Webコミュニティ活動に基づく
映像アノテーションとその応用に関する研究

山本 大介

目 次

図 目 次

表 目 次

概要

本論文では，Web 上で共有されている映像コンテンツをより高度に利用するために，映像コンテンツに対するアノテーション技術，とりわけ，Web 上の一般ユーザによるアノテーション技術とそれに基づく応用技術について提案する．

映像コンテンツはテキストコンテンツなどとは異なり，テキスト検索のように意味内容を考慮した内容検索などの高度な処理を行うことが一般に困難である．映像共有サービスの普及とともに，Web 上で扱われる映像コンテンツは氾濫しつつあり，いかに効率よくこれらのコンテンツの管理・検索・要約・宣伝などの応用を実現するかという問題が顕在化してきた．

従来，とりわけ，基礎研究の分野において，映像コンテンツの高度利用に対するアプローチは自動解析技術のみに基づく手法に傾倒してきた．映像解析技術は映像コンテンツの構造情報を理解することは可能であっても，映像コンテンツの意味を理解することは困難であり，将来にわたって信号処理技術のみでの実現は困難であると予想する．我々の提案手法は，人手によって意味内容を映像コンテンツと関連付けることを支援する仕組みである．いかにアノテーションのコストを低く，また，より正確なアノテーションを関連付ける仕組みを作るかということが問題になる．

そこで，本論文では，効率よく映像コンテンツに対するアノテーションを付与する手法について検討する．そのために，まず，これらの手法の基盤となる，アノテーションモデルについて検討する．具体的には，アノテーションに記述する内容の形式によって，様々なアノテーション型があることを示す．さらに，これらのモデルに基づき，効率よく映像コンテンツに対するアノテーションを付与する手法として以下の３つのアプローチを提案し，それぞれ具体的なアノテーションシステムを開発した．

1. VAE: オントロジーに基づく半自動アノテーション
2. iVAS: 閲覧者によるオンライン映像アノテーション

3. Synvie: Web コミュニティ活動に基づく映像アノテーション

オントロジーに基づく半自動アノテーションとは，映像コンテンツに対する 2 種類のアノテーションの作成を支援する仕組みである．1 つは，映像に対する構造アノテーションであり，信号処理技術とその解析結果の修正を支援する仕組みを備える．もう 1 つは，映像に対する意味アノテーションであり，映像に関するオントロジーと映像の要素を関連付けることによって行うオントロジー型アノテーションの仕組みを備える．具体的なシステムとして VAE を開発した．機械にできることは極力機械によって，人間にしかできないことは人間によって，協調的にアノテーションを作成する仕組みである．この方式のアノテーションは，専門家が商用コンテンツに対して詳細にアノテーションを付与することに適している一方，アノテーション作成コストが比較的高いので Web で共有されているようなコストをかけることが困難な映像コンテンツに対して同様の仕組みを適用することは困難であるという欠点もある．

閲覧者によるオンライン映像アノテーションとは，映像の閲覧者にアノテーションの参加を促す仕組みである．そのために，映像シーンを話題とした掲示板型コミュニケーションを支援するユーザコメント型アノテーションの仕組みを備える iVAS というシステムを開発した．また，不特定多数の閲覧者がアノテーションを作成する方式であるためアノテーションの信頼性が一様ではない問題があるため，アノテーション信頼度と呼ばれる評価手法を導入し，それに基づくアノテーションの選別手法を提案した．さらに，閲覧者のアノテーション履歴から閲覧者の嗜好を評価するための仕組みとして，印象距離と呼ばれる指標を提案し，これらの評価を行った．この方式はアノテーション作成のためのコストが低いという特徴があり，より多くのコンテンツに対して適用可能である．

Web コミュニティ活動に基づく映像アノテーションとは，基本的には閲覧者によるアノテーション方式と同じであるものの，映像を話題としたコミュニティ活動は映像の閲覧時だけではなく，ブログなどでも映像を話題とした記事が書かれている点に着目し，それらの情報もアノテーションとして獲得した．具体的には，映像シーンを引用したブログエントリーの執筆を支援する，引用型アノテーションの仕組みを提案した．引用型アノテーションでは，映像に関する意味情報だけではなく，引用の共起関係に基づく構造情報の獲得が可能である．さらに，これらの仕組みを実装したシステムとして Synvie を開発し，比較的大規模な公開実験を行った．2006 年 7 月から一般

公開し，284 名の登録ユーザと，223 個の映像コンテンツと，3534 個のユーザコメントに基づくアノテーションを含む 19182 個のアノテーションの取得に成功した．これらの取得されたアノテーションの定量的な分析を行うことによって，引用型アノテーションの優位性を示した．これにより，iVAS よりもより良いアノテーションの収集が可能になるという利点がある．

次に，Synvie によって獲得されたユーザアノテーションに基づいた応用について提案する．具体的には，Web コミュニティ活動に基づくアノテーションの仕組みによって獲得されたアノテーションに基づく検索システムを提案する．アノテーションのテキスト情報からタグ集合を生成することによって映像シーンに対する意味情報と，複数の映像シーンを同時引用することから得られる共起関係やアノテーションの参照頻度を用いた構造情報を利用する．これらのアノテーションは必ずしも信頼性が高いとはいえないという欠点がある．これらの欠点を補うために，タグクラウドの仕組みを利用した映像シーン検索システムを開発した．このシステムは，既存のタグクラウドと同様に，人間が提示されたタグを探索的に検索する仕組みであり，Synvie によって獲得されたアノテーションに適した仕組みである．本システムを利用した評価実験を行うことによって，必ずしも信頼性が高くないアノテーションを用いた応用システムにおいても，アノテーションの有用性を確認した．

Web ユーザやコミュニティによるコミュニティ活動から，映像コンテンツに対するアノテーションを間接的に作成するというアプローチが本論文の最大の新規性にあたる部分である．また，実際に実証実験や応用システムを開発することまで行うことによって，本論文の有用性を検証した．それにより，Web コミュニティからアノテーションを獲得することは有用であることを示した．Weblog を書く，映像を閲覧しながら掲示板型のコミュニケーションをするなどといった Web コミュニティ活動は，ユーザによって自発的に行われる活動であり，実質的なコストを 0 として捉えることができる．そういった意味で，映像コンテンツの自動解析技術と同等に，人的なコストが掛からず意味情報の抽出が可能になる点が新しい．しかしながら，Web コミュニティから取得されたアノテーションのみで十分な精度が得られるとはいえない．将来的には，信号処理技術的なアプローチと，人間が介在したアノテーション的なアプローチとを融合させた仕組みを実現させることが重要である．これらの仕組みが実現することによって，コンテンツ作成者にとっても，コンテンツ閲覧者にとっても，コンテンツ権利者にとっても有益なシステムになることを期待する．

第1章　序論

近年，コンピュータの性能向上やインターネット回線の高速化にともない，映像・音楽などのマルチメディアコンテンツがWeb上で頻繁に配信・共有されている．これらのコンテンツは，専門家が作成したコンテンツだけではなく，一般ユーザが撮影・作成したコンテンツも爆発的に増加しており，それらのコンテンツをいかに効率よく配信・管理するかといった問題が顕在化している．それらの映像コンテンツは，一つ当たりの品質は必ずしも高くなく，また，多くの人にとって必ずしも有用であるとは限らないものも多い．その一方で，ブログやSNS，Wikiなどの登場により個人やWebコミュニティからの情報発信が一般化し，影響力も増している．

従来，映像配信といえば，テレビ放送や映画館などにおいて大衆に向けて一方向的にコンテンツを提供してきた．このような放送形態は大衆（マス）に向けたメディアとしてマスメディアと呼ばれている．1895年にフランスのリュミエール兄弟がシネマトグラフと呼ばれる投影式の映画を発明して以来，映像コンテンツはマスメディア型の配信を続けてきた．一度に多くの視聴者に同じ情報を提供することによってコストを下げることが可能になり，視聴者はより安い費用で視聴できる利点があるためである．その一方で，視聴者はあらかじめ用意された映像を一方向的に視聴することしかできない欠点も持ち合わせていた．近年，インターネットの登場によりこの映像配信の仕組みは大きく変革している．YouTube[38]に代表される映像共有サービスの普及により，ユーザが作成したコンテンツを，ユーザに向けて，ユーザ自身が配信することが可能になった．従来のようにコンテンツを提供する側とコンテンツを提供される側に明確な境界線が存在しない点において，まったく新しい形のコンテンツ配信の形態であるといえる．つまり，いつでも誰でもどこでも映像を配信・閲覧できる環境が整えられてきている．

その一方で，映像コンテンツはテキストコンテンツなどとは異なり，テキスト検索のように意味内容を考慮した内容検索などの高度な処理を行うことが一般に困難であ

る．映像共有サービスの普及とともに，Web 上で扱われる映像コンテンツは氾濫しつつあり，いかに効率よくこれらのコンテンツの管理・検索・要約・宣伝などの応用を実現するかという問題が顕在化してきた．

Web の世界ではテキストキーワードを用いてコンテンツを検索することが一般的である．テキストコンテンツの場合はキーワードマッチングにより内容をそのまま検索することが可能であるが，映像コンテンツの内容を直接テキストで検索することはできない．そのため，映像コンテンツの内容に対して何らかのテキスト情報を関連付けることが必要になる．つまり，これらを検索することによって間接的にコンテンツをキーワードで検索したことと同様の効果を得る仕組みである．そのため，内容検索などの高度な応用を実現するためには，その映像コンテンツの構造や意味内容を記述したメタ情報，アノテーションの取得が必須 [23] である．とりわけ，映像シーンの内容を表現するキーワードの抽出や映像シーンの重要度といった指標が重要になる．アノテーションの取得に関する従来手法としては，映像認識や音声認識などの自動解析技術を利用する自動アノテーション方式 [36] や，MPEG-7[18] に代表されるように専任の作業者が専用のツールを用いて作成する半自動アノテーション方式 [6, 33] などが知られている．しかしながら，Web 上の映像共有サービスに投稿されるような個人が作成した映像コンテンツなどの場合，手ぶれ・ピンぼけ・雑音・不明瞭な声などといった撮影者の技能の問題や，カメラ付き携帯電話やデジカメといった撮影機器の性能問題から映像や音声の品質のばらつきが大きく自動解析に基づく処理は限定的にしか利用できない．また，専任の作業者による半自動アノテーションを行うためには，視聴者が限定され，費用対効果の問題から，すべての映像コンテンツに対するアノテーションを施すことは困難である．多様なコンテンツに対して有効な精度を持つアノテーションを作成するためには人間が介在する必要があるが，はたして誰がどのようにアノテーションを作成するかという問題が残る．

さらに，ここ最近の傾向として，Web2.0 という言葉に代表されるように，テキストや画像コンテンツを扱う従来型の Web サービスに関連する技術も大きな変革を遂げてきた．AJAX(Asynchronous JavaScript + XML) と呼ばれる技術や高い処理能力を持った Web サーバの普及にともない，従来の静的な Web サイトとは異なり，必要に応じて Web ページの内容や要素を動的に更新するなどのよりインタラクティブな処理が可能になった．これにより，スタンドアローンアプリケーションに近い高度なインタフェースを持つリッチな Web サービスを提供することができる．また，Wikipedia[37]

や Social Bookmark サービスに代表されるように，ユーザ自身がコンテンツ作成者の一員となるサービスが大きな成功を収めている．これにより，ユーザにより良いコンテンツを作る貢献者の一人であるという意識が芽生え，このようなユーザ中心・コミュニティ中心型の Web サービスが一般化してきている．さらに，より機械にとっても分りやすい表現を目指す，Semantic Web[31] という概念が一般化しつつある．例えば，Weblog においてコンテンツのメタ情報を XML 形式で配信する仕組みである RSS[3] や Atom[16] が Semantic Web の概念を実現した技術の一つであり，これらのメタ情報を利用した応用システムとして，Web の更新情報を効率的に閲覧することが可能なツールである RSS リーダや，ブログに特化した検索システムであるブログ検索システム，複数のサービスを統合してあたかも一つのサービスであるかのように機能するマッシュアップなどに関連する技術などの応用例も登場してきている．つまり，よりリッチな Web サービス技術の進化と，機械的な処理を可能にする Semantic Web 技術によって，ユーザやコミュニティによる協調的なコンテンツ作成や流通を効果的に支援している点が興味深い．コンテンツ作成や流通におけるコミュニティの力が従来以上に，重要度を増してきている．これらの技術は主にテキストや画像・Web ページに対して適用されている．例えば，オンライン地図サービスの Google Maps[13]，高度な機能を持った Web メールである GMail[11]，写真の部分に対してコメントを付与できる写真共有サービスである Flickr[10]，ブログ検索サービスである Technorati[34]，ソーシャルブックマークサービスである Del.icio.us[7] などといった各種サービスが実際に提供されており，Web2.0 型の Web サービスとして広く利用されている．本論文では，これらの技術や Web コミュニティの力を映像コンテンツに対するアノテーション基盤として利用できないかという点に注目した．

