がないという利点があると述べている。

ただし、Proxy modelにもいくつか解決しなければならない問題点がある。まず、アノテーションあるなしに関わらず全てのコンテンツがプロキシサーバを経由

するためプロキシサーバに大きな負荷をかけること、次に、動的にプロキシサーバを切り替えることが困難なため、一つのプロキシサーバで全ての Web コンテンツのアノテーション情報を管理しなければならないこと、さらに、プロキシサーバはユーザグループ毎に設定する仕組みであるが、アノテーション情報の二次利用を考える場合には、コンテンツごとにアノテーション情報を管理した方が効率が良いことなどが挙げられる。

いずれの方法にせよ、コンテンツの一部分とアノテーション情報を一対一に関連付け、それをユーザ間で共有可能にすることによって、オンラインアノテーションを行うことができる。

2.3.4 関連研究

最後に、オンラインビデオアノテーションに関連する研究をいくつか紹介する。ただし、筆者が調べた限りにおいて、筆者が想定するような形でのオンラインビデオアノテーションは未だ存在しない。そこで、まず、オンライン上でビデオコンテンツに対して何らかの議論や投稿が可能な関連研究について報告する。

まず、動画コンテンツに対して電子掲示板感覚で情報を付与することができるシステムとして SceneNavi[20] がある。これは、VOD (Video on Demand) など、インターネット上に存在している動画を閲覧しているコミュニティの間でコミュニケーションを図るツールである。動画上の任意の時点で閲覧者のコメントを関連付けるという点で、オンラインアノテーションの一つであるといえる。これらは全て一般的な Web ブラウザを用いて行うことができる。また、投稿された情報を基にした検索などの応用等も考えられているようである。

次に、ストリーミングメディアコンテンツに対してアノテーションを行うシステムとして MRAS (Microsoft Research Annotation System)[2] がある。これは、グループ間でストリーミングメディアコンテンツに対してアノテーションを行うシステムである。任意の時点の映像に対してコメントやそのコメントに対する回答などを投稿することが可能である。さらに講義映像向けシステム [3] では、Web ブラウザベースで PowerPoint のスライド情報なども表示可能であり、参加者や TA (Teaching Assistant) らが質問の回答などを投稿できるようになっている。

第3章　閲覧者によるオンラインビデオアノテーション

前章で、ビデオアノテーションに関して、オフラインビデオアノテーションとオンラインビデオアノテーションの二つのアプローチがあることを示した。本研究ではアノテーションコストを軽減できる点やビデオコンテンツとそのコミュニティの発展にもつながるという点から、オンラインビデオアノテーションの仕組みを採用した。そこで、本章では、コンテンツ閲覧者にWebブラウザを通してアノテーション作業に参加してもらう仕組み、および複数のアノテーション作成手法を提案する。

近年、ハードディスクの大容量化や携帯電話等の簡便な動画記録デバイスの大衆化、さらにはブロードバンド環境の普及に伴い、インターネット上に多数の動画コンテンツが公開されるなど、ネットワークを通してデジタルビデオコンテンツに容易にアクセス可能な環境が整備されつつある。さらには、デジタルビデオカメラの普及や動画編集ソフトウェアの低価格化などにより膨大なビデオコンテンツが氾濫することが予想される。それに伴い、ビデオコンテンツの意味的な検索や要約などを行いたいという欲求は日に日に高まっていると考えられる。

本研究では、動画の場合、多くの閲覧者を獲得することが比較的容易であるという点に着目し、その閲覧者からのフィードバックやアノテーションに参加してもらえる環境を整備すれば、より多くのアノテーション情報が集まるのではないかと考えた。そこで、一般的なWebブラウザを用いて、閲覧者による簡単かつ負担の少ない手段で動画に対してアノテーションを行うシステムが有用ではないかと考えた。この方法だと、たとえ一人当たりのアノテーションの量が少なくても、複数の閲覧者のアノテーション結果を融合させることにより、全体として高度なアノテーションとその活用（検索・要約など）が実現できるのではないかと考え

た。そのために、コンテンツを閲覧しつつついでにビデオコンテンツの時間軸に沿って電子掲示板感覚でアノテーションを行うシステムが有用であると考えた。

また、本システムによって得られたアノテーションを用いた応用例として、ビデオコンテンツの意味的検索および簡約をするシステムを試作した。また、アノテーションの副次的利用として、アノテーション情報をユーザ間で共有することにより、ビデオコンテンツを中心としたコミュニティ形成の手段の一つとして利用できるのではないかとも考えた。それを発展させることによって、新しい広告媒体として利用可能なシステムへの足掛かりとして利用できると考えている。

3.1　閲覧者によるビデオアノテーションとその問題

ここでは、閲覧者によるビデオアノテーションについての解説と、その問題点について述べる。一般に、ビデオコンテンツにアノテーションを行う方法には様々なものが考えられるが、筆者は大きく分けて以下の3種類があると考えている。

- 自動ビデオアノテーション
- 半自動ビデオアノテーション
- オンラインビデオアノテーション

一つ目の自動ビデオアノテーションとは、全く人間の手作業が介在しない方法であり、人的コストが極めて低いため、大量の映像をアノテーションするのに適している。自動ビデオアノテーションの例としては、音声認識技術や画像認識技術を用いたInformedia[18]や、画像認識技術とオブジェクト指向型状態遷移モデルを組み合わせた研究[31]など様々な研究が行われている。自動ビデオアノテーションは人間の手が介在しないためにアノテーションコストが低いという利点がある一方、コンテンツの種類によって解析手法や解析精度が異なるため一般的なコンテンツに対して適用することが困難だという欠点がある。

二つ目の半自動ビデオアノテーションとは、ビデオコンテンツに対してアノテーションを行う場合、人間がアノテーション作業の一部に参加するアノテーションである。前章で述べたように、半自動ビデオアノテーションの例としては、専用の

ツールを用いてアノテーションを行う研究がいくつか存在する [4, 14, 6, 11, 10, 9]。人間の手が介在する分、アノテーションコストがかかるものの、意味的な情報を付与することができるのは大きな利点である。筆者らも以前、カット検出やオブジェクトトラッキング、音声認識などの技術を用いたアノテーションエディタの試作を行った [25] が、詳細なアノテーションを行うためには、コンテンツの長さの数倍もの時間がかかるのが現状である。また、アノテーション情報をビデオ区間の包含関係に基づいて自動継承する OVID モデル [12] などもアノテーションコストを下げる点において有意義な方法である。

そして三つ目のオンラインビデオアノテーションとは、アノテーションに必要な情報をネットワークを通じてリアルタイムに収集し応用に反映させる手段である。これも前章で述べたように、他メディアと放送コンテンツをリンクさせ最新の情報を得る研究 [27, 30] や、動画像コンテンツに対して電子掲示板感覚で情報を付与する SceneNavi[20] などの研究が存在するが、アノテーションとしての利用や、その応用、さらには後述する問題などに対処できていない。

そこで本論文では、オンラインビデオアノテーションとして、閲覧者よりオンラインで情報を収集しアノテーションとして反映させる、閲覧者によるビデオアノテーションシステムを提案する。

一般に、ビデオコンテンツへのアノテーションには多大な人的コストがかかり、各種ビデオ解析処理の高速化や精度の向上を図ってもそれほど改善されない。それは、ビデオコンテンツの深い意味情報付与には人間の高度な判断や解釈を必要とするからである。そこで本研究では、人的資源の不足を補うために、閲覧者が閲覧時に比較的負担の少ないやり方でアノテーションを行う仕組みを提案する。つまり、ビデオコンテンツを解析するという問題を、複数のユーザからのアノテーション情報の収集と解析という問題に帰着させることにより、効率よく動画を解析しようとする試みである。しかも、電子掲示板感覚なインタフェースを採用することにより、閲覧者による自発的なアノテーションが促され、人的なアノテーションコストを最小化する試みである。

ここで、閲覧者によるビデオアノテーションを行う上で解決すべき問題はいくつかあるが、まず、アノテーション情報の量を確保すると同時に情報の質を確保する必要があるという一見相反する問題があげられる。閲覧者による映像コンテ

ンツへのリアルタイム掲示板書き込みの例として、大規模匿名掲示板である2ちゃんねるの実況板がある。人気のあるTVコンテンツに対しては1分間に数十から百以上もの書き込みがあることは珍しくなく、このような匿名掲示板システムならば、アノテーションの量を確保できる可能性が高い。しかしながら、信頼性の高い良質な情報もあるが多くは信頼性の低い情報や日本語として正しくないものであり質は低い。そこで、本システムでは、個々のアノテーションに対してアノテーション信頼度という指標を導入し、アノテーション情報の選別を行うことによって、この問題の解決を図る。

さらに、複数のアノテータのアノテーション情報をいかに統合するかという問題もある。Pradhanらの研究[13]では、複数のアノテータによるビデオアノテーション間に、矛盾や不完全性が認められた場合に、これらを抽象化することで解消を図る手法を提案しているが、本システムの場合、アノテーション記述内容の信頼性が低いものが多くあると想定されるため、直接この仕組みを適用することが困難である。この問題にもアノテーション信頼度を用いれば、信頼度付きアノテーション情報を基にして確率的な処理を行うことができ、より現実的な解決法が得られるのではないかと考えている。

また、本システムでは、アノテーションが増えれば、逐次反映され、検索などの応用例の精度向上に結びつき、この点でもオンラインであることの利点が活かせると思われる。

最後に、アノテーションによって得られた情報を用いて、様々な応用例を考察し、アノテーションの有用性を示すことで、従来研究からの差別化を図ることができる。

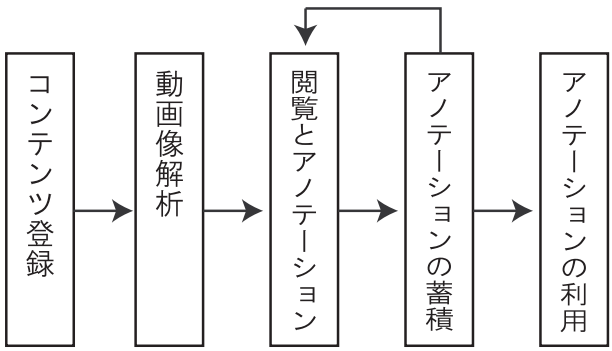

図 3.1: 処理の流れ

3.2 オンラインアノテーションシステムの構築

閲覧者によるアノテーションの有用性と、前章で述べた問題に対する解決法を検証するために、閲覧者によるビデオアノテーションシステムである iVAS(intelligent Video Annotation Server) を構築した。iVAS は単なるアノテーションをするためだけのツールではなく、アノテーションの蓄積・アノテーションの解析を行うことにより、ビデオ検索などのアノテーションを用いた各種応用も行う統合システムである。また、Proxy Model を採用するために、既存の Web サーバやクライアントのブラウザを一切変更することなく、インターネット上に存在するすべてのビデオコンテンツを対象として機能するシステムである。

本システムが行う具体的な処理の流れを図 3.1 に示す。まず、対象となるビデオコンテンツの登録を行う。登録されたビデオコンテンツはカット検出サーバによりカット検出や色ヒストグラム抽出などの動画像解析を行う。解析された情報を元にして、アノテーション編集ページの生成を行う。アノテーション編集ページでは、コンテンツの閲覧とともに各種アノテーションを行うインタフェースを提供する。アノテーション編集ページで収集された情報は XML データベースに蓄積され、ビデオコンテンツの検索や要約などの各種応用に利用される。さらに、ユーザは他のコンテンツを検索結果などから探し出し、閲覧及びアノテーションを再帰的に行うことができる。このような仕組みにより、アノテーションの蓄積とそ