これらの Web や映像技術の現状を踏まえて，映像配信に望まれることは，1) ユーザ作成コンテンツにも適した映像配信の仕組み，2) 映像コンテンツをより高度に扱うための仕組み，3) これらを実現する具体的なアプリケーションを示すことである．これらの問題を解決するための重要な技術の一つがアノテーション技術であると考えた．本論文におけるアノテーションとは，1) 機械処理のためのコンテンツの意味内容記述という側面と，2) コンテンツを話題としたユーザ間のコミュニケーション履歴という二つの側面がある．どちらも，コンテンツの内容に関連した情報の記述という点で同じである．本論文では，アノテーションの後者の特徴，つまり，コンテンツを話題としたユーザ間のコミュニケーションから，いかに，前者の特徴，つまり，コンテンツ

の意味内容記述を取得するかという問題についても議論する．別の見方をすれば，アノテーションは人と機械を結びつけるための技術であり，また，人と人を結びつけるための技術でもある．コンテンツとコミュニティとアノテーションは組にして扱うべき問題であると考えている．

そこで，本論文では，まったく新しい映像配信と映像アノテーションの仕組みを提案する．それは，映像コンテンツとそれらを取り巻く Web コミュニティとを効果的に融合させる仕組みを提案し，それらのコミュニティにおけるユーザの自然な Web コミュニティ活動からコンテンツに関する知識をアノテーションとして獲得する仕組みである．具体的には，映像を話題としたコミュニティ支援のための 2 つの仕組みを提案する．１つは，映像コンテンツの任意のシーンに対して，コンテンツの内容に対する感想や評価などの情報の関連付けを支援する掲示板型コミュニケーションの仕組みであり，もう１つは，任意の映像シーンを引用したブログエントリの作成を支援するブログ引用型コミュニケーションの仕組みである．これらの仕組みを提案することによって，ユーザ同士の映像を題材としたコミュニティ活動を支援する．さらに，コンテンツの内容とこれらの情報とを詳細に結び付けることによって，コンテンツに付随する様々な情報をアノテーションとして獲得する．このような方式ならば，映像の内容は人間によって理解しアノテーション作成にかかる実質的なコストは 0 となるため，前述した自動・半自動アノテーションの問題を回避できると考えた．これらを実現するための要素技術として，映像シーンに対するアノテーションに関する技術，映像シーンの引用に関する技術，映像コンテンツを Web 上で効率よく扱うための技術とこれらアノテーションに基づく応用に関する技術について議論する．またこれらの技術に基づく，より使いやすい Web ベースのインタフェースやシステムについて考察した．

これらの研究成果に基づく具体的な映像アノテーション基盤のための Web サービスとして Synvie を開発した．Synvie は，(Syndication, Synchronization, Synergy, Synthesis) などの単語で使われる接頭語の Syn と，(Movie, View) などの単語で使われる vie から成る造語であり，映像コンテンツがシナジー効果によって発展的に配信されていくことを期待している．2006 年 7 月から一般公開し，284 名の登録ユーザと，223 個の映像コンテンツと，3534 個のユーザコメントに基づくアノテーションを含む 19182 個のアノテーションの取得に成功した．また，本システムは，実際に一部の商用の映像共有サービスにも影響を与えるなどの成果もあげている．

本論文で提案する映像アノテーション基盤技術が一般社会に浸透することによって，

表 1.1: 既存の映像共有サービスとの比較

	Flickr	YouTube	ニコニコ動画	iVAS	Synvie
コンテンツ	写真	映像			
ユーザコメント対象	部分	全体	部分	部分	部分
引用対象	全体	全体	全体	全体	部分
アノテーション再利用性	×	×	×	○	○
公開時期	2004/2	2005/2	2006/12	2004/3	2005/3

全く新しいコンテンツの作成や流通が可能になると考えている．例えば，映像コンテンツに引用の仕組みを提供することによって，口コミでコンテンツに関する評判が伝わっていくことが可能になり，映像コンテンツの内容検索が少ないコストで可能になる，ユーザアノテーションも含めた形で一つのコンテンツとみなし新しい映像の視聴手段の提供が可能になる．これらの仕組みを実現することによって，コンテンツ作成者にとっても，コンテンツ閲覧者にとっても，コンテンツ権利者にとっても有益なシステムになることを期待する．

1.1 既存の映像共有サービスの系譜

既存の映像共有サービスと本研究の関係について述べる．アノテーションの作成や利用を目的として本研究で開発した Synvie と，動画共有を目的とした既存の映像共有サービスは本来異なる目的で開発された仕組みである．そのため，これらの仕組みを比較することは適切ではないが，近年の動画共有サービスの普及に伴い比較検討する必要があると考えた．

Web における映像配信の歴史は 1997 年に Real Networks 社によって発表された Real Video にさかのぼり，1998 年に Microsoft が Windows Media Video を発表した．また 1998 年に SMIL[35] と呼ばれる映像を含むコンテンツ記述言語が W3C によって標準化されるなど，Web における映像配信に関する関連技術が大きく進歩した．2000 年頃になると商用の映像コンテンツを Web 上でストリーミング配信するインターネット TV と呼ばれるサービスが普及してきた．しかしながら，Web におけるユーザ投稿型の映像共有サービスの歴史はまだ浅い．本研究を開始した時点において，映像の

Web 共有サービスは存在しなかった．最初の映像共有サービスとして有名な YouTube は 2005 年 2 月に登場したが，筆者が提案した最初の映像に対する Web アノテーションシステムは，2003 年の 4 月頃から検討しはじめ，2004 年 3 月に学会にて口頭発表を行った．そして，2005 年 1 月に最初の雑誌論文として掲載された．YouTube では映像全体のみにユーザコメントを付与することができたが，本研究によって開発した iVAS では映像の部分に対してコメントが付与可能であること，アノテーションとしての再利用を前提とした点において技術的にも研究的にも先行していた．その当時は，果たして本当に Web ユーザが映像の部分にコメントを自発的に付与するのかという疑問が多くの人に持たれたが，本研究によって開発した iVAS とその後継である Synvie を参考にし，映像の部分にコメントを付与することができるニコニコ動画 [39] と呼ばれるサービスが 2006 年の 12 月に登場したことによってその疑問を晴らした．ニコニコ動画の詳細については 7 章 4 節で説明するが，ニコニコ動画では，会員数 400 万人を超え，7 億以上ものユーザコメントが付与された．このコメント数は，YouTube などのコンテンツ全体にしかコメントを付与できないシステムよりも圧倒的に多く，本研究によって提案した映像の部分に対してコメントを付与するという仕組みが，アノテーションの再利用を目的とした研究的側面だけではなく，エンターテインメントや新しいコミュニケーション手段としても社会に多大なインパクトを与えたのは間違いない．また，YouTube は Flickr と呼ばれる写真共有サービスに大きく影響を受けたサービスであるが，Flickr が公開されたのは 2004 年 2 月であり，筆者が研究をはじめた後である．なお，本研究によって開発した Synvie や iVAS の技術的・研究的な優位性について表 1.1 にまとめ，映像共有サービスと本研究との時間的な流れを図 1.1 に示す．

1.2　論文構成

最後に，本論文の構成について述べる．

2 章において，映像コンテンツに対するアノテーションに関するモデルをいくつか述べる．これらのアノテーションモデルに基づき，次章以降に詳しく述べる．3 章において，オントロジーに基づく映像アノテーションの仕組みについて述べる．これは，Web に公開されていないクローズドな映像コンテンツに対して効果的なアノテーション手法である．4 章においては，閲覧者が映像アノテーションに参加する仕組みを提

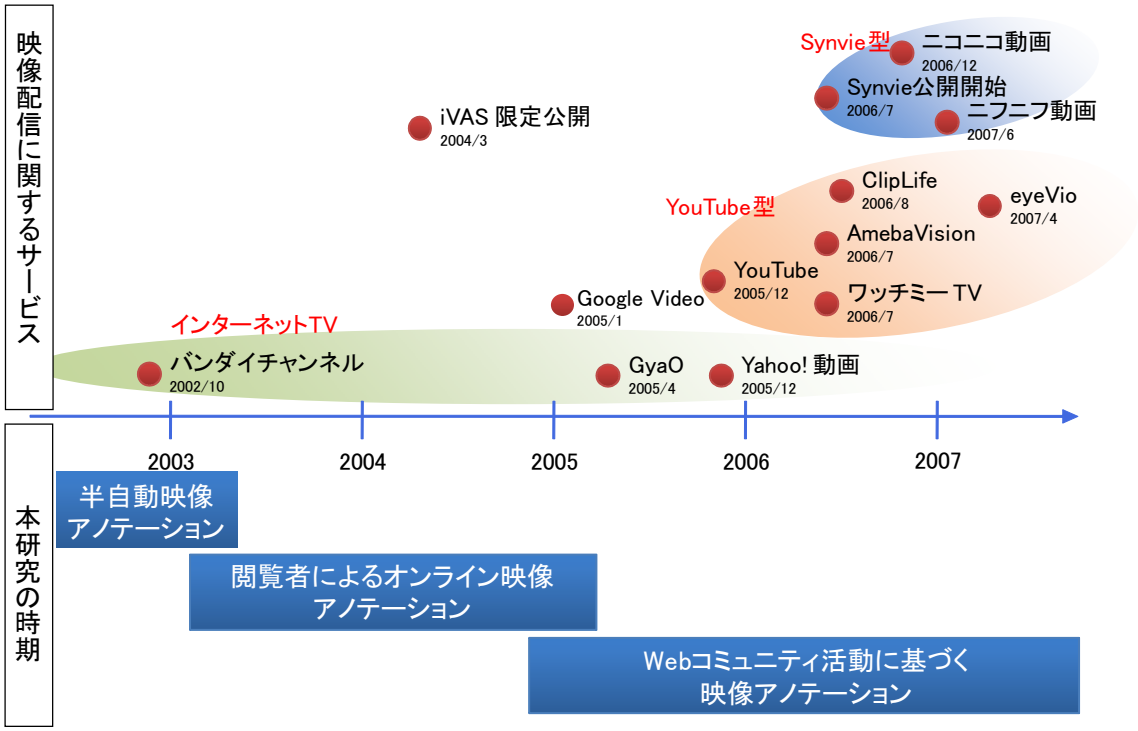

図 1.1: 映像共有サービスと本研究の系譜

案する．これは，Web に公開されているオープンな映像コンテンツに対して効果的なアノテーション手法である．また，アノテーション信頼度と印象距離という 2 つの指標についても提案する．5 章において，4 章の考え方に基づき，アノテーション参加者を映像を中心とした Web コミュニティ全般に拡張した仕組みを紹介する．具体的には，映像コンテンツの部分引用の仕組みとそれを実現するためのインタフェースを提案する．また，大規模な実証実験を行い，取得されたアノテーションに基づき量と質の観点から分析を行う．6 章では，5 章で獲得したアノテーションの分析に基づき，映像を中心とした Web コミュニティから獲得したアノテーションを用いたいくつかの応用システムについて提案する．7 章で関連研究を述べた後，8 章で本研究についてまとめる．なお，本論文において一番述べたい章は 5 章であり，時間的な余裕がない読者は 5 章を中心に読んで頂きたい．

第2章　映像コンテンツとWebアノテーション技術

本章では，映像コンテンツへのアノテーション [22] を作成するための基盤となる技術について述べる．本論文でいうアノテーションとは，コンテンツに関連付けられたメタ情報と定義し，多くの場合は，そのコンテンツの意味内容を記述したテキストや数値からなる構造化された情報である．しかしながらアノテーションに関する研究について系統的に述べられた文献は少ない．そこで，1) アノテーションをいかに作成するか，2) いかなるデータ構造にするか，3) どのようにコンテンツと関連付けるかという観点から議論する．

一般に，映像の内容検索や要約などの高度利用のためには，映像コンテンツの意味内容を抽出・付与する手段が必須である．例えば，テキストコンテンツと違って，映像コンテンツの内容検索を行う場合は，そのままではキーワードを用いて映像コンテンツの内容を検索することはできない．映像コンテンツの意味内容や構造を表すアノテーションを関連付け，そのアノテーションを検索することによって間接的に映像コンテンツを検索することができる．そのための手段として，2 つの手法が考えられている．1 つは，映像コンテンツの自動解析技術に基づく信号処理的アプローチであり，もう 1 つは，人手によるメタ情報を付与するアノテーション的アプローチである．前者は，人間によるコストがかからないという利点がある一方，解析精度がコンテンツに大きく依存し，また，抽出できる情報の種類も限定的であるという欠点がある．後者は，対象となるコンテンツへの依存度が低く作成できるアノテーションの種類が多いという利点がある一方，人間のコストが大きいという問題がなる．どちらのアプローチが優位かは常に論争の的になりやすいが，本論文は後者のアノテーション的アプローチを選択した．なぜならば，映像コンテンツは現在のところその作成過程において映像に関する意味情報を機械的に埋め込む手段がなく，人間の判断を利用しなければその意味を確定できないと考えているためである．