の応用が可能になる。

3.2.1 システム構成

本システムは、図3.2に示すように、アノテーションXMLデータベース、カット検出サーバ、登録・収集サーバ、ビデオアノテーションサーバ、アノテーションを用いた各種サービスサーバで構成される。ユーザは、ネットワークからアクセス可能な任意のビデオコンテンツに対して、iVASを通してアノテーション及び閲覧を行うこととする。また、ユーザは匿名、記名に関わらずアノテーションできる仕組みとなっている。iVASを通して閲覧したいコンテンツは、登録サーバを用いて明示的に登録する必要があるが、将来的にはコンテンツ収集サーバを用いて自動収集することを考えている。コンテンツを登録すると直ちに、カット検出サーバが連動しカット検出やヒストグラム情報の取得などの自動処理が行われる。閲覧者はビデオアノテーションサーバによって生成されたページを通してコンテンツを閲覧しながらアノテーションを投稿することができる。投稿されたアノテーションはアノテーションXMLデータベースに蓄積され、4章で述べる各種アノテーションを利用したサービスなどで利用される。アノテーションシステムとしては、Proxy Model[5]に相当する。

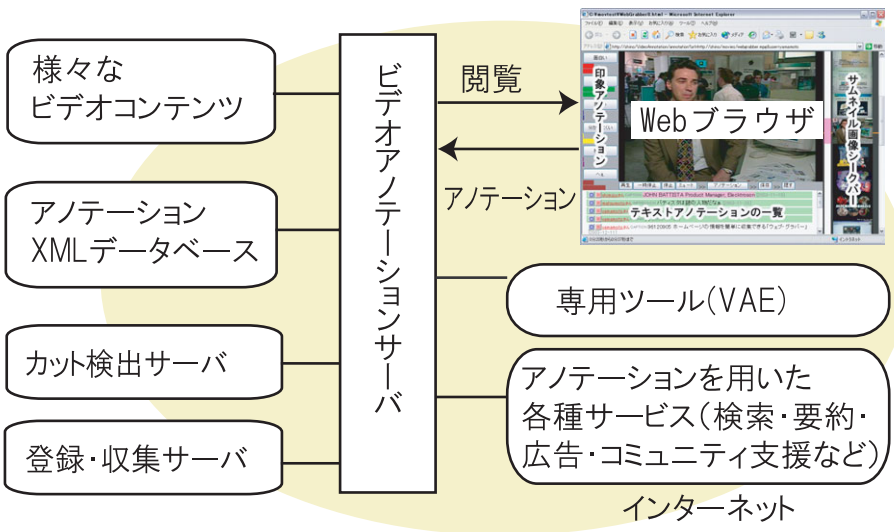

図 3.2: システム構成

3.2.2 想定するビデオコンテンツ

アノテーションが可能なビデオコンテンツは、PC からアクセスできるデジタルビデオコンテンツである。基本的には、インターネット上で公開されている Web ビデオコンテンツを対象とするが、個人的に HDD ビデオレコーダなどで大量かつ無作為に録り貯めた TV 映像などのホームネットワークに存在するコンテンツ、あるいは DVD などのパッケージメディアコンテンツなどにも適用可能であると (図 3.3)。つまり、コンテンツを一意に特定できさえすればよくコンテンツの形態によらない。アノテーションデータは Web 上で共有されるため、インターネットに接続可能な環境であれば良い。また、本システムはメタ情報のみを編集・蓄積するものであり、オリジナルコンテンツの編集・コピー・再配信を行うものではないため、著作権的な問題も発生しにくいと考えている。

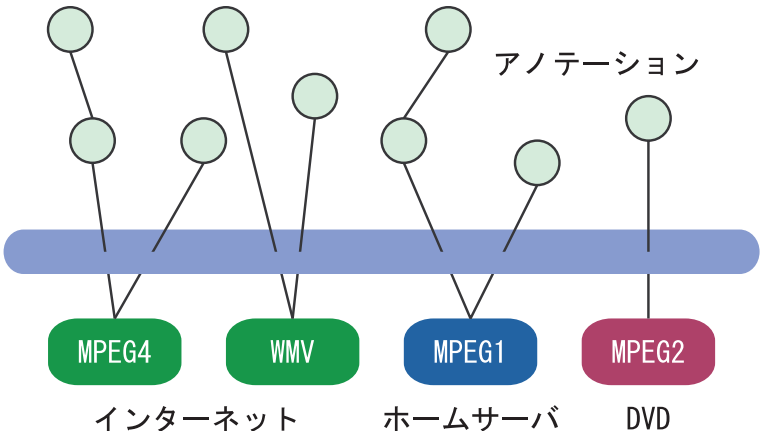

図 3.3: アノテーションを想定するビデオコンテンツ

3.2.3　コンテンツ登録

ビデオコンテンツは、コンテンツ登録サーバを通してあらかじめ解析を行う。コンテンツ登録時に、テンプレートの選択によるアノテーション作成環境を設定でき、コンテンツに適したアノテーションを作成することができる。具体的には以下のような項目を登録可能である。コンテンツ登録ページを図 3.4 に示す。

1. コンテンツのタイトル
2. コンテンツの URL
3. コンテンツの種類
4. 後述する印象アノテーションのボタンの種類
5. コンテンツの登録者の名前の E-mail アドレス

これらは手動により登録を行っているが、将来的にはコンテンツ収集サーバにより Web 上のコンテンツを RSS (RDF Site Summary)[15] などの情報を元にして、自動収集することによりコンテンツの自動登録も検討している。

投稿されたコンテンツは以下のような XML の形式で保存される。

```
<?xml version="1.0" encoding="x-sjis-unicode" ?>
<videoDB id="va00000006">
  <datetime>
    <date>2004-11-04</date>
    <time>16:40:00</time>
  </datetime>
  <media>
    <filename>news2_mini.mpg</filename>
    <url>http://fs02/ivas/movies/news2_mini.mpg</url>
    <title>ニュース映像</title>
  </media>
  <author>
    <name>山本大介</name>
    <email>yamamoto@nagao.nuie.nagoya-u.ac.jp</email>
  </author>
  <impSelection num="6">
    <item name="楽しい" no="1" />
    <item name="キター" no="2" />
    <item name="おいしそう" no="3" />
    <item name="悲しい" no="4" />
    <item name="嫌悪" no="5" />
    <item name="重要" no="6" />
  </impSelection>
</videoDB>
```

上記の例は、media 要素の中の url 要素に記述してあるサイトに存在するビデオコンテンツが、iVAS サーバにコンテンツ ID を va00000006 として登録されることを意味する。datetime 要素の子要素にコンテンツの登録日時、media 要素の子要素にコンテンツのタイトルなどのメディア情報、author 要素の子要素に登録者情報などが記述される。さらに後に述べる印象アノテーションのために使われる、このコンテンツに特化した印象ボタン情報を impSelection 要素の子要素に列挙する。

このような形式により、XML データベースに記述しコンテンツの管理を行う。

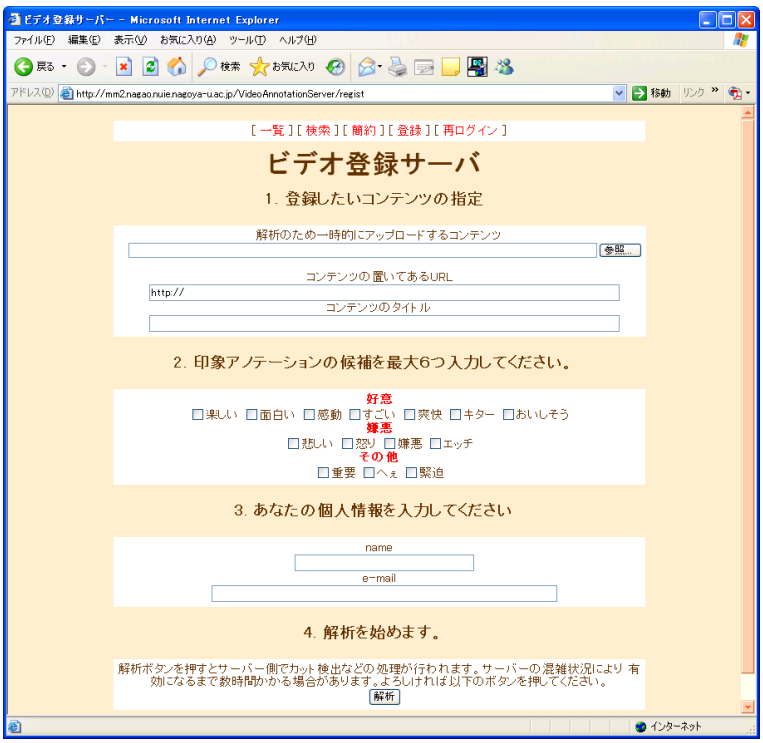

図 3.4: コンテンツ登録画面

3.2.4　動画像の解析

Web ブラウザを用いて動画のアノテーションを行う場合、インタラクティブに動画の解析処理をすることは処理速度などの点で好ましくないため、あらかじめ解析を行いたいコンテンツに対して前処理を行う必要がある。そのための事前処理としてカット検出を行い、動画からカットの時刻とサムネイル画像をサーバ上に保存するプログラムとしてカット検出サーバを用意した。

カット検出のアルゴリズムは、現在のフレーム間差分と直前のフレーム間の比較を行い、ある閾値を超えた時点でカットとした。また、フレーム間差分算出手法には、分割 χ 二乗検定法 [21] を採用した。また、フレーム間差分の変動が大きい区間（つまり動きの激しい区間）はひとつのカットと認識するなどの工夫をし

ている。

カット検出アルゴリズムとしては以下のようになる。

DirectShow の機能の一部としてある SampleGrabber の機能を用いて動画から静止画を 1 フレームごとにとりだして評価を行う。数フレーム置きに行っても良いが、MPEG などの圧縮形式は 1 フレームごとに前フレーム情報を利用してデコードする必要があるため、数フレームごとにやるのは効率がよくないためである。まず個々のフレームにおいてヒストグラムを生成する。本研究では RGB 要素それぞれ 4 ビット分合計 12 ビット、4096 分割して生成した。ヒストグラムでの前フレームとの比較により、ある閾値以上の差が生じた場合は新たなカットとして検出するプログラムである。さらに、動画の任意での位置で右クリックのコンテキストメニューを選択することにより、カット分割・カット併合をマウスで簡単に行えるようにもした。

カット検出の具体的なアルゴリズムは次のようになる。

1. DirectShow の SampleGrabber 機能を使い MPEG コンテンツから静止画をとりだす
2. RGB 要素をそれぞれ 4 ビット合計 12 ビット (4096) 分割をしてヒストグラムを計算する
3. 前フレームとのヒストグラムの要素ごとにの分割 χ 二乗検定法を用いて差分を求める
4. 3. の絶対値誤差の合計がある閾値以上になったら、カットであると認識し終了
5. メディア時間を 1 フレーム分進めたのち 1. に戻る

このままでもカットの検出は可能であるが、動きが激しいシーンではカットが大量に誤検出してしまう。そこで、動きが激しいカットをひとつのカットであると認識するために以下の手法を用いた。