2.1　映像コンテンツの構造化

映像の部分要素にアノテーションを関連付けるためには，そのコンテンツの内部要素を一意に特定する必要がある．そのためには，いかに映像の内部構造を構造化するか，また，映像の部分を一意に特定するための記述形式をどのように定義するかという点が問題になる．そこでまず，映像の構造化について議論し，映像の部分に対する記述形式として映像シーンへの Permalink[1] について提案する．さらに，映像シーンへのアノテーションの基本的な仕組みについても提案する．

2.1.1　映像の構造化

映像コンテンツを効率よく扱うためには，映像コンテンツを機械にとっても人間にとっても分りやすい形で構造化する必要がある．映像の構造に関連する用語として，シーンやフレーム，カットやショットなどといった各種用語が存在するが，これらの用語は日常生活の中では曖昧に用いられている場合が多い．そこで，これらの映像に関する用語について整理する．図 2.1 に示すように，映像コンテンツは，時間軸と連動する複数の画像列から成る．個々の画像をフレームと呼ぶ．カメラの切り替わりなどフレームの色調が大きく切り替わる境界をカットと呼び，カットから次のカットまでのフレーム列をショットと呼ぶ．さらに，映像コンテンツの意味的なまとまりかつ連続するショット列をシーンと呼ぶ．さらに，大きな意味的なまとまりを持つシーン列をシナリオと呼び，映像コンテンツは複数のシナリオから成立する．また，映像コンテンツには 1 つ以上の映像と連動する音声トラックが含まれている場合もある．映像コンテンツの中で扱う時間軸を特にメディア時間と呼び，一般に映像の最初のフレームを 0 秒とする．

しかしながら，すべての映像コンテンツはこのような単純な木構造で定義できるとは限らない．一般の映像コンテンツの中には，シーンが時間的に重複したり，複数のシナリオが同時に進行する場合もありうる．また，シーンやシナリオは作成者の主観によるところが大きく，人によって，または見方によって大きく変わる場合もあり，必ずしも一意に決定できるとは限らない．さらに，通常，これらの構造情報は映像のバイナリデータの中には含まれておらず，映像コンテンツとは別のところでこれらの情報を保持する必要があるなどの欠点も存在する．

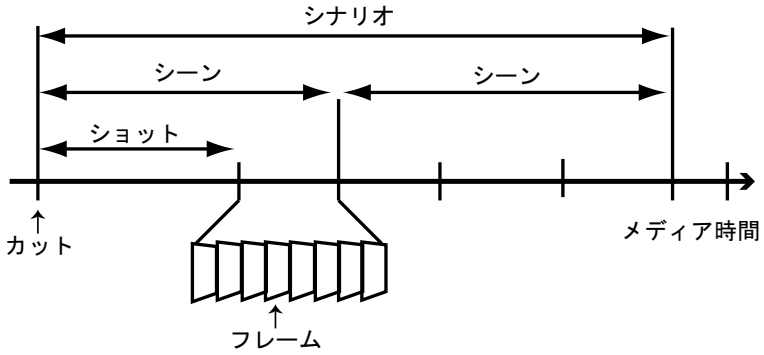

図 2.1: 映像の構造化

映像の任意の部分を一意に特定することを考えた場合，このような構造情報を用いて特定する場合と，単純にメディア時間によって特定する二つの手法が考えられる．前者はシーンやシナリオ単位で映像を特定する手法であり，映像に関連付けられたこれらの構造情報が存在する場合，より意味的に映像の部分を特定することができる利点がある．その一方で，これらの構造情報は主観や目的によって変わる場合があるため客観性という点では劣る．後者は，メディア時間によって映像の部分を特定する手法であり，開始メディア時間と終了メディア時間で特定する．映像コンテンツを意味的に扱う場合には不向きであるものの，客観的に映像の部分を一意に特定できるという利点がある．映像を一意に特定することのみに重点を置く場合には，後者の手法の方が汎用性が高く優れている．

2.1.2 映像シーンへの Permalink

映像の任意のシーンに対してアノテーションを関連付けるためには，それらのシーンに対して固有の ID を定義する必要がある．本論文では，Web 上に存在する映像コンテンツを対象としているため，既存の Web 技術との親和性が高い仕組みであることが望ましい．一般に Web サイトの場合は，URI(Uniform Resource Identifier) という形式で Web サイトを一意に特定することができる．既存の Web サイトは URI が同一でも内容が頻繁に書き換えられる可能性があるため，URI と Web サイトの内容は必ずしも一致しないことが多い．しかしながら，最近は Weblog の普及に伴い Permalink という概念が一般化しつつあり，リンク切れ防止のため URI の内容は一度公開したら

原則変更しない，1 つの URI には 1 つのコンテンツのみ存在している方が良いという考え方が定着してきた．リンク切れ防止という観点だけでなく，アノテーションという観点からもこの考え方は有用である．なぜならば，アノテーション対象の内容が変更されるとそのアノテーションとの整合性が保たれなくなるためである．そのため，Permalink という概念はアノテーションにとって一番重要な仕組みである．この概念を映像コンテンツの，しかも内部要素に対しても適用するために，本研究では梶ら [45] によって提案されている Element Pointer の仕組みを採用した．Element Pointer は任意のコンテンツの部分要素に対して URI を関連付ける仕組みであり，それぞれのコンテンツの URI が一意であることが保証されている．

映像コンテンツ全体に対する Permalink は以下のように，固有の ID を用いた URI を記述する．

```
http://[server]/[content ID]
```

任意のシーンに対する Permalink は，以下のように固有の ID とその時間区間を記述する．

```
http://server/[content_id]#epointer(urn:aps:timeline(begin,end))
```

また，複数の時間区間に対する Permalink を記述する場合は，コンマで区切って複数記述する．

```
http://server/[content_id]#epointer(
    urn:aps:timeline(begin,end),urn:aps:timeline(begin,end), ...)
```

これらの仕組みにより，映像の任意の時間区間に対して，固有の Permalink を記述することができる．この URI をキーとすることによって映像の任意の要素へアクセスすることが可能になる．

2.1.3　映像シーンへのアノテーション

Web コンテンツにアノテーションを関連付ける枠組みとして RDF(Resource Definision Framework)[21] と呼ばれる仕組みが存在する．RDF は，URI で記述可能な任意の Web コンテンツを主語，関連付けるメタ情報を目的語，主語と目的語の関連性を

述語とし，主語・目的語・述語の 3 つ組で記述される．通常の RDF の主語に映像シーンへの Permalink を記述することによって，RDF の仕組みを映像コンテンツに適用可能になる．アノテーションを記述するための枠組みとしては他に，MPEG-7 を利用することも考えられるが，Semantic Web[9] との親和性や，アノテーションのためのより汎用的な枠組みである RDF を採用することによって，映像コンテンツと Web コンテンツを統一的に扱うことが可能になる利点がある．

記述形式は以下のような XML になる．

```
<rdf:Description rdf:resource="主語 (Permalink)">
  <述語>目的語 (アノテーション)</述語>
</rdf:Description>
```

例として,「長尾研究室」という情報を，映像シーンの 10 秒から 20 秒まで,「place」という関係で関連付ける場合は以下のように記述する．

```
<rdf:Description rdf:resource=
   "http://server/a.avi#epointer(urn:aps:timeline(10,20))">
  <place>長尾研究室</place>
</rdf:Description>
```

2.1.4 映像シーンの引用

次に映像シーンの引用 [50] について提案する．本論文で述べる引用とは，任意のコンテンツの任意の要素から任意のコンテンツの任意の要素へと関連付ける行為であると定義する．引用されたコンテンツの要素を引用元と定義し，他方のコンテンツの要素を引用先と定義する．HTML ではハイパーリンクを作成することによって任意のコンテンツへのリンクを作成することが可能になり Web の発展において重大な影響を与えてきた．引用とは，ハイパーリンクと対を成す考え方であり，コンテンツの任意の要素を的確に引用できる仕組みを作成することによって，アノテーションの技術とともにさらなる Web の発展に大きな影響を技術であると考えている．しかしながら，技術的困難さから引用を支援する仕組みは今まで存在しなかった．我々が想定する引用の例としては，映像の任意のシーンを引用したブログエントリーの執筆があげ

られる．具体的には，映像シーンの内容を表すサムネイル画像及び映像シーンへのリンクとそれに関連するコメントの組から成ることを想定している．この場合，引用元のシーンと引用先のブログエントリーの段落を関連付けることができる．論文の引用などとは違い，引用元コンテンツ全体を引用するのではなく，引用元コンテンツの該当する箇所のみを切り取って引用する仕組みである．

映像コンテンツを引用する場合，テキストの引用などと違って，いくつか解決しなければならない問題がある．これらの問題に関する具体的なシステムは 5 章において詳しく述べる．アノテーションの観点から見た場合，映像シーンの引用は，映像シーンへのアノテーションを包含する情報と見なせる．また，コンテンツとコンテンツを関連づけるリンクとしての側面も持ち合わせている．

2.2　映像アノテーションとデータ構造

次に，映像アノテーションにおけるデータ構造について議論する．本論文におけるアノテーションとは，コンテンツに関連付けられたメタ情報と定義しており，メタ情報の形式によりいくつかのアノテーション型が考えられる．従来型のアノテーション型としては，スキーマ型アノテーション，オントロジー型アノテーション，タグ集合型アノテーションの 3 つの方式がある．スキーマ型アノテーションは MPEG-7 などの規格で採用されており，一般的に利用されているアノテーション手法の一つである．オントロジー型アノテーションは，Web コンテンツを知的に扱うための枠組みである Semantic Web を実現するための重要な要素であるオントロジー [47] を用いており，推論や学習などの知識処理に向いたアノテーション型である．タグ集合型アノテーションは，画像共有サービスである Flickr やブックマーク共有サービスである Del.icio.us などで採用された，Web コミュニティ向けに利用されているアノテーション型である．また，本論文で提案する仕組みとして，ユーザコメント型アノテーションと引用型アノテーションを提案する．これらは，映像を中心とした Web コミュニティユーザによって付与されるアノテーションであり，ブログを書く，映像の部分引用をする，マーキングをするといった自然な活動によって暗黙的に作成される．それぞれのアノテーション型には一長一短があり，用途に応じてこれらを複合的に利用することが望ましい．

2.2.1 スキーマ型アノテーション

スキーマとは，項目とそれらの型について定義したデータ構造の1つである．リレーショナルデータベーススキーマやＸＭＬスキーマなど，データ構造の形式としては幅広く使われており直感的かつ実用的である．たとえば，映像のとあるシーンに登場する人物についてのアノテーションを作成することを想定した場合，人物スキーマというデータ構造を定義し，そのスキーマに従って情報を入力する．人物スキーマの例として，図 2.2 に示すように，名前，年齢，性別，所属，現在の状態や行動の項目からなるスキーマをあげる．それぞれの項目には，文字列型，数値型などの適切な型が定義されておりそれ以外の情報を記述することができない．一般的に，一人ないし少数の専門家がこれらの項目に手作業で値を入力し，映像の部分と関連付けることによってアノテーションを作成する．

スキーマ型アノテーションの利点としては，このようなスキーマをあらかじめ定義することによって，アノテーション情報を形式的に扱うことができる，リレーショナルデータベースや XML などとの親和性が高い，応用目的にそったスキーマを定義することによって応用システムの開発が容易であるという利点があげられる．欠点としては，人間が情報を記述することが一般的であるため，表記ゆれや入力ミス等の人為的なミスが発生しやすいという点があげられる．

スキーマ型アノテーションを用いたシステムや研究の例としては，MPEG-7[18]，それを記述するための関連ツールとして MovieTool[29] や VideoAnnEx[20] などがあげられる．

2.2.2 オントロジー型アノテーション

オントロジーとは，概念の関係をグラフで記述した知識ベースであり，Semantic Web の中核を成す技術として注目されている．例えば，教員という概念を表現する場合，図 2.3 で示すようになる．教員という概念の下位概念には，大学教員，高校教員などが存在し，大学教員の下位概念には，教授，准教授，講師，助教などが存在する．さらに大学組織オントロジーを考えると，大学という概念の下位概念には，国立大学，私立大学が存在し，国立大学の下位概念として名古屋大学が存在し，名古屋大学の下位概念には，情報科学研究科や情報メディア教育センターなどが存在する．長尾確教授は，教員オントロジーの教授という概念を持つと同時に，大学組織オントロジーの