1. 上記アルゴリズムでカット検出を行う

2. 1. を繰り返し、8 フレーム以内に再びカット検出を行ったら、それをカット検出とはせずに、2 へ

3. 1 へ飛ぶ

8 フレームとしたのは、30 フレーム/秒とすると、0.26 秒にあたり、人間の目では、これ以上短いカットはカットと認識しない限界であると考えたからである。

カット検出サーバの対象とするコンテンツは、カットが多く存在するコンテンツであり、サッカー中継や個人で撮影したホームビデオや講演ビデオなどカットの少ないコンテンツにはカット検出が適用できない。これらのコンテンツは、一定時間ごとに区切り、形式的なカットとし、アノテーションを行うこととする。

カット検出サーバによって取得された情報は以下のような XML 形式でデータベースに保存される。

```
<?xml version="1.0" encoding="x-sjis-unicode" ?>
<videoAnnotation id="va00000006">
  <media>
    <url>http://fs02/movie/news2_mini_2.mpg</url>
    <filename>news2_mini_2.mpg</filename>
    <length unit="sec">317.962444</length>
    <winsize height="240" width="352" />
  </media>
  <movie num="58">
    <cut etime="3.704" image="t000000.jpg" no="0" stime="0" />
    <cut etime="8.07437" image="t000001.jpg" no="1" stime="3.704" />
    <cut etime="14.782" image="t000002.jpg" no="2" stime="8.07437" />
    <cut etime="22.089" image="t000003.jpg" no="3" stime="14.782" />
    <cut etime="25.526" image="t000004.jpg" no="4" stime="22.089" />
    <cut etime="27.9944" image="t000005.jpg" no="5" stime="25.526" />
    ...
    <cut etime="317.962" image="t000057.jpg" no="57" stime="317.651" />
  </movie>
```

```
</videoAnnotation>
```

上の例では、コンテンツ ID va00000006 のコンテンツに対して行ったカット検出結果が movie 要素の子要素である cut 要素の形で列挙されている。カットと検出された時間を stime、そのカットが終了したと判断された時間を etime、さらにそのコンテンツのサムネイル画像を image 属性によって記述する。

3.3 閲覧者によるオンラインビデオコンテンツへのアノテーション

前節までに、オンラインビデオアノテーションを行うためのシステムを説明した。本節では、実際にユーザがどのようにビデオコンテンツに対してアノテーションを行うかどうかを説明する。閲覧者はコンテンツを閲覧するのと同時に、iVAS のアノテーション編集ページを用いて、テキスト入力を主としたアノテーション（テキストアノテーション）とマウスクリックを主にしたアノテーション（印象アノテーション）及びテキストアノテーションに対する○×評価（評価アノテーション）の 3 つのアノテーションを行うことができる。

また、アノテーションモデルとしては、代数ビデオ [19] の stratified model に相当する。

3.3.1 アノテーション編集ページ

アノテーション編集ページのブラウザは図 3.5 のようになる。画面左に印象アノテーションインタフェース、中央部上部に動画閲覧画面、画面中央下部にテキストアノテーションの一覧、右側にサムネイル画像を用いたスクロール可能なシークバーを配置した。

このシークバーは、マウスのスクロールボタンによってシームレスにシーク可能なバーであり、アノテーションを行う時に頻繁に繰り返されるビデオのシークを直感的に支援する。このシークバーには、ビデオコンテンツのカットから 20 フレーム後のサムネイル画像を表示している。20 フレーム後の画像をサムネイルと

する理由はカット直後の画像は乱れていることが多いためである。このシークバーにより、ユーザは目的のシーンを容易に探すことが可能になり、スムーズにアノテーションをすることが可能になる。

基本的には閲覧することが主目的であるので、なるべく動画閲覧画面を大きくとる構成にしている。印象アノテーションインタフェースは、後述する印象アノテーションのためのインタフェースであり、ボタンを押すという簡便なインタフェースである。また、テキストアノテーションの一覧は、現在のカットに関連する情報を時間軸に応じて表示している。また、後に述べる重要度に応じて、重要度の高い情報が上位に来るようにソートして表示することにより、多数の情報を効率よく表示している。

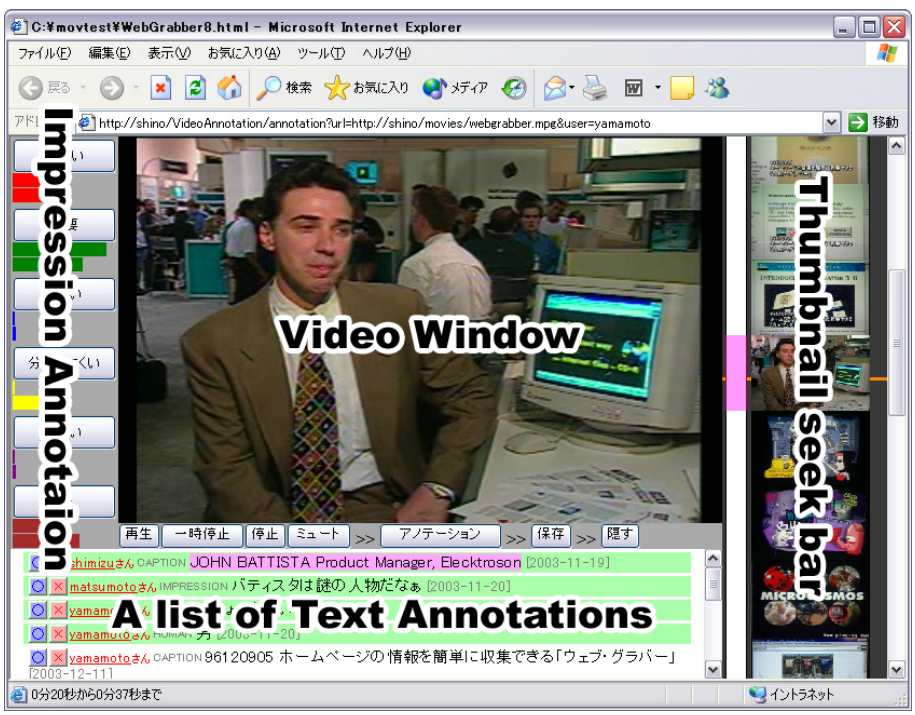

図 3.5: アノテーション編集ページ

3.3.2 テキストアノテーション

テキストアノテーションは、ビデオ内の連続するシーンやオブジェクトに対して、テキストで情報を付加するものである。具体的には、任意のビデオコンテンツのシーンの開始時刻・終了時刻・位置に対してテキスト情報を付与することが可能になる。

まず、ユーザはアノテーションしたい場面で動画上のアノテーションしたいオブジェクトのある部分をクリックする。このクリックした位置を取得することにより、オブジェクトの画面上のおおよその位置が取得可能である。次に、カット検出サーバ等であらかじめ検出されているカット単位でアノテーションしたい時間範囲を選択する。さらに、全てのアノテーションには、後の検索や要約等の機械処理をしやすくするために、コメントの対象（全体・映像・キャプション・音声・音楽・登場人物・オブジェクト・場所など)、種類（名前・状況説明・補足情報・感想など）の選択肢をわかりやすく表示し、容易に選択できるようにした。さらに、他の閲覧者によって入力されたテキストアノテーションに対し、閲覧者が評価する仕組み（○・×のボタンを押す）を用意し、閲覧者によって個々のアノテーションに対する評価を行うことができるように配慮した。コメントの対象と種類のカテゴリは適宜追加・編集可能であるが、どのようなカテゴリを用意するかは今後の課題とする。

具体的な例を図 3.6 に示す。この例において、ユーザは「このシーンに登場する人物はジョン・バティスタだ」という内容を投稿したいとする。まずユーザは、この人物の中心付近をマウスでクリックする。すると、アノテーションを行いたい時間範囲が連続するカット単位で選択可能になる。次にアノテーションの対象となるものの種類の選択を行うが、ジョンバティスタは人物であるので、人物を選択する。次に、アノテーションの種類を選択するのだが、人物の名前を記述するために、「名称」というアイコンを選択する。最後に、この人物の名前を入力する。なお、この例ではこのコンテンツで過去に入力された人物の一覧が表示されており、該当する人物名があれば、それを選択することによってアノテーションを行うことができる。

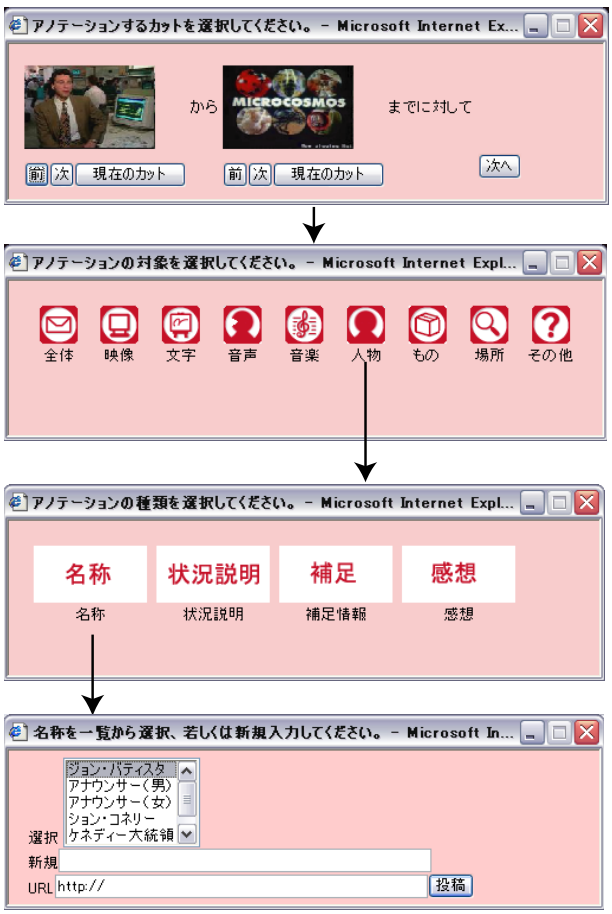

図 3.6: テキストアノテーションの例 (「この人物の名称はジョン・バティスタだ」という内容を投稿している)

テキストアノテーションによって図 3.1 のような情報を取得することができる。これらの情報は、後述するアノテーションを用いた応用に利用される。このようにアノテーションされた情報は、以下のような XML の形式でアノテーション XML データベースに蓄積される。

```
<?xml version="1.0" encoding="x-sjis-unicode" ?>
<annotation id="an1090211400084-0"
            object="OBJECT" type="SITUATION" vid="va00000010">
  <timecode ecut="8" etime="50.2502" scut="8" stime="44.3109" />
```

```
  <position x="0.44745762711864406" y="0.4762979683972912" />
  <annotator handle="yamamoto" id="yamamoto@nagao.nuie.nagoya-u.ac.jp" />
  <description>おいしそうな牛肉だ</description>
  <date>2004-07-19</date>
  <time>13:29:28</time>
  <evaluation bad="1" good="2"/>
  <semantics>
    <word class="ADJECTIVE">おいしい</word>
    <word class="NOUN">そう</word>
    <word class="NOUN">牛肉</word>
  </semantics>
</annotation>
```

上記の例が一つの投稿にあたる。個々のアノテーションには自動的に固有の ID が付与される。その ID が annotation タグの id 属性に当たる。アノテーションの対象が object 属性、アノテーションの種類が type 属性に当たる。また、アノテーションの対象となるコンテンツのコンテンツ ID が vid 属性にあたる。timecode 要素において、アノテーションの対象となる時間範囲を指定でき、scut,ecut 属性はカット単位でのアノテーションの対象時間範囲を示し、stime,etime において秒単位での開始時刻と終了時刻を示す。position 要素の x 属性と y 属性において、マウスポインタがクリックされた位置を示す。また、アノテーションを行った人の情報を annotator 要素で記述する。投稿内容の本文は description において記述し、その情報を後の検索などで使いやすくするために日本語形態素解析器茶筌 [29] を用いて形態素解析した結果を semantics 要素の中に形態素ごとに記述をしている。また、後に述べる、評価アノテーションの結果を evaluation 要素で記述している。投稿時刻と日時を、date 要素と time 要素で記述している。