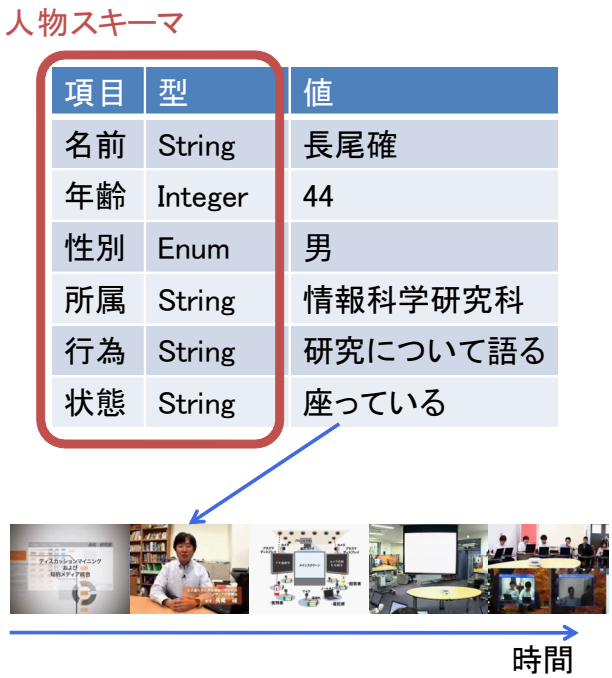

図 2.2: スキーマ型アノテーション

情報科学研究科や情報メディア教育センターの概念を持つ．この長尾確というインスタンスと映像シーンを関連付けることによってアノテーションを作成する．

オントロジー型アノテーションの利点としては，概念グラフと映像コンテンツの要素と関連付けるという作業を行うだけであるので，タイプミスに由来する入力ミスや表記ゆれという問題が発生しない，概念はグラフ構造で表現されているため，概念間の関係を利用した推論や学習などの知識処理が可能であること，RDF ベースで開発されたシステムとの親和性が高い点があげられる．欠点としては，オントロジーを作成することはコストが高いこと，また，高い知識をもった人でなければ正しくオントロジーを作成できないという点である．先の例は極めて分りやすいオントロジーであるが，例えば現実社会全般のオントロジーを作成することは哲学的な困難さもあり一筋縄ではいかないのが現状である．この問題を解決するためには，特定の領域と特定の目的に限定することによって作成するオントロジーの量を少なくする領域オントロジーという考え方が存在する．

オントロジー型アノテーションの例としては，我々が 3 章で解説するビデオアノテー

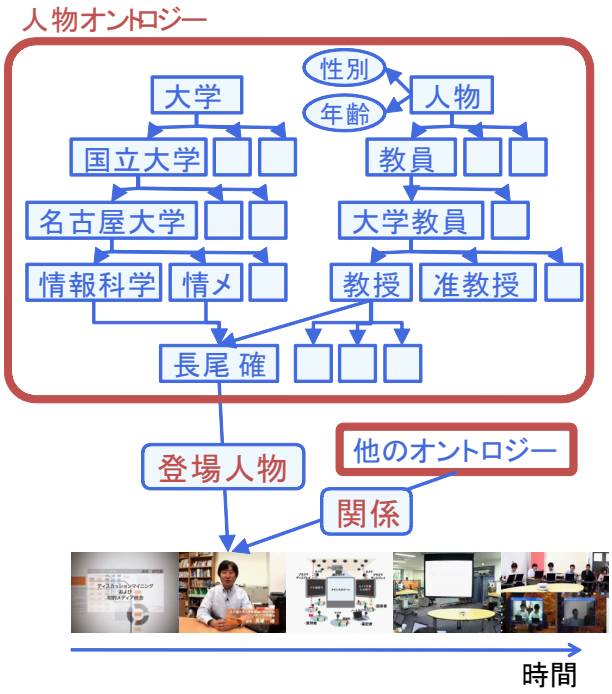

図 2.3: オントロジー型アノテーション

ションエディタ [48] があげられる.

2.2.3 タグ集合型アノテーション

タグ集合とは，コンテンツの閲覧者や製作者がコンテンツの分類やブックマークなどのために付与したコンテンツの内容に連想するキーワードの集合である．これらのキーワードをタグという．複数のユーザによってタグが付与されるという性質上，複数のユーザによって同一のタグが付与されることがあり，より多くのユーザによって付与されたタグほど重要視する．タグ集合型アノテーションの例としては，図 2.4 で表すように，長尾先生が登場している映像シーンに対しては，ユーザがタグとして，「長尾教授」「名古屋大学」「研究紹介」「長尾研究室」「ＩＴ」「面白い」などといったタグを付与することを想定している．タグ集合を可視化したものをタグクラウド [30] と呼ぶ.

タグ集合型アノテーションの利点としては，キーワードを付与するという簡単なア

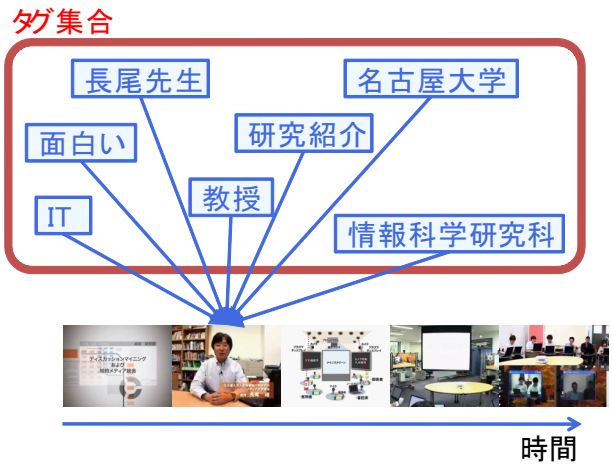

図 2.4: タグ集合型アノテーション

ノテーションであるため，誰でも容易にタグを付与することができること，ソーシャルブックマークシステムの Del.icio.us や映像や写真共有サービスの Flickr でも実用化されている点，さらにそれらのサービスにおいて，検索やタグクラウドといった応用例が運用されている点があげられる．欠点としては，タグ集合の品質はタグを付与するユーザの数や質に大きく依存するため，場合によってはノイズが大きいこと，タグ集合という曖昧な情報記述手段であること，新しい研究分野であるため学術的な理論が確立されていない点があげられる．また，映像シーンに対してタグを付与するという習慣がないため，ユーザが映像シーンにどれだけタグを付与してくれるかが未知数という問題点もある．

タグは一般的に単語によって構成されているため，それらの単語とシソーラスを関連付けることによって，上位語，下位語などの関係を取得することが可能になり，言語的なオントロジーと関連付けることが可能になる．

タグ集合型アノテーションの例は，多くの Web2.0 型サービスで利用されており，たとえば，Flickr[10] や YouTube[38] などのコンテンツ共有サービスなどがあげられる．

2.2.4 ユーザコメント型アノテーション

ユーザコメントとは，コンテンツに関連した情報や感想などをユーザがコミュニケーション目的によって投稿する自由コメントのことである．つまりユーザコメントをコンテンツの要素と関連付ける仕組みが，ユーザコメント型アノテーションである．ユーザコメントはタグとは違って多くの場合短い文章から成る．アノテーションの仕組みとしては一番単純な仕組みではあるものの，ユーザコメントの自然言語解析を行うことによってより多くの知識を取得できる可能性がある．利用例としては，図 2.5 のように，長尾先生が登場しているシーンに対して,「名古屋大学情報科学研究科教授の長尾先生のＩＴ関連の話です．面白い.」というコメントを投稿することを想定している．

ユーザコメント型アノテーションの利点としては，誰でも簡単にコメントを投稿できること，自然言語解析技術を適用することによって何らかの意味抽出が可能であるという点があげられる．欠点としては，ユーザの自由度が高いため関連のない話題や誤字や文法的に正しくない文章が入力されるなどのノイズが多く，また，それらに高度な自然言語解析処理を適用した場合必ずしも解析結果の精度が高いとはいえない点があげられる．具体的な自然言語処理の適用例としては，ユーザコメント文からキーワード集合を抽出し，それをタグ集合として扱うことによって，間接的にユーザコメント型アノテーションをタグ集合型アノテーションとして扱うことを検討している．

これらの仕組みの具体例については 4 章で説明する．

2.2.5 引用型アノテーション

本論文で述べる引用とは，先で述べたとおり，任意のコンテンツの任意の要素から任意のコンテンツの任意の要素へと関連付ける行為であると定義する．また，引用されたコンテンツの要素を引用元と定義し，他方のコンテンツの要素を引用先と定義する．

たとえば，テキストの引用の場合は，自身の主張を展開する上で有用な他の文献の一部を自身のテキストに引用符を付けて引用することが一般的に行われている．書評を書くことを考えた場合，引用されたテキストの内容と，引用先である自身の書評の内容には何らかの関連性があるとみるのが自然であろう．同様に，映像コンテンツに関するブログエントリーを執筆する場合には，映像の部分を適切に引用することさえできれば，その映像の部分の内容とブログエントリーの内容に何らかの関連性がある

図 2.5: ユーザコメント型アノテーション

と考えるのは妥当であろう．つまり，引用という行為には，多くの場合，引用元の部分と引用先の部分の内容に関連性があると仮定することができる．

引用型アノテーションとは，引用元と引用先の内容に関連性があると仮定した場合，引用先の周囲にある情報を引用元の映像コンテンツの部分要素に対してアノテーションとして捉え，関連付ける仕組みである．半ば，強引な手法であるように思われるが，5 章において引用型アノテーションの有意性を示す．本論文では引用先にはブログエントリーを想定しているが，それ以外のコンテンツでも構わない．たとえば，図 2.6 のように，料理番組においてある店の紹介に関するシーンを引用したブログエントリーを執筆した場合，引用された箇所に近い文を映像シーンに対するユーザコメント型アノテーションとして利用することを想定している．

引用型アノテーションの利点としては，コンテンツに関連するブログエントリーから詳細な情報をユーザコメント型アノテーションとして取得可能であること，文章内容が掲示板やチャット文よりも詳細に記述されていることが期待できる点などである．さらに，映像の部分と外部 Web サイトの部分を引用によって関連付けることによって，

図 2.6: 引用型アノテーション

既存のハイパーリンクによる Web ネットワークよりもより詳細かつ広範囲な，部分引用に基づく Web ネットワークの構築が可能になる点も重要である．これらの Web ネットワークを解析することによって，コンテンツ間の関係や類似性の発見などの応用に利用可能である．

これらの仕組みの具体例については 5 章において説明する．

2.3 Web アノテーションの共有

Web コンテンツに対するアノテーションの共有手法について述べる．アノテーションを保持するデータベースをどこで保持するかによって，Hirotsu[15] らは，クライアントモデル，プロキシモデル，サーバモデルの 3 つの方式を提案している．クライアン

トモデルは，アノテーションの情報をクライアント側，つまり Web ブラウザを用いてアノテーションを管理し，クライアント間で互いに通信を行うことによってアノテーションを共有する方式である．既存のインターネットの仕組み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 Web ブラウザを用意する必要があり，またアノテーションの共有が技術的に難しいという欠点がある．サーバモデルは，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，クライア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，すべ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．プロキシモデルは，Web 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ものの，プロキシにアクセスが集中しボトルネックになりやすいという欠点がある．

どのモデルを利用するかは用途次第であるが，比較的小規模なユーザグループにおいて非常に多くの Web サイトにアノテーションを付与したい場合はプロキシモデルを，動画共有サービスのように非常に多くのユーザ間でアノテーションを共有する必要がある場合にはサーバモデルが有用である．クライアントモデルは，現状では利用できる状況が限定的であると考える．

2.4　アノテーションに基づく応用

次にアノテーションに基づく応用について議論する．映像コンテンツの内容に基づく応用，たとえば，映像シーン検索を実現するためには，映像コンテンツの意味内容を抽出する必要がある．しかしながら，自動解析処理によって取得できる映像の意味内容は限定的であり，シーン検索などの用途には十分ではない．アノテーションに対応する映像シーンの内容を表す情報が含まれているので，これらのアノテーションを対象に検索などの処理を行うことによって，間接的に映像の検索などの応用が可能になる．

具体的な応用例については 6 章で述べる．

2.5 本章のまとめ

本章では，映像コンテンツとアノテーションに関する研究の基礎となるモデルや考え方について述べた．

はじめに，映像コンテンツに関する用語について定義した．次にアノテーションについて述べた．アノテーションとはコンテンツの意味内容や構造情報を記述したメタ情報であり，アノテーションの内容に基づく処理を行うことで，間接的にコンテンツの内容に基づく処理を行うための仕組みである．次に，いくつかのアノテーション型について述べた．アノテーション型とは，アノテーションを格納するデータ形式であり，具体的には，スキーマ型アノテーション，オントロジー型アノテーション，タグ集合型アノテーション，ユーザコメント型アノテーション，引用型アノテーションの5つの仕組みを述べた．前3つのアノテーション型は，我々が3章で解説するものも含めて，多くのシステムや研究で採用されている従来型のアノテーション形式である．後2つのアノテーション型は，Webユーザが自由コメントによってアノテーションを付与する形式であり，アノテーションの仕組み自体は単純であるものの，アノテーションとしての再利用を目的とした研究を行った例は少ない．特に引用型アノテーションに関してはオリジナリティが高く，詳細な意味情報を獲得できる可能性を秘めているだけではなく，他コンテンツとの関連性や構造情報を取得できる可能性がある点において重要な技術である．これらのアノテーション技術についてまとめると表2.1のようになり，特に作成コストの点や拡張性の点において，我々の提案手法の優位性がある．タグ集合型アノテーション，ユーザコメント型アノテーション，引用型アノテーションの作成コストが○である理由は，それぞれ，ソーシャルブックマークやブログ執筆などといったユーザ自身の目的を達成する結果としてアノテーションが作成されるため，アノテーション作成に関わる実質的なコストが低いためである．

また，アノテーションを共有するためのモデルやアノテーションに基づく応用についても言及した．

本章では映像に対するアノテーションのモデルのみしか述べておらず，具体的な実装例は次章以降に述べる．スキーマ型アノテーションとオントロジー型アノテーションに関しては3章で，ユーザコメント型アノテーションについては4章で，引用型アノテーションに関しては5章において述べる．

次章では，まず，スキーマ型アノテーションとオントロジー型アノテーションの実

表 2.1: アノテーション型別の比較

	スキーマ	オントロジー	タグ集合	ユーザコメント	引用
作成者	専門家		Web ユーザ		
作成コスト	×	×	○	○	○
拡張性	×	△	△	○	○
信頼性	○	○	△	×	△
再利用性	○	○	△	△	○
表現力	△	△	△	△	○
リンク情報	×	○	×	×	○
ターゲット	商業コンテンツ		Web 共有コンテンツ		

装例である，オントロジーに基づく半自動映像アノテーションについて述べる．

第3章　オントロジーに基づく半自動映像アノテーション

本章では，映像コンテンツに対するアノテーションを作成するアプローチの1つとして，オントロジー型アノテーションとスキーマ型アノテーションに基づく半自動映像アノテーションの仕組みを提案する．とりわけ，Web上に公開されていないクローズドな映像コンテンツに対してアノテーションを作成するのに適した仕組みである．

一般に，信頼性の高いアノテーションを作成するためには，自動解析技術のみに頼るのではなく，人手によってコンテンツの意味情報をアノテーションとして作成することが有効である．つまり，映像コンテンツに対して，コンピュータによる自動解析と人間によるアノテーション作成を効率よく連携させる仕組みが有効である．映像コンテンツに対するアノテーションに関する研究はいくつかあるが [6, 29, 33, 23]，Webオントロジーの概念を導入したアノテーションの仕組みはあまりない．映像コンテンツに対してアノテーションを作成するうえで問題となる点は，大きく分けて3つある．1つはアノテーションを作成する負担が大きいこと，2つ目は作成したアノテーションが特定のアプリケーションのみに依存しがちであること，3つ目は作成したアノテーションの内容が人間にとっても機械にとってもに一意に内容を把握できる仕組みでとはいえない点である．アノテーションを作成する負担を下げるアプローチとして，機械と人間が協調的にアノテーションを作成できる仕組みを用意すること，さまざまな場面で活用できるためのアプローチとしてXML形式での情報の蓄積を行うこと，人間にも機械にも内容が把握できるための仕組みとして同義語リストと関連付けられたオントロジーを導入した点があげられる．前2つのアプローチはアノテーションツールに共通する仕組みであるが，最後のアプローチは本研究の新規性の部分である．

そこで，本章では，人手によって映像の内部要素にアノテーションを効率よく関連付ける仕組みとして，ビデオアノテーションエディタを開発した．さらに，それらのツールを用いて作成したアノテーションに基づく，映像シーン検索システムを開発し，

図 3.1: ビデオアノテーションエディタ

我々の作成したツールの有効性について議論する．

3.1　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とは，映像コンテンツに対するアノテーションを効率よく作成するための専用のツールである．おもに，専任の作業者が，重要なコンテンツに対してアノテーションを作成することを想定している．たとえば，放送局には膨大な映像アーカイブが存在するが，それらの映像を効率よく検索する手段は整備されておらず，タイトルなどから内容を推定し，ユーザが実際に映像の中身を閲覧しなければ内容の把握はできない．将来，膨大な映像コンテンツが Web 上で公開されることも期待されるため，これらの映像コンテンツの内部要素を効率よく検索できることは有益である．

3.1.1 システム構成

本システムはWeb上の映像コンテンツに対してアノテーションを作成することを想定している．

本システムは，映像コンテンツを蓄積する映像データベース，アノテーションを蓄積するアノテーションデータベース，後述するオントロジー情報を保持するオントロジーデータベース及びアノテーションを作成するビデオアノテーションエディタから成る．

アノテーション作成者は，映像データベースのコンテンツに対してビデオアノテーションエディタを用いてアノテーションの作成を行う．作成されたアノテーションは専用のアノテーションデータベースに蓄積される．オントロジー型アノテーションの作成を行うために，映像に関するオントロジーを保持した専用のデータベースを用意し，映像コンテンツの内部要素と関連付ける．アノテーションやオントロジーはXML形式で記述されるため，これらのデータベースはフリーのXMLデータベースであるXindice[2]を採用した．これらのデータベースは複数のユーザによって共有することを前提としており，協調的にアノテーションの作成を行うことが可能になる．

映像シーン検索などの応用アプリケーションでは，これらのアノテーションを用いて実現可能になる．

3.1.2 ビデオアノテーションエディタの機能

ビデオアノテーションエディタの機能について説明する．

主に3つの機能から成る．1つは，映像の構造に関する情報を付与する構造アノテーションを支援する仕組みである．主に映像に関する自動解析技術とその解析結果を人手によって修正するインタフェースから構成される．2つ目は，映像の内容に対して意味情報を付与する意味アノテーションを支援する仕組みである．主に人手による情報の付与であり，スキーマ型アノテーションとオントロジー型アノテーションを支援する．3つ目はアノテーションの蓄積や管理を支援する仕組みであり，アノテーション情報をＸＭＬとして出力する仕組みや外部サーバにアップロードする仕組みである．

具体的には，構造アノテーションを支援する機能として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. オブジェクトトラッキング機能

5. 色ヒストグラムの自動抽出機能

意味アノテーションを支援する機能として以下のものを備える．

1. 映像シーンやオブジェクトに対するスキーマ型アノテーション機能

2. 映像シーンやオブジェクトに対するオントロジー型アノテーション機能

アノテーションの蓄積や管理をするための機能として以下のものを備える．

1. XML 形式での出力機能

2. アノテーションデータベースへの登録機能

以上の機能を実現することによって，アノテーションを作成するために必要な機能を網羅する．

3.2　映像の自動解析に基づく構造アノテーション

本章では，アノテーションの取得を目的とした映像コンテンツの自動解析技術に関して議論する．アノテーションに基づく応用システムを作成するために必要な情報は，構造情報と意味情報に分類される．構造情報を取得するために，映像のカット検出，ショットの重要度の推定，シーンの推定，オブジェクトトラッキングなどの技術が有用である．

しかしながら，自動解析に基づく解析結果には少なからずエラーが存在する．質の高いアノテーションを作成するためには，これらのエラーを効率よく修正するための仕組みが必要になる．そこで，自動解析システムとともに，人手による効率の良い修正インタフェースを提案する．

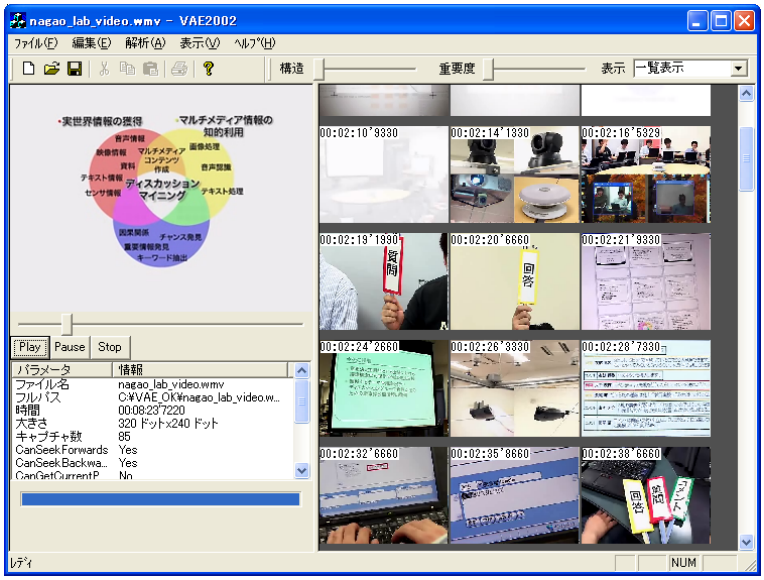

図 3.2: カット検出画面

3.2.1 カット検出

映像コンテンツに対するカットの自動検出機能について議論する．本論文におけるカットとはカメラの切り替わりなど，フレームの画像が大きく切り替わる境界のことを言い，カットの自動検出とはカットを自動画像解析により検出する手法である．また，カットとカットで分割された映像区間をショットと呼ぶ．カット検出画面例を図 3.2 に示す．

本論文で用いたカット検出のアルゴリズムは，色ヒストグラムに基づく現在のフレーム間差分と直前のフレーム間差分の比較を行い，ある閾値を超えた時点でカットとした．ここで述べるフレーム間差分とは，あるフレームとその次のフレームの色ヒストグラムの個々の値の差分の絶対値の合計のことを示す．また，フレーム間差分算出手法には，精度の高い算出手法である分割 χ 二乗検定法 [41] を採用した．

具体的には，Microsoft 社の映像解析ライブラリである DirectShow ライブラリを用いて開発した．DirectShow では MPEG1 や MPEG4，Windows Media Video などの多くの映像フォーマットに対応し，映像から順次フレーム情報を取得することが可能である．多くの映像フォーマットは 1 画素を 24 ビットで表現しているが，色ヒスト

グラムを作成するには色数が多すぎるので，RGB 各色要素 4 ビットづつの 12 ビット，4096 色に色数を圧縮し色ヒストグラムを作成する．また，フレーム間差分の変動が大きく連続する区間（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．具体的なカット検出アルゴリズムの手順は以下のとおりである．

1. DirectShow ライブラリを利用しフレーム毎に静止画像を獲得
2. 色情報を 12 ビットに圧縮し，色ヒストグラムを作成
3. 前フレームとの色ヒストグラムの要素ごとの絶対値誤差の合計を計算
4. 絶対値誤差の合計が閾値以上の場合カットであると認識
5. 以上の処理を繰り返す

このままでは，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．人間が知覚できないほど素早くカットが切り替わる場合にはそれをカットするのは適切ではない．そこで，連続的なカットを検出した場合，それらをカットとして認識しないような工夫を加えた．具体的なアルゴリズムは以下の手順である．

1. カットを検出
2. 最後にカットを検出してから n フレーム以内にカットを検出したらそのカットをカットとしない．
3. 以上の処理を繰り返す

今回の実験では n=8 とした．これにより，カット情報の取得が可能になる．

カット検出の仕組みはニュース映像などカメラ映像が頻繁に切り替わるコンテンツに向いているものの，サッカー中継や講義ビデオなど同じカメラで延々と録画しつづけた映像コンテンツなどには，そもそもカットが存在しない．しかしながら，これらの映像コンテンツにおいても意味の切れ目が存在するはずであり，それらの情報を抽出することは一般に困難である．またカット検出アルゴリズムにおいて誤検出が発生する場合も考えられる．そのため，人間の手によってカット検出結果を修正する仕組みを用意する必要がある．

カット検出結果の修正手法は以下のとおりである．本来カットではないフレームに対してカットと認識した場合，そのカットを表すサムネイル画像を選択し，コンテキストメニューから「前と結合」を実行すると，そのショットと前のショットが結合されることによって修正される．本来カットであるべきフレームがカットと認識されなかった場合は，ビデオウィンドウのシークバーを操作して対応するフレームに移動し，コンテキストメニュから「このフレームをカットとする」を選択すると新しいカットが作成される．

3.2.2 ショット重要度の推定

ショットの重要度を自動処理によって計算する手法について述べる．一般に，映像コンテンツの意味内容を自動処理によって抽出することは困難であり，意味内容を考慮した重要度の推定は困難である．そこで，映像コンテンツ全体が一様に重要であると仮定し，その映像ショットの時間の長さに応じて重要度が推定できるものとした．かなり強引な手法であるものの，人間が後から修正することを前提とした場合には有用であると考えた．具体的には以下の計算式により計算する．重要度をI，ショット時間をt，ビデオコンテンツの長さをLとする．

$$I = \frac{t}{L}$$

ショット重要度の修正インタフェースについて提案する．具体的な画面例を図3.3に示すように，ショットを表すサムネイル画像の大きさによってそのショットの重要度を表す．さらに，そのサムネイル画像をマウスで操作することによって容易に重要度を修正することが可能である．サムネイル画像の大きさとショットの重要度が対応しており，直感的に重要度の修正が可能なインタフェースである．

3.2.3 シーンの推定

一般に映像シーンは複数のショットから成る．1つのショットが1つのシーンである場合も考えられるが，映像表現手法の多様化により，1つのシーンを複数のカメラで切り替えつつ表示することが一般的に行われている．1つのシーンは1つの撮影場所で撮影されている場合が多く，環境光や背景など全体的な色味が類似している場合が