これらは独自の XML 形式で保存されているが、容易に MPEG-7 などの他の XML 形式に変換可能であるため、他のアノテーションソフトとの連携することも可能である。

表 3.1: 取得可能なアノテーション情報とその取得方法

アノテーション情報	取得手段	作成条件
アノテーション固有 ID	投稿時に生成	自動
個人情報	Cookie で入力	自動
オブジェクトの位置	動画上をクリック	暗黙的
時間範囲	カット単位で指定	必須
コメントの対象と種類	対話的に選択	必須
コメント	テキスト入力	必須
名称	選択または記述	必須
評価	○・×ボタン	任意

3.3.3 印象アノテーション

印象アノテーションとは、ビデオコンテンツの雰囲気や閲覧者の主観的印象、例えば、面白い・緊迫・悲しいなどをマウスクリックでボタンを押すという簡便なインタフェースで入力できる仕組みである。より印象深いシーンでは印象ボタンの連打度合いによって印象の強弱を表現できる。テキストアノテーションの場合は、閲覧を一時的に中断してアノテーションを施す必要があるが、印象アノテーションは閲覧しながらできるという利点がある。

印象アノテーションの対象となる各印象を $I_1 I_2 ... I_n$ とする。その時、クリックした時間を中心にして、正規分布 $N(\mu, \sigma^2)$ で印象情報をつけるとすると、各印象 I_k は以下の式でパラメータを付与する。印象 I_k のボタンを押したアノテーションの集合 S_k とすると、

$$I_k(t) = \sum_{S_k} N(t_i, m) \tag{3.1}$$

ただし t_i は印象アノテーションを行ったメディア時間である。m は定数であり、ボタンを押した時の前後の時間にもアノテーションの効果を与える。これによって、ボタンを押す間隔が短いシーンほど、印象深かったということがわかる。

また、自分のアノテーション結果だけでなく、閲覧者全体のアノテーションの結

果も棒グラフによって表示している。図 3.7 に示すように、それぞれの印象ボタンの下に日本の棒グラフが表示される。上の段がユーザ自身の現在の印象情報を表示し、下段では、閲覧者全員の統計的な印象情報を表示する。このインタフェースによって、ビデオコンテンツの時間軸にそって閲覧者がどのような印象を感じているかの情報を付与することができる。例えば、「おいしそう」という印象情報がある閾値以上の値があるシーンだけを集めることができれば、おいしそうなシーンのみを収集することが可能になる。なお、ボタンの数は最大 6 個まで可能であり、これは、登録サーバで指定可能である。どのようなボタンが有効であるか、また何個必要かといった検証はコンテンツの種類に依存することから今後の課題である。

なお、個人個人の印象アノテーション結果は以下のような XML 形式で保存される。

```
<?xml version="1.0" encoding="Shift_JIS" ?>
<impressionAnnotationLog
  email="matsumoto@nagao.nuie.nagoya-u.ac.jp"
  name="matsumoto" vid="va00000008">
  <impressionSelection num="6">
    <item name="楽しい" no="1">
      <push time="10.11" />
      <push time="12.02" />
      <push time="16.94" />
      ...
    </item>
    <item name="キター" no="2">
      <push time="49.65" />
      <push time="50.01" />
      <push time="50.27" />
      ...
    </item>
    <item name="おいしそう" no="3">
```

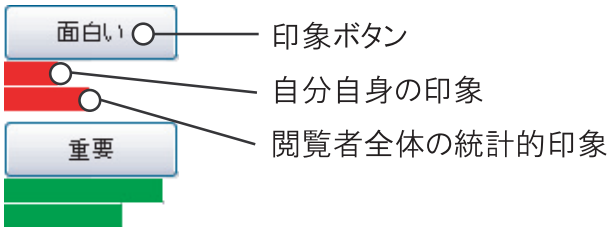

図 3.7: 印象アノテーションのインタフェース (印象ボタンの下に 2 段の棒グラフがあり、ユーザの現在の印象情報の表示と、閲覧者全員の統計的な印象情報を表示する)

```
      <push time="35.39" />
      <push time="36.37" />
      <push time="39.61" />
      ...
    </item>
    <item name="悲しい" no="4">
      <push time="364.36" />
    </item>
    <item name="嫌悪" no="5" />
    <item name="重要" no="6">
      <push time="290.17" />
    </item>
  </impressionSelection>
</impressionAnnotationLog>
```

それぞれのビデオコンテンツに登録されている印象ボタンに対してその印象ボタンを押したメディア時間（秒）を記述可能である。上の例では、matsumoto がコンテンツ ID va00000008 に対しての印象アノテーション結果を意味する。impressionSelection 要素の子要素である item 要素がそれぞれのボタンを意味し、item 要素の子要素の push 要素の time 属性がそのボタンを押したメディア時間を意味する。なお、上の例は簡略化のため一部記述を省略してある。

3.4 アノテーション信頼度

前説までで、ユーザがアノテーションを行う方法を説明した。しかしながら、ただ単にアノテーション情報を収集するだけでは不十分である。アノテーション情報を解析する必要がある。解析をする上で問題になるのは、そのアノテーションが信頼に足りうるものかどうかという点である。不特定多数のユーザによるアノテーションを扱うとどうしても、信頼性の低い情報が混在する可能性がある。そのため、各々のアノテーションに対する信頼度を計算し、情報の選別を行う必要がある。信頼度の計算方法として、「信頼できる情報を多数入力した人の情報ほど信頼できる」という原則に基づいて次の方法で計算している。

まず、A_k に対する単純評価 e_k を以下のように求める。おのおののアノテーションに○ (good) の評価をした人の数 g_k と× (bad) の評価をした人の数 b_k とする。また、選択項目に矛盾がないか、日本語として正しい記述をしているかどうかということにより機械的に決まる評価を c_k とする。ただし、c_k は $-1 < c_k < 1$ となり、c_k の値が大きいほど良いアノテーションと機械的に判断されたこととする。なお、自然言語文によるアノテーションの場合、後の検索や要約などの応用において形態素解析が重要となる場合が多い。そこで、形態素解析を行い、文全体の形態素数のうち未知語の含まれる割合（未知語率）や、文の長さ、構文解析の結果等を総合的に評価することにより c_k を求める必要がある。

これにより、単純評価 e_k は以下の式になる。

$$e_k = s \cdot \frac{g_k - a \cdot b_k}{g_k + a \cdot b_k} + t \cdot c_k \tag{3.2}$$

s は閲覧者による評価の割合、t は機械による評価の割合であり、$s + t = 1$ とし、機械的な評価の精度に依存して t の値を大きくする。

また、a は○評価と×評価の割合を補正する係数であり、すべてのアノテータが行ったすべてのアノテーションに対する評価の総数を g_{all}, b_{all} とすると、

$$a = \frac{g_{all}}{b_{all}} \tag{3.3}$$

と表す。e_k は、$-1 < e_k < 1$ の値をとる。

(2) は機械的な評価と人間の評価を組み合わせた直感的な式であるが、このままでは、アノテーションを行う個人（アノテータと呼ぶ）の信頼性が考慮されてい

ない。

そこで、アノテータに対する信頼度 p を求める。これは今までアノテータが行った全てのアノテーションに対する○ (good) 評価数を G、× (bad) 評価数を B とすると、アノテータ信頼度 p は、

$$p = d(G+B)\frac{G - a \cdot B}{G + a \cdot B} \tag{3.4}$$

により求める。ここで、$d(x)$ はサンプル数が少ない場合に評価値を低く抑える関数であり

$$d(x) = 1 - \exp(-\tau \cdot x) \tag{3.5}$$

とする。ここで、τ はどの程度評価値を抑えるかを決める定数であり $\tau > 0$ である。

また、動画像の時間軸にそったアノテーションをリアルタイムに閲覧者が評価する場合、刻一刻とアノテーション情報が変化するために、閲覧者が誤った○×評価をする場合も多く、閲覧者による評価が集まっていない場合にはアノテータ信頼度を重視し、閲覧者による評価が集まっている場合には閲覧者による評価を重視する必要がある。そこで、アノテーションに対する信頼度 r_k を以下のようにする。

$$r_k = (1 - d(g_k + b_k)) \cdot p + d(g_k + b_k) \cdot e_k \tag{3.6}$$

これにより、アノテーション信頼度 r_k を求めることができる。r_k は、$-1 < r_k < 1$ の値をとり、値が大きいほど相対的に信頼できるコンテンツであると言える。

なお、匿名の書き込みの場合は、アノテータ信頼度は最低の $p = -1$ とする。アノテータに対する信頼度を計算する意義は、機械的にアノテーションを評価するのは難しいこと、ユーザ評価が集まっていない段階ではその情報の信頼性が不明なこと、さらに、信頼性の低い人の大量書込みを防ぐこと（いわゆるアラシ対策)、ユーザに信頼性を公開し、信頼される情報を入力するように暗黙的に誘導するところにある。

今回、アノテーション信頼度はテキストアノテーションのみを扱っている。印象アノテーションと評価アノテーションの信頼度を計算しない理由は、これらの

情報は統計処理されるので、ある程度の信頼を得た情報であると考えているためである。

第4章　アノテーションの応用

前章までに、ビデオコンテンツに対してアノテーションを行う手法について述べた。しかしながら、単にビデオコンテンツに対してアノテーションを行うだけでは不十分である。アノテーションを行う理由は、アノテーションされた情報を用いてなんらかの応用例を期待するためである。そこで、本章では、検索や要約・コミュニティ支援などのアノテーションを用いた応用例を提示する。さらにそれらの有効性を示すためのシステムの構築も行った。システムの評価については次章で述べる。

4.1　本システムで取得可能なアノテーション情報

オンラインビデオアノテーションシステム iVAS で取得可能なアノテーション情報は以下の通りである。

最初にカット検出サーバで取得可能な情報を述べる。カット検出サーバでは文字通り、カット情報の取得が可能である。カットとは、ビデオカメラの切り替わりの時間を意味し、カットとカットで囲まれた映像をショットと呼ぶ。カット検出サーバにおいて、カットのメディア時刻・それぞれのショットに対するサムネイル画像・ヒストグラム情報などが取得可能である。ただし、ビデオショットはそれ自身ではカメラの切り替わりを意味するだけで必ずしもシーンの切り替わりを意味しないことに注意する必要がある。カラーヒストグラム情報は、画面全体が暗い・明るい・青っぽいといった情報を取得できるため有用である。将来的には、キャプション等の文字認識・顔画像認識などのパターン認識技術や音声認識技術などと組み合わせることにより、より多くの情報が取得可能になる。

次に、アノテーションによって取得可能な情報について述べる。テキストアノテーションと評価アノテーションを組み合わせることにより、信頼度付きテキス

トアノテーション情報をビデオコンテンツの任意のシーンに関連付けることが可能である。テキストアノテーションではテキスト情報以外に、個人情報・オブジェクトの位置情報・時間範囲・コメントの対象と種類・○・×ボタンによる評価結果を得ることが可能になる。印象アノテーションにおいては、時間軸に対して、誰がどのようなボタンを押したかという情報が取得可能であり、それらの統計的な情報も計算により取得可能である。テキストアノテーションはコメントベースであるのでより詳細な情報を取得可能であり、印象アノテーションではユーザの主観的な印象を取得可能である。