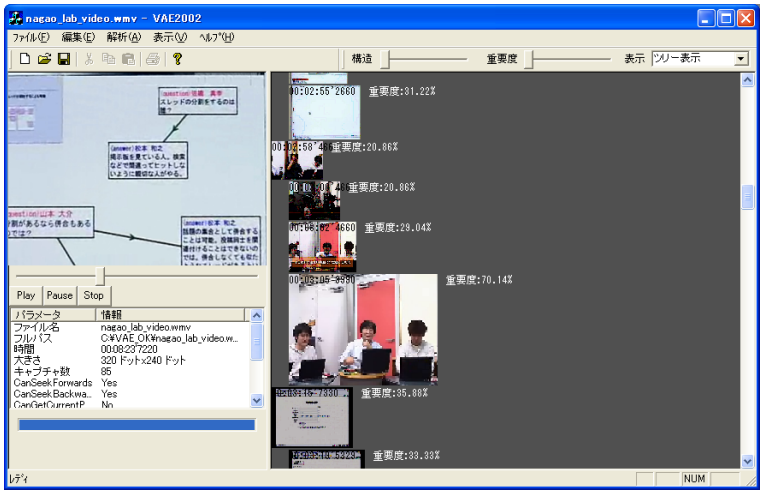

図 3.3: 構造解析と修正画面

多い．そのため，色味が類似し連続するショットを 1 つのシーンとして推定するための手法を提案する．

シーンを推定する仕組みは以下のとおりである．まず，連続するショットを $S_1, S_2, S_3, \ldots$ とすると，S_x と S_{x+1} が同一のシーンであるとの判別には，ショット S_x における全フレームのヒストグラムの合計を H_x，ショット S_x のフレーム数を m_x のとき，A_x と A_{x+1} が同一シーンである度合である P_x は，以下の式で表す．

$$P_x = |\frac{H_x}{m_x} - \frac{H_{x+1}}{m_{x+1}}|$$

P_x が閾値以下ならば同一シーンであるとした．これによって，映像シーンの推定を行う．しかしながら，当然推定誤りが存在する可能性が高い．

そこで，シーン推定を修正できるインタフェースを提案する．図 3.3 のように，重要度推定インタフェースと同じインタフェースを利用する．上下方向が時間的なつながりを示し，左右方向が構造情報を示す．マウスで左右にドラッグすることによって，多段階にサムネイル画像を左右に動かすことができる．複数のショットからなる 1 つのシーンは，最初のシーンのみ 1 つ上位の段にサムネイル画像を表示させ，残りのショットを 1 つ下位の段に表示することによってシーンを表現可能である．さらに，シーンを複数まとめてシナリオを構成したい場合には，最初のシーンのみを 1 つ上位の段に表示させ，続くシーンを 1 つ下位の段に表示させることによって表現する．このよう

に階層的に表現することによって，直感的に構造を把握しつつ容易に編集可能なインタフェースである．

3.2.4 オブジェクトトラッキング

オブジェクトトラッキングとは，映像中に現れる任意のオブジェクトを時間軸方向に追跡する技術である．追跡するオブジェクトは人や動物，物やテロップなどである．オブジェクトを追跡することによって，そのオブジェクトが登場する時間範囲を特定したり，そのオブジェクトの動作を推定することが可能になる．オブジェクトトラッキングのアルゴリズムは，パターンマッチングやオプティカルフローを計算する手法などが存在するが，本研究では，パターンマッチング法を採用した．

具体的には，ユーザが任意のフレームのオブジェクトを矩形範囲で指定する．その矩形範囲の画像をキーとして，メディア時間軸に対して未来方向と過去方向にトラッキングを行う．トラッキングが失敗した時点でトラッキングを終了し，その時間範囲がオブジェクトの存在時間範囲とする．オブジェクトトラッキングの操作画面を図3.4に示す．トラッキングした時間が正確であるかどうかを認識するために，開始フレームと終了フレームそれにその前後0.1秒のフレームを表示することにより一目で確認できるように工夫している．

3.3 オントロジーに基づく意味アノテーション

映像に関する意味情報を機械処理のみで獲得することは極めて困難である．そこで，これらの情報を主に人間の手によって付与することを支援するための仕組みが必要になる．そこで，スキーマ型アノテーションとオントロジー型アノテーションの2種類のアノテーション手法を用いて，映像シーンやオブジェクト，映像全体に対する意味アノテーションの仕組みを開発した．

3.3.1 映像シーンアノテーションのためのオントロジーの作成

アノテーションを付与する対象に基づき2種類のオントロジー集合を作成した．1つは，オブジェクトやその動作や状態を記述するためのオブジェクトに関するオントロ

図 3.4: オブジェクトアノテーションの操作画面

ジー集合であり，他方は，映像シーン全体の雰囲気や場所などを記述するためのシーンに関するオントロジー集合である．

オブジェクトに関するオントロジーの例を図 3.5 に示す．オブジェクトに関するオントロジーとしては，大きく分けて，オブジェクトの種別や所属などを表すオブジェクトオントロジーと，そのオブジェクトの行為や行動を表す行動オントロジー，オブジェクトの形状や属性を表す状態オントロジーが存在する．より正確には，オブジェクトの属性として行動オントロジーと状態オントロジーが付与可能であるといえる．図 3.5 は，それぞれのオントロジーを体系化したものである．本来ならば，これらのオントロジーのグラフと映像シーン中のオブジェクトとを直接関連付けることが望ましいが，我々のシステムでは関連付けインタフェースを二段階のプルダウンメニューからの選択によって実現しているため，それぞれのオントロジーのグラフの葉ノードとそれらの親ノードの組を最少単位の領域オントロジーとして分解した．具体的には，図 3.5 の黒背景のノードを親とする，静止オントロジー，移動オントロジー，方向オントロジー，発声オントロジー，行動オントロジー，人物オントロジー，動物オントロジー，植物オントロジー，食物オントロジー，機械オントロジー，形状オントロジー，画像オントロジー，アイコンオントロジー，文字オントロジーを定義した．それぞれのオントロジーの概念はその領域オントロジーの子要素のリストとなる．

同様の考え方において，シーンに関するオントロジーの例を表 3.6 に示す．シーンに関する領域オントロジーには，場所オントロジー，天気オントロジー，カテゴリオントロジー，雰囲気オントロジーが含まれる．これらのオントロジーにはそれぞれ多数の概念が含まれており，たとえば，天気オントロジーには，晴れ，雨，くもり，雷，雪などが含まれている．

さらに，これらの図には記述されていないが，それぞれの概念には同義語辞書とも関連付けられており表記ゆれや同義語にも対応する．

当然の疑問として，これらのオントロジーでは全ての映像を記述するためには役不足であると考えられるかもしれない．これらの問題を解決するためには，必要に応じて必要な概念を追加できる仕組みを用意する必要がある．たとえば，動物オントロジーには，うさぎや亀なども追加する必要があるであろう．また，動物オントロジーにおける犬や猫や哺乳類という共通の概念を持ち合わせているので，犬や猫の上位概念で動物の下位概念に哺乳類という概念を追加する必要があるかもしれない．そのために，本来ならばオントロジーを動的に追加するためのインタフェースも提供する必要

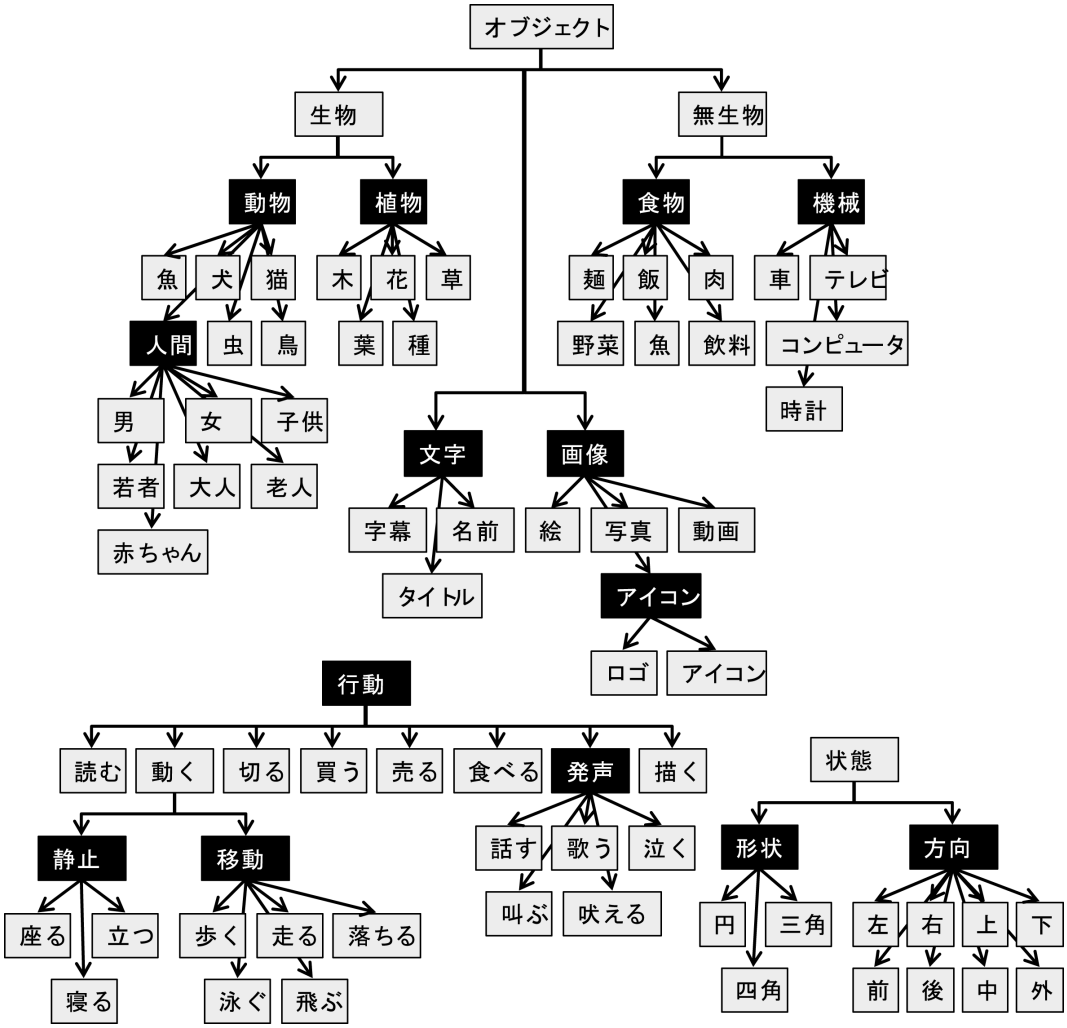

図 3.5: オブジェクトに関するオントロジー集合

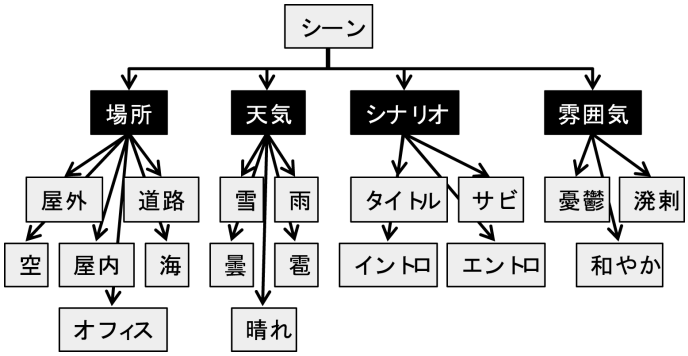

図 3.6: シーンに関するオントロジー集合

があるが，本論文ではそこまでは言及していない．また，別の解決案として，特定の用途に限定した，より小さいオントロジーを作成する．特定の領域に限定したオントロジーを領域オントロジーと呼ぶ．たとえば野球映像に対象を限定したアノテーションを作成する場合，野球に関する領域オントロジーを作成する．

拡張性を考慮し，オントロジーは XML 形式で蓄積する．オブジェクトに関するオントロジー集合は，objectOntology.xml という定義ファイルに，シーンに関するオントロジー集合は sceneOntology.xml という定義ファイルで XML データベースに保存される．具体的な形式は以下のとおりである．それぞれの概念には，その概念と同じ意味の同義語リストも記述することができ，文字の多様性に基づく表記ゆれや多言語の問題に対応している．

```
<?xml version="1.0" encoding="Shift_JIS"?>
<ontology-sets>
  <ontology id="ontology 名">
    <item id="概念名">
      <synonym word="同義語"/>
              ...（synonym が複数存在）
    </item>
     ...（item が複数存在）
  </ontology>
     ...（ontology が複数存在）
```

```
</ontology-sets>
```

リスト1 objectOntology.xml の定義ファイル形式

また，オントロジーや概念は，ユーザによって追加される場合も想定している．これは，あらかじめ用意されたオントロジーでは適切に表現できないオブジェクトやシーンがあるためである．しかしながら，単に新しい概念を追加しただけでは，コンピュータにはその概念の意味を理解できない．今回は動画コンテンツを自然言語レベルで処理することを想定しているため，新しく追加される概念に対し，新しい概念の同義語リストを列挙する形でその概念の意味を定義した．同義語リストを記述することによって，キーワード検索などで一致する可能性が広がる．現段階では複数の意味を手入力で記述しているが，同義語辞典であるシソーラスなどと連携させれば，自動入力も可能である．