これらの自動処理によって取得可能なアノテーション情報と、人手によるアノテーション情報を組み合わせることによって、より高度なアノテーションを用いた応用が実現可能になる。

4.2　自然言語による Web 検索

本研究で得られたアノテーション情報の有用性を確認する例としてビデオコンテンツの意味内容に即した検索を行う。カットに対する色ヒストグラム情報やテキストアノテーション情報を元に検索を行う (図 4.1)。

基本的な検索アルゴリズムとしては、筆者らが以前の研究 [25] で提案した、自然言語による意味的ビデオ検索の手法を改良した。具体的には、検索文、テキストアノテーション文、共に茶筌 [29] を用いて形態素解析を行い動詞・形容詞・名詞・未知語を取り出し、それぞれの単語の基本形を基にしてコサイン距離（マッチする単語が多いほど点数を加算）を求める。そして、テキストアノテーションを適用するそれぞれのショットに対してその点数を加算する。さらに今回の手法では、印象アノテーション情報とアノテータ信頼度の情報を用いている。ショットに対してテキストアノテーションの情報を加算する時、アノテータ信頼度と、映像の状況を直接的に表現しているであろうと思われる度合い、つまり、アノテーションの種類が名称 > 状況説明 > 補足説明 > 感想の順に比例した点数を加点している。

さらに、各ショットに対して、印象アノテーションの値が大きいショットほど重要度を上げている。

具体的には以下のようなアルゴリズムによって検索を行っている。

1. 検索キーワードから茶筌 [29] を用いて動詞・形容詞・名詞・未知語を取り出す。未知語とは茶筌に登録されていない単語であり、名詞や英語などの場合が多い。今回は未知語を全て名詞と仮定して使用している。

2. 形容詞や名詞から色にあたる単語（たとえば、赤い・黒い・青い・暗い・明るい等）を色名詞として、各シーンの色ヒストグラムを利用して検索結果を絞りこむ。

3. また、色名詞以外の名詞・形容詞・動詞は、アノテーション情報に記述されたテキスト情報の記述との部分もしくは完全一致に応じて加点する。なお、アノテーションの種類が、名称 > 状況説明 > 補足説明 > 感想の順に重みをつけており、またアノテーション信頼度が高いアノテーションほど大きい重みをつけるようにしている。また、アノテーション範囲の長さの逆数に応じてカットに重みをつけている。

4. 検索文に「おもしろい」、「悲しい」などの印象アノテーションに関連する言葉が入っている場合は、それに対応する各シーンの印象アノテーションの値に応じて加点する。

5. 「人」「男」「人物」など人を現す言葉はテキストアノテーションのコメントの対象の「人物」と対応させ加点する。

6. 最後に全ての印象アノテーションの合計値を各カットに割り当て、なんらかの興味のあるシーンとして、それに応じて加点する。

7. このようにして加点された得点に応じてソートし、順位づけをした上でユーザに Web ブラウザを使って検索結果を示した。

このようにして加点された得点に基づいて表示順序を決定し、検索結果をユーザに提示する。検索結果例を図 4.2 に示す。

4.3 アノテーションによるビデオ要約への応用

アノテーションを用いた要約システムは、音声トランスクリプトに付与された言語構造から重要度を計算してビデオ要約を行う長尾ら [11, 10, 9] の研究や、TFIDF

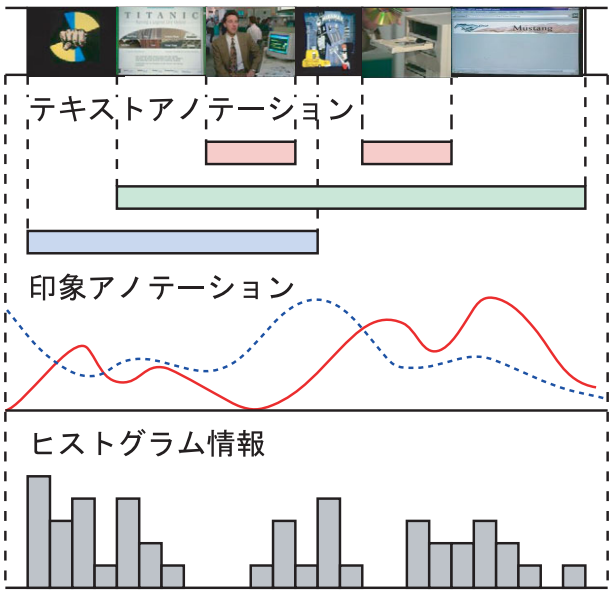

図 4.1: 検索に用いるアノテーション情報

を用いた Informedia プロジェクトなどいくつか存在する。

しかしながら、閲覧者の盛り上がりや視聴者情報を用いていないために、的確な要約を行っているとはいえない。そこで、もし本研究のシステムで得られたアノテーション情報を適用すれば、視聴者情報を活かしたより的確な要約ができると考えた。

具体的には、それぞれのカットに対して、印象アノテーションの各アノテーションの印象度の合計値の時間平均を正規化したものと、そのカットに関連するテキストアノテーションの数を足し合わせたものを、そのカットの閲覧者による重要度と考えて、要約をする場合のシーンの重要度として加算すればよいと考えられる。

本システムでは、「盛り上がっているシーンほど重要」という簡単な規則により上記の方式による閲覧者によるアノテーション情報のみを利用してビデオコンテンツから重要シーンを抽出して時間短縮を行うシステムを試作した。具体的には、上記の方法によって求めた重要度を用いて、指定した簡約時間を超えるまで、重要度の高いカットから順に選択していった。本来ならば、ストーリのつじつまが合うように要約をする必要があるが、今回はそこまで行っておらず、簡約、すなわちコンテンツの時間的縮退に留めた。簡約結果を図 4.3 に示す。

図 4.2: Web ビデオ検索結果

4.4 アノテーションによるコミュニティ支援

本研究のアノテーションの場合、コンテンツに対するアノテーションとの側面以外に、コンテンツを中心とした掲示板コミュニティを形成していると考えられる。また本システムはユーザ認証を用いたアノテーションシステムとしても動作可能であるため、それぞれのユーザに適応したシステムの構築が可能である。そのためアノテーションによって副次的に得られる個人情報を用いた応用として、コミュニティ支援が考えられる。

既存のコミュニティ支援やコンテンツ推薦システムは、コンテンツを見たか見ていないか、あるいはコンテンツを購入したか、していないかによって形成されるものが中心である。例えば、ソーシャルフィルタリングなどの研究はこれにあたる。しかしながら、コンテンツの意味的内容に応じたコミュニティ形成支援やコンテンツ推薦を実現している仕組みは少ない。コンテンツの内容を加味しなければ正確なシステムの実現が難しい例が多い。例えば、サッカー番組でサッカー

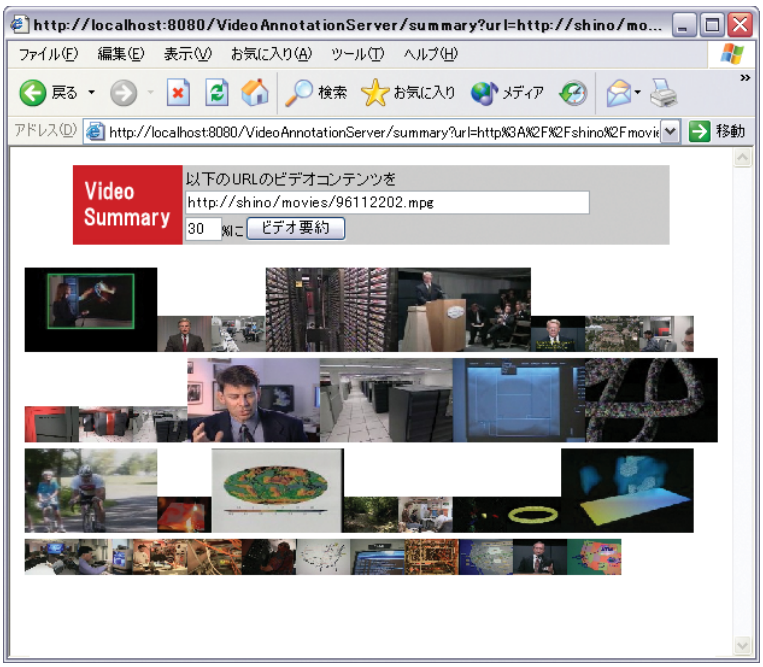

図 4.3: ビデオ簡約の例（小さいサムネイル画像の部分のカットが省略される）

の試合が好きな人たちと特定のサッカー選手のみが好きな人たちとは本質的に違うコミュニティであるし、また、だれもが閲覧しているような有名な映画では単に見たというだけでは情報量が小さく、アクション部分が好きなのか、恋愛シーンが好きなのかによって属するコミュニティが異なる。そこで、印象アノテーション情報を用いて、人と人との興味やものの感じ方の近さの一つの指標となる、印象距離を求める方法を提案する。コミュニティ発見やコンテンツ推薦のための一つの指標となれば良いと考える。

まず、あるコンテンツ C をみた人 P_j と P_k に関する印象距離 $D_C(P_j, P_k)$ を印象アノテーション情報を用いて、以下の式で定義する。

$$Dc(P_j, P_k) = \int (I_{pj} - I_{nj}) \cdot (I_{pk} - I_{nk}) dt \tag{4.1}$$

ここで、I_{pj} は式 3.1 で計算される P_j のポジティブな印象アノテーション情報であり、I_{nj} はネガティブな印象アノテーション情報、I_{pk} と I_{nk} はそれぞれ、P_k のポジティブ印象アノテーション情報とネガティブ印象アノテーション情報を示す。

なお、印象アノテーション情報はマウスクリックの頻度に対する個人差が少なくなるように最大値 1、最小値 0 で正規化を行う必要がある。

4.5 まとめ

本章において、アノテーションを用いた応用のいくつかを提案した。閲覧者によるオンラインビデオアノテーションシステムを構築した場合、大きく分けて二つのアノテーションの応用例がある。一つ目は検索や要約などコンテンツそのものを賢く・使いやすくするやり方であり、二つ目はそのコンテンツを取り巻くコミュニティに関する研究である。このように、コンテンツとそのコミュニティは切っても切れない関係であり、このあたりを意識したアノテーションの研究が重要である。なお、システムの評価は次章で述べる。

第5章　実験と考察

本章では、本論文で作成したシステムの評価を行うために実験とその評価について述べる。被験者は大学生30人であり、映像処理評価用映像データベース[28]のコンテンツを用いて実験を行った。評価する内容は、アノテーション信頼度の妥当性に関する実験と、印象距離の視覚化に関する実験である。また、評価に使われたコンテンツに対してカット検出の実験も行い、カット検出サーバの有用性を示した。

5.1　実験環境

本実験では、映像処理評価用映像データベース[28]のコンテンツを用いて実験を行った。これらのコンテンツは電子情報通信学会のパターン認識・メディア理解研究会によって作成された研究用のコンテンツであり、ビデオコンテンツに関する様々な研究を行うための標準的な映像が収録されている。

本実験では、この中のコンテンツから表5.1で示すコンテンツを、5分程度に編集した評価用映像コンテンツa,b,c,dの4つを用いて実験を行った。5分程度に編集した理由は、評価実験をスムーズに行うためである。大学生男女30人に対してアノテーション及び評価に関する実験を行った。それぞれ、ニュース、ドラマ、バラエティ、料理番組の一部である。

印象ボタンは、ポジティブなボタンとして「楽しい」「おいしそう」「重要」、ネガティブなボタンとして「嫌悪」「悲しい」を用意した。なお、1コンテンツにつき、1名あたり平均10分程度の時間をかけてアノテーションを行っている。