オントロジー定義ファイルの記述形式には，このようなXML形式による表記だけでなく，RDF[21]などグラフ構造を用いた定義の表現方法が存在するが，今回はより簡略化するために採用を見送った．

3.3.2 オブジェクトと映像シーンに対するオントロジー型アノテーション

カットやオブジェクトに対してアノテーションを作成するためには，自然言語で無秩序に記述するよりも，あらかじめ登録されている概念を関連付けたほうが，意味内容が一意に決まり都合がよい．また，複数の概念を同時に関連付けることによって，より正確な意味内容を記述できる．たとえば,「若い男が話している」という概念をオブジェクトに対してアノテーションを作成する場合,「若い」「男」「話す」という3つの意味属性を同時に関連付ければよい．

オブジェクトに対するアノテーションの操作画面を図3.7に示す．

3.3.3 映像に対するスキーマ型アノテーション

次に映像全体に対して，アノテーションを付与する仕組みについて議論する．タイトルや著作権情報といった基本的な情報を入力するコンテンツタイトル情報や，アノ

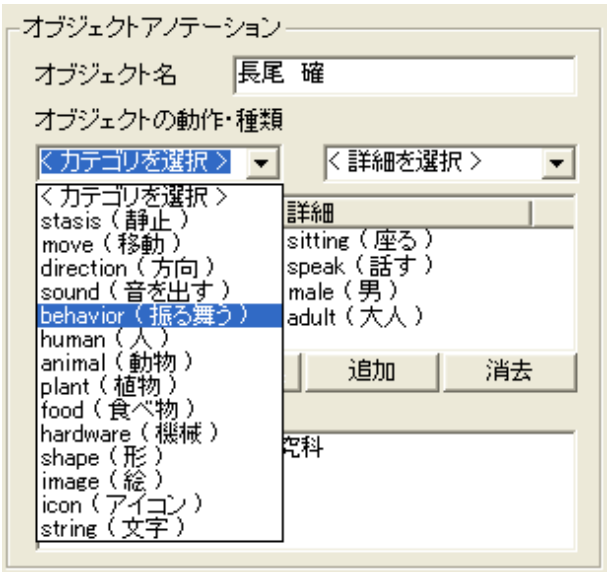

図 3.7: オブジェクトに対するオントロジー型アノテーションの操作画面

テーションを作成した人の情報を表すアノテータ情報を付与する必要がある．これらの情報はあらかじめ用意されたスキーマに基づいてテキストで情報を入力する，スキーマ型アノテーションの形式である．それぞれのスキーマを定義し，そのスキーマに基づく入力フォームをユーザに提示する．コンテンツの書誌情報に対する既存のメタデータスキーマとして Dublin Core[17] が存在するのでこれを参考にした．具体的には，コンテンツタイトル情報のスキーマは以下の通りになる．

```
<?xml version="1.0" encoding="Shift_JIS>
<scheme id="コンテンツタイトル">
  <item name="タイトル" type="String"/>
  <item name="サブタイトル" type="String"/>
  <item name="著作権" type="String"/>
  <item name="作成者" type="String"/>
  <item name="ファイル URL" type="String"/>
  <item name="詳細情報" type="String"/>
</scheme>
```

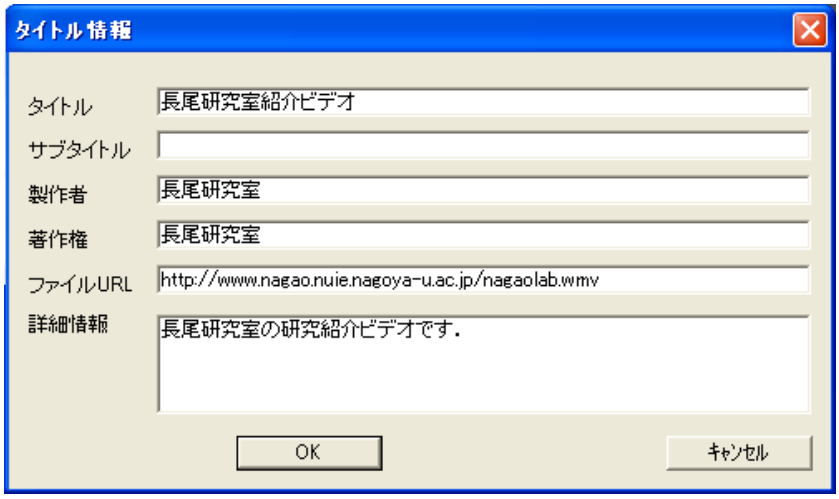

図 3.8: コンテンツタイトルダイアログ

このスキーマを元にして，コンテンツタイトル入力ダイアログは図 3.8 のようになる．同様にして，アノテータ情報入力ダイアログは図 3.9 のようになる．両ダイアログとも Windows の標準インタフェースを用いている．

3.4　アノテーションの管理

次に，アノテーションを管理する仕組みを提案する．具体的には，アノテーション作成結果を XML 形式で出力し，XML データベースに登録することを想定している．具体的なフォーマットの例について議論する．

3.4.1　XML による出力

編集結果を保存する形式として XML を採用した．これは，将来的に MPEG-7 や RDF への対応のしやすさと拡張性の高さを配慮したためである．XML の利点は，拡張性の高さと機種やアプリケーションに依存しない形式であり，また，標準的なテキスト形式で記述されているため，長い時間が経過してもデータとして利用可能な永続性の高さがあげられる．具体的なフォーマット例は以下の形式である．

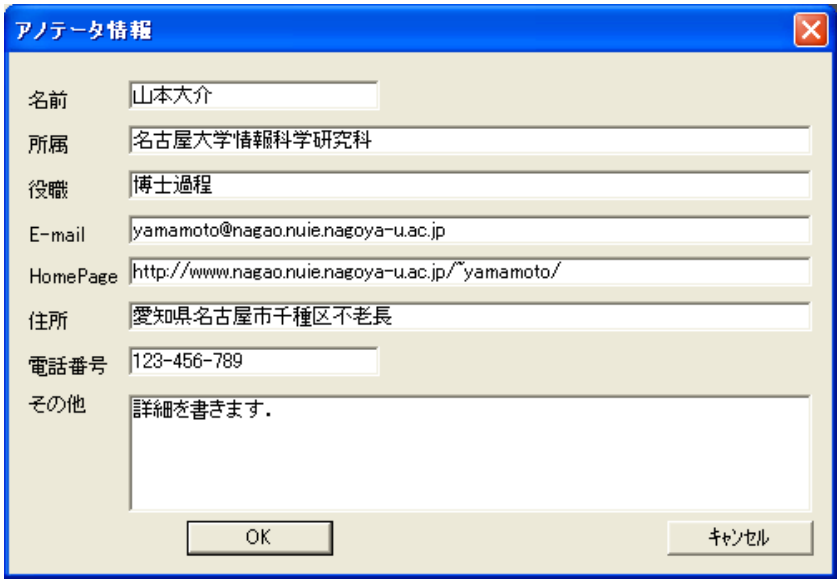

図 3.9: アノテータ情報ダイアログ

```
<?xml version="1.0" encoding="Shift_JIS"?>
<VideoAnnotation version="0.1">
  <VideoHeader>
    <filepath>ローカルファイルのパス</filepath>
    <url>インターネット上コンテンツの URL</url>
    <filesize unit="byte">ファイルサイズ</filesize>
    <winsize width="352" height="240"/>
    <description>詳細情報の記述</description>
  </VideoHeader>
  <VideoTitle>
    <title>ビデオコンテンツのタイトル</title>
    <subtitle>サブタイトル</subtitle>
    <creator>製作者</creator>
    <copyright>著作権表示</copyright>
    <description>詳細情報の記述</description>
  </VideoTitle>
```

```
<Annotator>
  <author>アノテーションをした人</author>
  <email>foo@hoge.com</email>
  <group>author の所属グループ</group>
  <url>annotator の HP</url>
  <address>住所</address>
  <tel>電話番号</tel>
  <post>ポスト</post>
  <description>詳細情報の記述</description>
</Annotator>
<VideoShots>
  <shot stime="開始時間" etime="終了時間" keytime="キーとなるフレームの
時間">
    <histogram r="4" g="4" b="4"> // ヒストグラム情報
      <item r="0" g="0" b="0">0.1112</item>
      ...（item が２の r 乗×２の g 乗×２の b 乗個存在）
    </histogram>
    <place>このシーンの場所</place>
    <annotation>
      <item ontology="オントロジー" detail="概念"/>
        ...（item が複数存在）
    </annotation>
    <description>詳細情報の記述</description>
  </shot>
   ...（shot が０以上複数存在）
</VideoShots>
<Objects>
  <object stime="開始時間" etime="終了時間" keytime="キーの時間"> // object
以下は省略
    <histogram r="4" g="4" b="4"> // ヒストグラム情報
      <item r="0" g="0" b="0">0.1112</item>
```

```
        ... （item が２の r 乗×２の g 乗×２の b 乗個存在）
      </histogram>
      <name>このオブジェクトの名前</name>
      <trackedRect left="左端" right="右端" top="上" bottom="下"/> // オブジェクトの存在矩形範囲
      <annotation>
        <item ontology="オントロジー" detail="概念"/>
          ... （item が複数存在）
      </annotation>
      <description>詳細情報の記述</description>
    </object>
     ... （object が 0 以上複数存在）
  </Objects>
</VideoAnnotation>
```

リスト３　出力 XML のフォーマット

このフォーマットは，主に５つの項目からなる．

1. ヘッダ情報を示すタグの <VideoHeader> タグ．ファイルの存在する位置やファイルサイズ動画フォーマット等を記述．
2. タイトル情報を示す <VideoTitle> タグ．ビデオのタイトルや著作権情報を記述．
3. アノテーションした人の情報をあらわす <Annotator> タグ．アノテーションした人の名前やメールアドレス等を記述．
4. シーン情報をあらわす <VideoShots> タグ．カットごとにカットの開始時刻と終了時刻およびカラーヒストグラムに選択式アノテーション情報の意味属性等を列挙．
5. オブジェクト情報を示す <Objects> タグ．オブジェクトの存在する開始時刻と終了時刻それにオブジェクトの存在矩形範囲，選択式アノテーションや記述情

報を列挙.

VideoTracks タグと Objects タグの中身を説明する．VideoShots タグの中には shot タグが存在する．ひとつの shot タグでビデオコンテンツのひとつのカットからカットまでのまとまりを示す．Objects タグはビデオコンテンツの中に現れるオブジェクトを示す.

3.4.2 データベースへの保存

作成されたアノテーション XML ファイルは，XML データベースである Xindice に蓄積される [2]．アノテーションは他のユーザと共有され，また，アプリケーションから随時利用可能である.

3.5 議論

本システムで作成したビデオアノテーションエディタについての議論する．本システムは主に，自動解析技術に基づく構造アノテーション部分と，意味アノテーション部分から成る．構造アノテーションについて，本来ならば，すべての項目について評価することが好ましいが，我々の研究は信号解析技術を主たる対象としているわけではないので，カット自動検出の精度の評価のみにとどめた．また，意味アノテーションの有効性に関する議論に関しても言及する.

3.5.1 カット検出の評価

カット検出についての評価を行う．評価に利用した映像コンテンツは，56 個のニュース映像から無作為にとりだした 6 つのニュース映像を利用している．対象コンテンツをニュース映像のみに限定しているため，すべての種類の映像コンテンツに対してこの評価が成立するとは必ずしもいえない点には注意が必要である．実験はビデオアノテーションエディタのカット検出機能を利用して行い，人手による修正等は行わない.