テキストアノテーションに限り、なるべく正確かつ有用な情報を真面目に書き込んでもらうグループA、半分普通に半分不真面目な情報を書き込んでもらうグループB、不真面目な情報を積極的に書き込んでもらうグループC、自由に書き

表 5.1: 評価用映像コンテンツ

	コンテンツ名	ジャンル	利用メディア時間	カット数
a	ニュース 19	ニュース	5 分 33 秒〜10 分 56 秒	56
b	ピエロの涙	ドラマ	0 分 30 秒〜5 分 30 秒	31
c	女王様のブランチ	バラエティ	0 分 00 秒〜6 分 12 秒	72
d	料理番組	料理番組	0 分 20 秒〜3 分 50 秒	32

込んでもらうグループ D、テキストアノテーションを行わないグループ E の 5 つのグループに分けて実験を行った。ここでいう、不真面目とは、全く関連のない情報や間違った情報、あるいは記述した内容自体は妥当であるが、表現に不備があることをいう。

また、この実験を行う前にこと前アンケートとして、NHK ONLINE MEMBERS[1]の中にある好きな番組ジャンル設定 92 項目に対して、好き・普通・嫌いの 3 段階評価を行った。これは、印象距離に関する考察のために用いた。

5.2 アノテーションの結果

被験者 30 名によるアノテーション結果は表 5.2 のとおりである。コンテンツ a,b,c,d ともにほぼ満遍なくアノテーションがされている。テキストアノテーションの総数が 653 個であるのに対して、印象アノテーションの数は、6743 個と 10 倍以上のアノテーションが行われている。これは印象アノテーションの方が簡便なインタフェースであるために、気軽にアノテーションできるためであると考えられる。また、○評価の方が×評価よりも多い結果となったが、それは悪い投稿はテキストアノテーションのリストの下のほうに下がっていくため目立たなくなり、×評価がされなくなると考えられる。そのために、単純に○評価と×評価を比較することはできないと考えられる。

[1]https://members.nhk.or.jp/

表 5.2: 得られたアノテーション結果

	評価人数	テキストアノテーション数	印象アノテーション数	○評価	×評価
a	30 人	174	2153	239	68
b	30 人	160	1225	234	70
c	30 人	222	2132	334	61
d	30 人	97	1224	159	44

表 5.3: カット検出の精度

	ジャンル	検出カット数	過検出	未検出
a	ニュース	56	8	8
b	ドラマ	31	0	2
c	バラエティ	72	1	3
d	料理番組	29	0	3

5.3 カット検出に関する実験

今回の実験で用いたコンテンツに対するカット検出の精度を表 5.1 に示す。なおこの表で、検出カット数は本システムでカットと認識した数、過検出は本来カットではない部分をカットとして認識した部分、未検出は本来カットである部分を検出できなかった数をいう。今回、ニュースコンテンツでの精度が極端に悪いのは、花火大会に関するニュース映像のコンテンツでの誤検出が多かったためである。これは、花火のフラッシュを誤ってカットと認識してしまう例があったためである。カット検出の全体の適合率は 89.3 %であり、再現率は 91.7 %であった。

5.4　アノテーション信頼度に関する実験と考察

次にアノテーション信頼度に関する考察を行う。アノテーション信頼度はアノテータ信頼度と閲覧者による○×評価を元にして求める。ここでの○×評価は一般的なものであり、特にバイアスがかかっているわけではない。一方、アノテータ信頼度に関しては、よく考察する必要がある。そこで、アノテータ信頼度の妥当性を考察するために、真面目グループ (A)、半分不真面目グループ (B)、不真面目グループ (C)、自由グループ (D) のアノテータそれぞれに関するアノテータ信頼度を式 3.4 より求め図 5.1 に図示した。なお今回の実験では、式 3.2 における機械的な評価の精度による誤差を無くす為に $s = 1, t = 0$ とした。また、○の評価が×の評価に比べて 4 倍以上多く、式 3.3 より $a = 4$ とした。また、閲覧者による評価数が 4 個で式 3.5 の値がおよそ 0.5 になる $\tau = 0.2$ とした。

図 5.1 から分かるように、アノテータ信頼度の値は $A > B > C$ の傾向があり、これは直感と一致する。信頼度にばらつきが生じる理由は、アノテータによって不真面目さの基準が異なるため、不真面目な書き込みであっても、それは映像から想起される情報であるため、内容によっては閲覧者がその書き込みを容認し×の評価をしない場合があるためである。

また、自由に書き込んだグループのアノテータ信頼度のばらつきが大きい。これは、なんら制約なく自由に書き込むために個人差が出やすくなるためであると考えられる。そのため、アノテータ信頼度を求める意義があると考える。

この結果得られたアノテータ信頼度を元にして、D グループ 6 人、153 個のテキストアノテーションに対する式 3.6 によって求めたアノテーション信頼度の値と、式 3.2 の単純評価を求め妥当性を比較した。まず、筆者が主観で全てのテキストアノテーションを良い・悪いの二つに分類した。分類基準は、シーンの内容を的確に表現していない、あるいは、正しい日本語の記述ではないものを悪いとし、それ以外は良いとした。悪い分類をしたものは 23 個、良い分類をしたものは 130 個あった。このうち単純評価で明らかに間違っているもの (悪い分類なら $r_k > 0.4$、良い分類なら $r_k < -0.4$) は 17 個であるのに対し、アノテーション信頼度では 3 個と、アノテータ信頼度を考慮しない場合に比べて、間違った評価をした信頼度が減少しており、その点で改善されているといえる。ただし、今回の予備実験では、テキストアノテーションの数が 153 個なのに対して、○×評価の数が 289 個と少

ないため、正確な評価は今後の課題としたい。

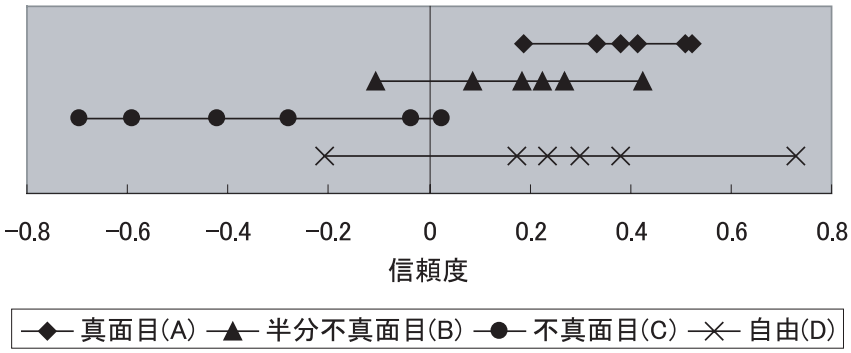

図 5.1: アノテータ信頼度 (数値が大きいほど信頼できるアノテータである)

5.5 印象距離に基づくコミュニティの視覚化に関する実験

次に印象距離に関する実験を行った。内容がバラエティに富んでいるほど、印象距離を求める意義があるために、女王様のブランチ六本木編を用いた。女王様のブランチ六本木編に対して行った印象アノテーション情報 26 人分を式 4.1 を用いてそれぞれの印象距離を求め、ばねモデルを用いて図示したものを図 5.2 に示す。一つの円が一人のアノテータを意味し、距離が近いアノテータほど今回のコンテンツに関しては興味が近かったことを意味する。背景が灰色の円はアンケートでファッション番組に興味があると答えた人、太字の名前は歴史・紀行に関する番組に興味がある人を示す。女王様のブランチ六本木編は、アメリカ風カフェ、和風料理店、バー、ファッション、カラオケ、映画館、喫茶店、神社関連のお店や名所を順に 1 分程度づつ紹介していく番組である。事前アンケートにより、ファッション番組に興味があると分かっている人について、図 5.2 に示すように、印象距離を見てみると比較的近い距離に密集する傾向があり、一つのコミュニティを形成していると考えられる。

図5.2のように、歴史・紀行に興味がある人も印象距離が近い場合があるが、ファッションに興味がある人に比べて密集度が小さく、コミュニティを形成しているとは考えにくい。女王様のブランチのように複数の話題があるコンテンツの場合、人によって興味のもつ話題が複数に分散している場合が多く、そのために式4.1の計算手法では全体的に印象の薄い話題が印象の強い話題でかき消されてしまう可能性があるからである。明確なコミュニティを視覚化するにはより多くの閲覧者を必要とするか、あるいは、印象距離を話題ごとに時間区間で区切って個別に考える必要があると考えられる。

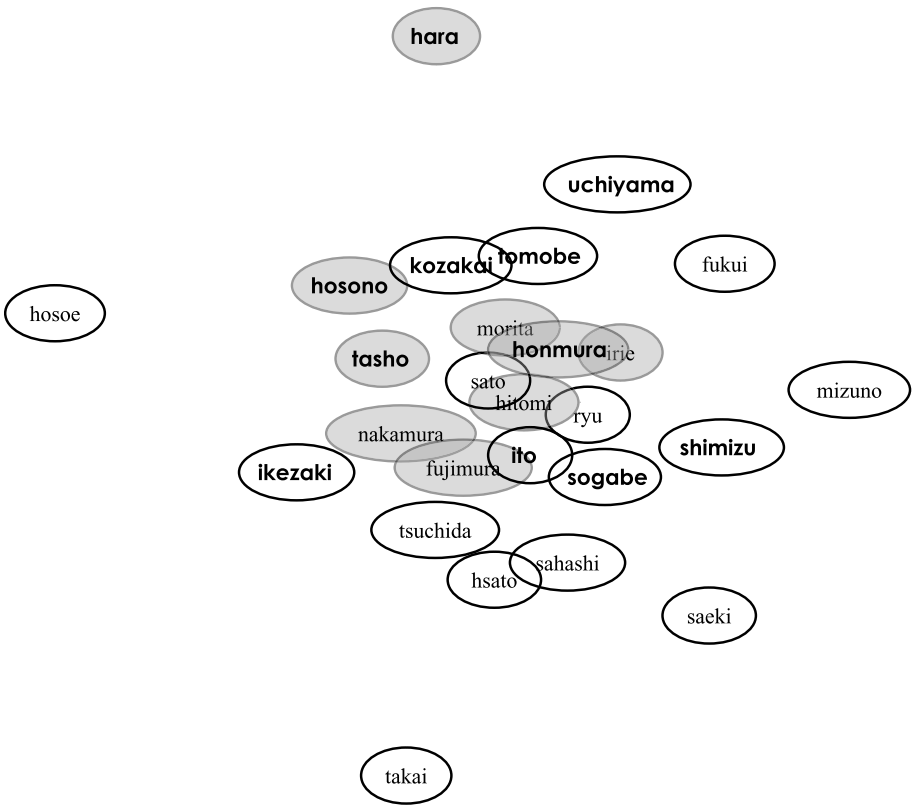

図 5.2: 印象距離

表 5.4: 5 段階評価によるアンケート結果

評価指標	1	2	3	4	5	平均
テキストアノテーションの使いやすさ	0	3	7	12	2	3.50
印象アノテーションの使いやすさ	0	3	8	11	8	3.80
正確に付与できたか	0	2	11	16	1	3.53
取っ付き易さ	0	2	7	10	11	4.00
iVAS を使いたいか	1	1	11	11	6	3.67

5.6 アンケートによる評価

本システムは、より多くのユーザに使ってもらえてこそ意義があるシステムである。そこで、システム全体の使いやすさに関するアンケートを行った。アンケート結果を表 5.4 に示す。アンケートは、評価実験後に行い、5 段階評価で集計をした。なお、この表において数字が大きいほど高い評価をしたことを意味する。テキストアノテーションを行っていない 6 人に関してはテキストアノテーションの評価をしていない。

母集団が大学生ということもあるが、iVAS の使いやすさを示す「取っ付き易さ」の項目で多くの人が普通よりも良い項目をつけており、本システムのインタフェースはなかなか評判のよいものであった。また、iVAS を使ってアノテーションをしたいかという問いに対しても同様に高評価であり、本システムが比較的多くのユーザに使ってもらえる可能性を示唆する結果となった。また、テキストアノテーションの使いやすさよりも印象アノテーションの使いやすさの方が良い結果となっており、印象アノテーションの使いやすさを示すことができたと考えられる。しかしながら、使いにくいという判断をする被験者もいるため、誰にとっても印象アノテーションは使いやすいとは限らないことも分かった。これは、全く新しいインタフェースであるためであろうと推測できる。

5.7　実験のまとめ

評価実験によって、アノテータ信頼度・アノテーション信頼度の妥当性を示した。これによって、アノテーション信頼度付きテキストアノテーションを取得可能になった。また、印象距離の実験も行い有用性を示したと考えている。

また、今回の実験では、検索や要約等に関する実験は行っていない。今回の実験によって得られる情報は、被験者実験によって得られた情報であり、ある程度の恣意性が残るため必ずしも実験に適したデータではないからである。特に検索は、アノテーションの正確さと量によって精度が大きく変わってしまう項目であり、これらの評価は今後の課題としたい。

第6章　今後の方向性

本論文では、マルチメディアコンテンツに対する、閲覧者によるアノテーションの仕組みと、それによって得られたアノテーション情報を用いた応用についての提案と実験および考察を行った。本章では、本研究をより有意義にするための新たな方向性について述べる。

6.1　Weblogを用いたビデオアノテーションとそのコミュニティの活性化

近年ブロードバンドネットワークの発達と共に、インターネット上でビデオコンテンツを柔軟に取り扱える環境が整いつつある。現状ではビデオコンテンツを閲覧することしかできず、インターネットの双方向性という利点をあまり活用していない。そこで、ビデオコンテンツにWeblog[7]の仕組みを取り入れることによって、コンテンツとそれを取り巻くコミュニティを活性化する仕組みが考えられる。

6.1.1　オンラインビデオコンテンツのWeblog化

Weblogは、インターネット上で記事（エントリーと呼ばれる）を効率よく配信するための仕組みである。Weblogにはそれぞれのエントリーに対して、閲覧者がコメントを付ける、トラックバックでリンクを張る、更新情報をRSS(RDF Site Summary) [15]で配信する仕組みが備わっている。これらの仕組みを用いて、各々のエントリー同士をトラックバックなどのリンクで結びつけることにより、Weblogエントリーを中心としたコミュニティを活性化することができる。本研究では、Weblogのエントリーをビデオコンテンツ（の任意のシーン）に置き換えることにより、ビデオコンテンツとそのコミュニティの活性化を促す仕組みを提案する。コ

ンテンツを Weblog 化することを Weblogize と呼び、Weblogize されたビデオコンテンツを Videoblog と呼ぶ。

Weblogize のための一つ目の仕組みとして、ユーザがコンテンツに対して容易にコメントを付与できる仕組みが必要である。そこで、先に述べた iVAS の仕組みを利用する。

Weblogize のための二つ目の仕組みとして、トラックバックの仕組みを iVAS に取り入れた。トラックバックとは、記事の参照先から参照元への間に自動的にリンクを張る仕組みである。通常のサイト単位の相互リンクとは違って、記事単位のリンクであるため、よりピンポイントに、コンテンツに依存したリンクを張ることができる。さらに、このリンクは人間の主観によって張られるものであり、記事内容を意識した関連性の強いリンクである。

Videoblog のトラックバックの仕組みは、Weblog と同じであり、trackback ping を送ることによって成立する。trackback ping の送り先 URL は次のようになり、対象となるコンテンツの ID と関連付けたいシーンの開始時間と終了時間を指定することによってトラックバックを受け入れる。

```
http://[ivas server address]/tb/
   [content ID]/[begin time]-[end time]
```

そのため、Videoblog は既存の Weblog には一切の改変を加えることなく、Weblog のトラックバックネットワークに参加することが可能である。

6.1.2　トラックバックに基づくビデオアノテーション

以上のように、ビデオコンテンツにトラックバックの仕組みを導入した。トラックバック元の記事内容は、トラックバックの張られているビデオコンテンツのシーンの意味内容に関連しており、アノテーションの一種として扱うことができる。もし、ユーザにとってトラックバックを張る十分な利益があれば、極めて低い人的コストでビデオアノテーションを行うことができる。そのための一つの手段として、以下の映像視聴記事作成インタフェースを提案する。

ユーザはコンテンツを閲覧する時、iVASの電子掲示板風インタフェースを用いて閲覧することにする。このインタフェースでのアノテーションは閲覧時にリアルタイムに作成することを想定しているため、必ずしも十分な内容のアノテーションを期待できない。しかしながら、そのユーザにとって良いコンテンツであった場合、後からそのコンテンツに対する評価または感想の記事を書きたいという欲求がある。もし、後で記事を書くことを前提した閲覧の場合、簡便な印象アノテーションなどは施していることが期待できる。

そこで、iVASを用いてリアルタイムにアノテーションされた情報から、映像視聴記事対象シーンを推定する。具体的には、ユーザにとって興味深かっただろうシーンを推定し、そのコンテンツに対するWeblogのエントリーを書くためのテンプレートを提供する。また、ユーザにとって興味深いシーンへのリンクと、対応するVideoblogのシーンへのトラックバックを自動的に張る。その推薦結果に基づき、ユーザはそのシーンに対する感想等の記事を記述する。

映像視聴記事対象シーンの推薦の仕組みは概ね以下の通りである。まず、ユーザがそのシーンに対してどれくらいの興味を持っているかを示す度合いとして、興味度という指標を導入する。興味度は、そのシーンに対してどの程度アノテーションを施したかに応じて興味度が高くなる。具体的には、テキストアノテーションを記述したシーンに対して興味度は高く、また印象アノテーションのボタンがたくさん押してあるシーンも興味度が高くなるようにする。そして、興味度の高い順にシーンn個を推薦する。さらに、同じビデオシーンの他のユーザの興味度と比較し、他のユーザよりもそのシーンに対して相対的に興味が高いシーンも推薦する。これは、シーンごとのアノテーションの数をある程度均等にするためである。

ユーザは、Weblogなどで通常のエントリーを書くのと同様な形式で、半自動生成されたエントリーを修正できる。この修正結果はiVASにテキストアノテーションとして反映され、検索などの各種応用の精度をあげるのに利用される。修正できる項目は以下の3つである。1つ目は、アノテーションの対象となるシーンの選択範囲の修正であり、あらかじめカット検出されたカット単位でシーンの開始カットと終了カットを修正することができる。2つ目は、テキストの修正であり、3つ目はシーンごとにエントリーを公開/非公開するかの選択である。画面例を図6.1に示す。修正された結果は、修正前のアノテーションに対して、上書き属性のア

ノテーションとして XML データベースに記録される。これにより、人間の手が加わったより信頼度の高いアノテーションとして保持される。

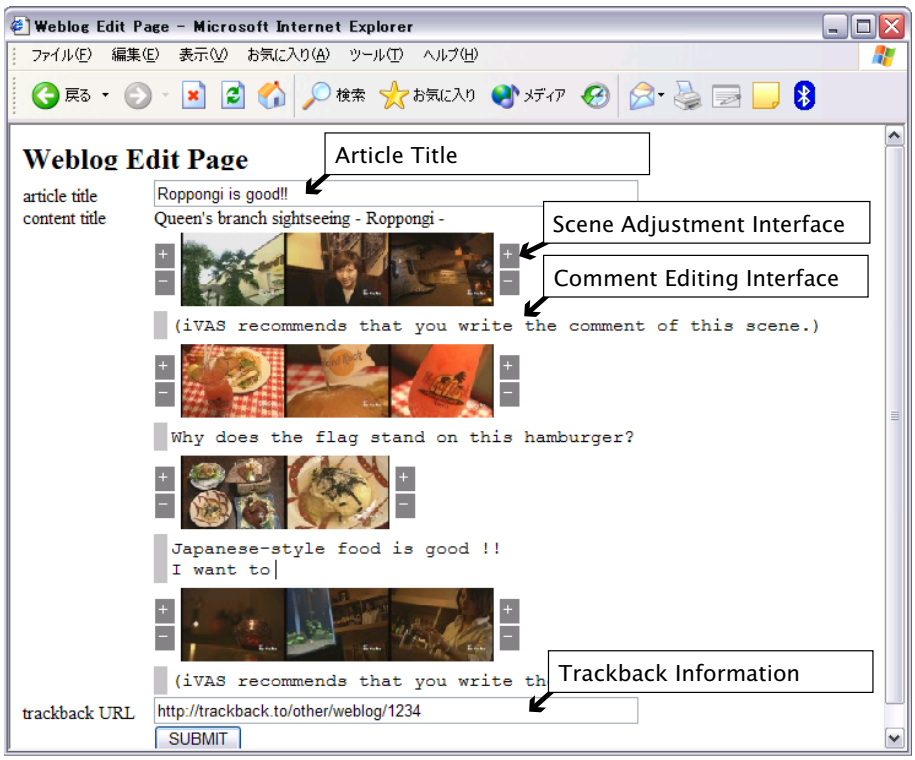

図 6.1: Weblog 編集画面

この仕組みを用いることにより、ビデオコンテンツと Weblog とユーザとを結びつけることが可能になった。これらの仕組みは二つの利点がある。一つはユーザが普段よくみる Weblog とビデオコンテンツとさらにそれに関連するコンテンツを結びつけることにより、コンテンツの流通の促進と活性化を促すことができる。さらに、コンテンツに対するビデオアノテーションとして、コンテンツの意味内容に基づく検索や、要約などの既存の研究などに活かすことができると考えている。

6.2 リアルタイム TV コンテンツへの適用

iVAS では、Web 上にあるコンテンツに対して掲示板型の非同期でコミュニケーションをするシステムであった。現在放映中の TV コンテンツに対するアノテーションに対しては対応していない。もちろん、掲示板形式のインタフェースをチャット形式のインタフェースに変更すれば同様な考え方を適用することは可能である。TV コンテンツに対してチャット形式においてコメントを付与する例として、大規模匿名掲示板群である 2 ちゃんねるの実況板が上げられる。実際に 1 分間あたりに書き込まれた数の例を図 6.2 に示す。この例では、2004 年 9 月 30 日の NHK 総合放送に対して付与されたコメント例である。このように人気のあるコンテンツでは 1 分間に 100 以上のコメントが付与されている。筆者は、以前これらのコメントを用いてアノテーションに応用できないかと考えたが、あまりにも大量なコメントが存在しているために、解析に時間がかかり断念した経緯がある。これらの情報を効率良く解析し、アノテーション情報として利用する手法については今後十分な検討が必要である。

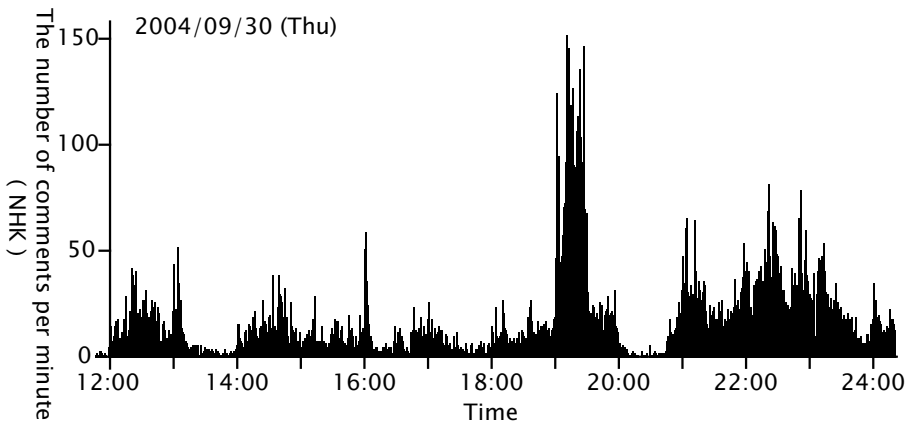

図 6.2: 2 ちゃんねるの書き込み数の変化

6.3　iVAS の公開と運用

本研究で作成したアノテーションを用いたシステムを検証するためには、一般の人に実際に iVAS を公開し、実際のデータを収集する必要がある。そのためには、iVAS を実システムに添った形で作り直し、ソースコードと共に公開する必要があると考えている。また、Web 上で公開可能なビデオコンテンツを収集し、著作権的に問題ないように配慮する必要がある。

第7章　まとめ

本論文ではまず、ビデオアノテーションに関する関連研究を紹介し、その一般的な概念を解説した。ビデオアノテーションには大きく分けて二つある。一つ目は、従来からあるオフラインビデオアノテーションであり、二つ目は、筆者らが提案しているオンラインビデオアノテーションである。オフラインビデオアノテーションでは、専任のアノテータがローカルにあるコンテンツに対してアノテーションを行うのに対し、オンラインビデオアノテーションではオンラインにある任意のコンテンツに対して複数人でアノテーションを行うことが可能である。特に、本論文では、閲覧者によるオンラインビデオアノテーションについて解説した。つまり、そのコンテンツを見ている人にアノテーションに参加してもらえれば、より効率良くアノテーションが作成できるのではないかと考えたためである。

そこで本研究では、以上の考察に基づきオンラインビデオアノテーションシステム iVAS(intelligent Video Annotation Server) を開発した。本システムはカット検出情報や色ヒストグラムを得るために動画像解析を行う。そしてまた、自動的に掲示板風のアノテーション編集 Web ページが生成される。閲覧者は一般的な Web ブラウザを通して、オンライン上にあるビデオコンテンツに対して以下の 3 つの方式でアノテーションを行うことができることを示した。1 つ目は閲覧者によって、人やオブジェクトの名前や状況説明や感想を対話的にアノテーションする方式であり、2 つ目は、マウスのボタンをクリックするという簡便なインタフェースで、閲覧者の主観的な印象を任意のビデオのシーンに対して関連付ける方式である。3 つ目はそれぞれのテキストアノテーションに対して○×ボタンによって評価を行う評価アノテーションである。生成されたアノテーション情報は iVAS に接続された XML データによって蓄積・管理される。このようにして、閲覧者がインターネット上に存在する任意のビデオコンテンツに対してアノテーション（メタ情報）を関連付けることができることを示した。

また、不特定多数の閲覧者によるアノテーション作成を可能にすると、どうしても信頼性の低いアノテーションが混在するという問題点がある。そこでユーザからのフィードバック情報（○×ボタン）からそれぞれのアノテータの信頼度と個々のアノテーションの信頼度を計算する手法を提案した。またその手法の妥当性を確認するために、被験者30人による評価実験を行った。

また、本システムで収集したアノテーション情報の有効性を確認するために、アノテーションに基づく応用として、ビデオ検索、ビデオ簡約、ビデオコンテンツコミュニティ支援に関するシステムを構築した。これらのシステムは、一般的なWebブラウザを用いてアノテーションを作成することが可能である。

本研究で開発したシステムの優位性は、人手によるアノテーションと機械によるアノテーションを組み合わせている点、および、その人手によるアノテーションは、閲覧者がコンテンツに対して掲示板風なインタフェースによるユーザ間でのコミュニケーションの結果により副次的に得られるものであり、人的コストはほとんど必要としないという点である。また、コンテンツとアノテーションを関連付けて蓄積しているので、将来的にアノテーションの解析技術が向上すればより精度の高い応用技術が期待できる。

将来、これらの基盤的技術がビデオコンテンツの発展とそれによって生成される新しいコミュニティの形成に貢献できるものと考えている。

なお、本研究で作成したシステムが以下のページで公開されているのでぜひ参照して頂きたい。

http://www.nagao.nuie.nagoya-u.ac.jp/ivas/

謝辞

本研究を進めるにあたり、指導教官である長尾確教授、大平茂輝助手には、研究の心構えなど基礎的なことから、 ゼミ等を通しての貴重なご意見、論文指導等を賜り大変お世話になりました。御礼申し上げます。また、学部生時代にお世話になった、末永康仁教授、森健策助教授には、研究を行う上での基本的なスタンスを御指導いただき有難う御座いました。研究室メンバーの友部博教さん、梶克彦さん、松本和之くん、小酒井一稔くん、本村可奈子さん、佐橋典幸くん、土田貴裕くんとは研究室活動を通じて大変お世話になりました。さらに、元研究室メンバーの、山根隼人さん、清水 敏之くん、松浦 真治くん、細野 祥代さん、加藤範彦くん、田中 和也くん、鬼頭 信貴くんにも大変お世話になりました。最後に、長尾研究室元秘書の兼松英代さん、現秘書の金子幸子さんには学生生活ならびに研究活動のための様々なサポートをいただきました。この場を借りて御礼申し上げます。

参考文献

[1] *Apache Xindice*. http://xml.apache.org/xindice/, 2004.

[2] David Brgeron, Anoop Gupta, Jonathan Grudin, and Elizabeth Sanocki. Annotations for streaming video on the web: System design and usage studies. In *Proceeding of The Eighth International World Wide Web Conference*, pp. 61–75, 1999.

[3] David Brgeron, Anoop Gupta, Jonathan Grudin, Elizabeth Sanocki, and Francis Li. Asynchronous collaboration around multimedia and its application to on-demand training. In *Proceeding of the 34th Hawai International Conference on System Sciences*, 2001.

[4] Marc Davis. An iconic visual language for video annotation. In *Proceedings of IEEE Symposium on Visual Language*, pp. 196–202, 1993.

[5] T Hirotsu, T Takada, S Aoyagi, K Sato, and T Sugawara. Cmew/u-a multimedia web annotation sharing system. In *TENCON 99. Proceedings of the IEEE Region 10 Conference*, Vol. 1, pp. 356–359, 1999.

[6] Ching-Yung Lin, Belle L. Tseng, and John R. Smith. *VideoAnnEx Annotation Tool*. http://www.research.ibm.com/VideoAnnEx/, 2002.

[7] Charlie Lindahl and Elise Blount. Weblogs: simplifying web publishing. *IEEE Computer*, Vol. 36, No. 11, pp. 114–116, 2003.

[8] MPEG. *MPEG-7*. MPEG-7 Consortium, http://www.mp7c.org/, 2002.

[9] Katashi Nagao. *Digital Content Annotation and Transcoding*. Artech House Publishers, London, 2003.

[10] Katashi Nagao, Shigeki Ohira, and Mitsuhiro Yoneoka. Annotation-based multimedia summarization and translation. In *Proceedings of the Nineteenth International Conference on Computational Linguistics (COLING-02)*, pp. 702–708, 2002.

[11] Katashi Nagao, Yoshinari Shirai, and Kevin Squire. Semantic annotation and transcoding: Making Web content more accessible. *IEEE MultiMedia*, Vol. 8, No. 2, pp. 69–81, 2001.

[12] E. Oomoto and K. Tanaka. Ovid: Design and implementation of a video-object database system. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 5, pp. 629–643, aug 1993.

[13] Sujeet Pradhan, Keishi Tajima, and Katsumi Tanaka. Querying video databases based on description substantiality and approximations. In *Proceeding of the IPSJ International Symposium on Information Systems and Technologies for Network Society, World Scientific*, pp. 183–190, sep 1997.

[14] Ricoh. *Movie Tool*. http://www.ricoh.co.jp/src/multimedia/MovieTool/, 2002.

[15] RSS-DEV Working Group, http://web.resource.org/rss/1.0/spec. *RDF Site Summary (RSS) 1.0*, 2001.

[16] G. Salton, A. Wong, and C.S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, Vol. 18, No. 11, pp. 613–620, 1975.

[17] SemanticWeb.org. http://www.semanticweb.org/, 2005.

[18] D. Wactlar, T. Kanade, A. Smith, and M. Stevens. Intelligent access to digital video: Informedia project. *IEEE Computer*, Vol. 29, No. 5, pp. 140–151, 1996.

[19] Ron Weiss, Andrzej Duda, and David K. Gifford. Content-based access to algebraic video. In *Proceedings of IEEE First International Conference on Multimedia Computing and Systems, Boston, MA*, 1994.

[20] 山田一穂, 宮川和, 森本正志, 児島治彦. 映像の構造情報を活用した視聴者間コミュニケーション方法の提案. 情報処理学会研究報告・グループウェアとネットワーク, 第 43 巻, 2001.

[21] 長坂晃朗, 田中譲. カラービデオ映像における自動索引付け法と物体探索法. 情報処理学会論文誌, Vol. 33, No. 4, pp. 543–550, 1992.

[22] 橋田浩一. Gda: 意味的修飾に基づく多用途の知的コンテンツ. 人工知能学会誌, Vol. 13, No. 4, pp. 528–535, 1998.

[23] 是津耕治, 上原邦昭, 田中克己. 時刻印付オーサリンググラフによるビデオ映像のシーン検索. 情報処理学会論文誌, Vol. 39, No. 4, pp. 923–932, 1998.

[24] 山本大介. 半自動ビデオアノテーションとそれに基づく意味的ビデオ検索に関する研究. 名古屋大学工学部情報工学専攻 卒業論文, 2003.

[25] 山本大介, 長尾確. 半自動ビデオアノテーションとそれに基づく意味的ビデオ検索. 情報処理学会第 65 回全国大会, 第 3 巻, pp. 95–96, 2002.

[26] 山本大介, 長尾確. 閲覧者によるオンラインビデオコンテンツへのアノテーションとその応用. 人工知能学会論文誌, Vol. 20, No. 1, pp. 67–75, 2005.

[27] 谷田部智之, 佐藤隆, 坂内正夫. 放送間のリアルタイムリンクを可能とするアドバンスト tv の構想. 電子情報通信学会情報・システムソサイエティ大会 D-279, 1996.

[28] 馬場口登, 栄藤稔, 佐藤真一, 安達淳, 阿久津明人, 有木康雄, 越後富夫, 柴田正啓, 全炳東, 中村裕一, 美濃導彦, 松山隆司. 映像処理評価用映像データベースについて. 電子情報通信学会技術研究報告 PRMU2002-30 June, 2002.

[29] 奈良先端科学技術大学院大学自然言語処理学講座. 形態素解析システム 茶筌. 2003.

[30] 佐藤隆, 坂内正夫. ライブハイパーメディアに基づく放送映像のアクセス. 情報処理学会第 51 回（平成 7 年度後期）全国大会講演論文集, 第 3 巻, pp. 301–302, 1995.

[31] 佐藤隆, 坂内正夫. ライブハイパメディアにおける映像情報の獲得. 電子情報通信学会論文誌（D-II）, 第79巻, pp. 559–567, 1996.