具体的な評価は再現率と適合率を用いて行う．本実験における適合率とは，カットとしてシステムが認識した箇所のうち実際のカットである割合を示し，再現率とは,

表 3.1: カット検出に使用した動画ファイル

ファイル名	画面解像度	時間	カット数
WebGrabber.mpg	320x240	1 分 20 秒	14
MapPlanner.mpg	352x240	45 秒	7
NewsPage.mpg	352x240	2 分 40 秒	11
Aspire.mpg	352x240	3 分 04 秒	28
LED.mpg	352x240	3 分 13 秒	58
VirtualWall.mpg	352x240	3 分 34 秒	24

表 3.2: カット検出の実験結果

ファイル名	カット数	適合率	再現率
WebGrabber.mpg	14	100.0%	100.0%
MapPlanner.mpg	7	100.0%	100.0%
NewsPage.mpg	11	91.7%	100.0%
Aspire.mpg	28	96.4%	96.4%
LED.mpg	58	98.3%	95.0%
VirtualWall.mpg	24	95.8%	88.5%
全体	-	97.2%	95.2%

実際のカットのうち，システムによって検出できたカットの割合を示す．実験結果を表 3.2 に示す．全体の適合率は 97.2%，再現率は 95.2%になる．

エラーの例としては，図 3.10 のように同じカットではあるものの，信号の色が赤になるなどの場合は，カットとして検出してしまう例や，図 3.11 のように，画面全体の色変化がおきないでカットが出現する場合などである．このようなエラーの主要因は，画面全体の色ヒストグラムを求めているので局所的な変化に対応していない点である．単純な色ヒストグラムではなく位置情報も考慮した手法を考えなければならないと考えている．このように，すべての映像コンテンツに対して高い精度をもつカット検出のアルゴリズムを考案することは難しい．そのため，人手による修正インタフェースがより重要になる．

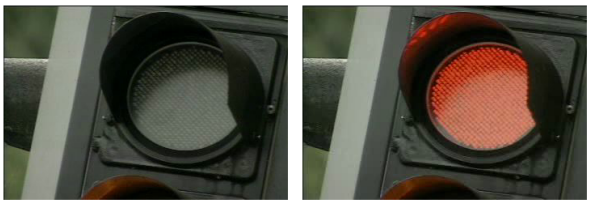

図 3.10: 誤検出例：信号の色が時間とともに変化することにより新たなカットであると誤検出している例

図 3.11: 未検出例：映像が左から右へと徐々に変わっていく例

3.5.2　意味アノテーションについての議論

次に意味アノテーションについての議論する．本来ならば，本ツールの使い勝手を評価する必要があるが，比較対象となるツールが無かったため，今後の課題としたい．次に，本ツールを用いて作成したアノテーションの質についての評価を行う．アノテーションの質に関してはアノテーションを作成する人に大きく依存するため，本ツールの客観的な評価を行うのは難しい．そこで，本ツールによって作成されたアノテーションに基づくビデオシーン検索システムを作成することによってアノテーションの評価とする．

3.5.3　アノテーションに基づく応用

アノテーションに基づく応用例として，自然言語によるビデオシーン検索システムを取り上げる．これは，自動解析技術のみでは十分な精度が達成できない例であり，本ツールによって作成されたアノテーションの有用性を示す良い例である．具体的なシステムとして，ビデオアノテーションエディタによって作成され，XML データベー

スに蓄積されたアノテーションを元に，一般的な Web ブラウザを用いてビデオコンテンツを検索するシステム (図 3.12) を開発した．検索クエリーは，ユーザが入力した自然言語文から成る．

検索アルゴリズム

検索は，本研究で作られたアノテーションを XML データベースに登録したデータを元に処理する．データベースに登録されている情報としては，動画コンテンツに対するタイトル情報と構造アノテーション情報や意味アノテーション情報，カットとオブジェクトに対する色ヒストグラム情報などである．

これらのデータに基づき，以下のアルゴリズムを用いて検索を実現する．

1. 茶筅 [46] を用いて検索クエリーの形態素解析を行い，形態素に分解する．その中で，動詞・形容詞・名詞・未知語を検索キーワードとして抽出する．なお，未知語とは辞書に登録されていない単語であり，経験的に固有名詞や英単語などのキーワードとなりうる単語である場合が多い．

2. 次に，形容詞・名詞から色や明暗を表す単語（たとえば，赤い・黒い・青い・暗い・明るい等）を色を表現する単語として利用した．色を表現する単語（便宜上，色形容詞と呼ぶことにする）は，アノテーションに含まれる色ヒストグラム情報との比較を行うために利用する．

3. 検索アルゴリズムには，意味アノテーション，およびその同義語リストと，検索キーワードとの完全一致により実現する．一致した場合，そのオブジェクトやシーンに対して加点し，得点が高い順に検索結果を並び替えてユーザに提示する．検索キーワードの中に色を表現する単語が含まれている場合は，色ヒストグラムとの相関関数を用いた比較を行い，その相関度に基づき加点する．得点の重み付けは，1 つのショット及びオブジェクトにおいて，キーワードが 1 つヒットしたら 1 点とし，色ヒストグラムに関しては，最大 1 点最小 0 点の実数値で加点した．また，検索クエリー文の文頭のキーワードの重みを小さく，文末のキーワードの重み付けを大きくした．本来ならば，構文解析に基づく係り受け関係などの情報を用いて重み付けを決定する必要があるが今回の実験ではその処理を省略した．

図 3.12: ビデオの検索画面例

4. オブジェクトやシーンを得点に基づきソートしユーザーに提示する．

評価

検索結果に対する評価を行う．本来ならば十分に多くの動画コンテンツに対してアノテーションを行う必要があるが，現段階では試作段階であり，少数のビデオコンテンツに対してのアノテーションを詳細に検討した方がよいと考えたため，３つの映像コンテンツだけにアノテーションを作成し検索を行った．アノテーションはオブジェクト及び映像シーンに対して付与し，また，検索等に耐えられるように十分な量のアノテーションを作成した．

今回，アノテーションを作成した映像コンテンツは図 3.3 に示す３つであり，いずれもニュース映像である．それぞれのコンテンツに対する映像シーン検索のクエリー例に基づいて検索の有効性について議論する．

表 3.3: アノテーションを作成した映像コンテンツ

ファイル名	画面解像度	時間	カット数	オブジェクト数	シーン数
WebGrabber.mpg	320x240	1 分 20 秒	14	12	1
MapPlanner.mpg	352x240	45 秒	7	7	4
NewsPage.mpg	352x240	2 分 40 秒	11	9	6

最初の検索クエリー例として,「ウェブについて話している男」について議論する．この例では，形態素解析に基づき,「ウェブ」・「話す」・「男」のキーワードを抽出した．この中で,「話す」,「男」はそれぞれ該当する概念が存在しそれに応じて検索を行う．それらと一致した概念を持つオブジェクトやシーンの一覧を作成し，順位づけしてユーザに提示した．具体的には,「男」,「話す」の両方の概念を持つオブジェクトやシーンが検索結果のより上位になる．検索結果の例を図 3.13 に示す.

次のクエリー例として,「ウェブについて話している若い野郎」について議論する．この例では，形態素解析に基づき,「ウェブ」・「話す」・「若い」・「野郎」とキーワードを抜き出している.「話す」,「若い」は直接概念と関連付けられ，また,「野郎」という概念は存在しないが,「男」の同義語リストの 1 つとして「野郎」が存在するため,「野郎」という単語は間接的に「男」という概念と対応する．このように，1 つの意味属性に複数の単語が対応する例である．検索結果の例を図 3.14 に示す.

最後のクエリー例として,「暗いシーンのテレビ」について議論する．この例では，形態素解析に基づき,「暗い」・「シーン」・「テレビ」という 3 つのキーワードを抽出した．シーンの前に色や明度の情報をあらわす「暗い」が存在し，映像の色ヒストグラム情報を利用し,「暗い」シーンの検出も行い，その度合いに基づき加点する．検索結果は図 3.15 に示している.

これらの例により，適切なアノテーションが作成されている場合は，適切に映像シーン検索ができることが分かる．しかしながら，検索のために用意したアノテーションを作成することによって映像シーンの検索ができるということは，同義語情報や色情報を利用するなどの工夫はしているものの自明であり，具体的な評価については省略する.

図 3.13: 検索例「ウェブについて話している男」

図 3.14: 検索例「ウェブについて話している若い野郎」

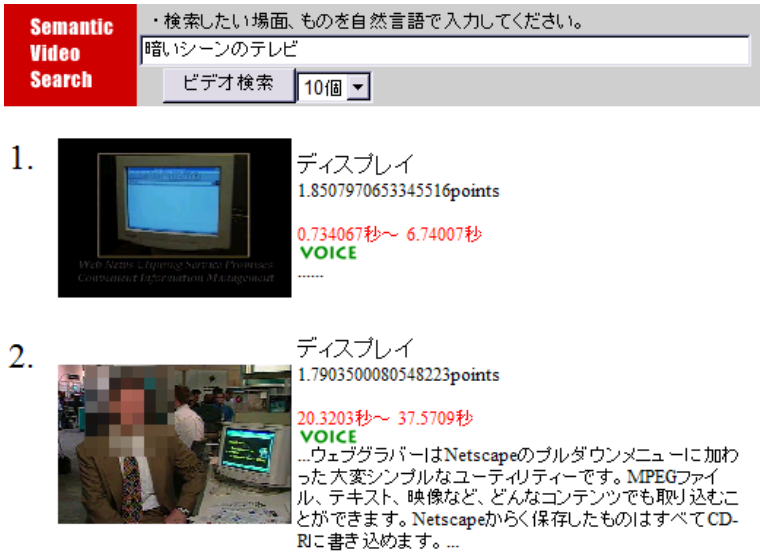

図 3.15: 検索例「暗いシーンのテレビ」

3.6　本章のまとめ

本章では，映像コンテンツに対してオントロジーに基づく半自動アノテーションシステムを提案した．映像に対するアノテーションとしては，構造アノテーションと意味アノテーションの 2 種類が必要である．構造アノテーションの作成には，カット検出技術などの自動解析技術が有効であることを示した．しかしながら，それらには少なからずエラーが存在するが，エラーを人手で容易に修正できる仕組みも提案した．意味アノテーションの作成に関しては，信号処理のみに基づく処理では一般に困難である．そこで，人間が容易にアノテーションを作成できる仕組みを開発した．具体的には，あらかじめ作成したオントロジーと映像の内部構造とを容易に関連づけられる仕組みである．それぞれのオントロジーには，同義語リストが関連付けられており，キーワード検索などにおける再現率の向上が見込める．さらに，これらの仕組みを用いて作成した映像シーン検索システムを試作した．また，映像シーン検索を行う上で十分なアノテーションを作成することが可能であり，十分なアノテーションが作成されているコンテンツに対しては，適切に検索が可能なことを示した．

このような人手と機械処理を組み合わせたアノテーションの方式は応用システムを開発する上で極めて有効である．信号解析技術の向上により，将来的にはよりコスト

を下げることも可能になるであろう．しかしながら，専門家が映像コンテンツに対してアノテーションを作成するのには多大な人的コストがかかる．経験的に，映像コンテンツの長さの 10 倍程度の時間が必要であり，計算機が映像コンテンツの意味を完全に理解することができない限り，大幅なコスト低減は見込めない．また，どこまで詳細にアノテーションを付与すればいいかの見極めも難しい．そのため，費用対効果に見合う映像コンテンツ，たとえば商用コンテンツにしか現実的には適用できないという欠点も持ち合わせている．

近年，YouTube などに代表される動画共有サービスの普及により，膨大な映像コンテンツが Web 上で共有されている．それらの映像コンテンツの一つ一つの価値は必ずしも高くない．そのため，本章で述べたアノテーションの仕組みのみでは費用対効果の問題からこれらのコンテンツに適用できない．この問題を解決するための仕組みとして，次章以降に述べる，Web コミュニティや閲覧者による映像アノテーションの仕組みが有効である．つまり，本章で作成された理想的なアノテーションと同等な情報を，Web コミュニティからいかに獲得するかが問題になる．

次章では，閲覧者が容易に映像アノテーションに参加できる仕組みとして，閲覧者によるオンライン映像アノテーションについて述べる．

第4章　閲覧者によるオンライン映像アノテーション

本章では，映像コンテンツに対するアノテーションのアプローチの 1 つとして，閲覧者がアノテーションを容易に作成可能な仕組みを提案する．従来のアノテーションシステムは，前章のように専門家が専用のツールを用いて映像コンテンツに関するアノテーションを作成することが一般的であった．本章では，一般の閲覧者がアノテーションに参加する仕組みを提案することによって，アノテーション作成作業を分散化しようとする試みである．

一般に，より多くの閲覧者が気軽に映像アノテーションに参加することを考慮した場合，1) 専用のアプリケーションをインストールしてアノテーションへの参加を強いることは困難である，2) アノテーションの作成に高度な知識や煩雑な操作を強いることは困難である，3) 質の低いアノテーションが生成される懸念があるなどの問題点が挙げられる．これらの問題に対するアプローチとして，1) 一般的な Web ブラウザを用いてコンテンツの閲覧とアノテーションの作成が可能なシステムを開発，2) 前章で述べたオントロジーに基づくアノテーションのような高度なアノテーション手法ではなく，ユーザの自由コメント付与によるアノテーション手法の提案，3) 個々のアノテーションに対するアノテーション信頼度の提案を行う．

具体的なシステムとして，オンライン映像アノテーションシステム iVAS を開発した．iVAS とは，Web 上の映像コンテンツに対して，コンテンツを閲覧しつつ映像コンテンツの内部要素に対して電子掲示板感覚でアノテーションを作成するシステムである．さらに，このシステムを用いて，4 つの映像コンテンツに対して 30 人の被験者によるアノテーション作成に関する評価実験を行った．実験結果を元に，アノテーション信頼度に関する評価及びユーザ同士の興味や印象の近さを示す印象距離の評価を行った．これらの実験により，閲覧者によるアノテーションの有効性について議論する．

従来，マルチメディアコンテンツに対して閲覧者がアノテーションを作成するという試みはあまりなく，映像の要素に対して閲覧者がユーザコメントを書き込みアノテーションとして再利用する点が新規であり，また，アノテーション信頼度や印象距離といった指標の提案と評価も本研究の新規な部分である．

4.1　はじめに

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映像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映像コンテンツが氾濫することが予想される．それに伴い，映像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要求は格段に高まっている．

動画コンテンツの意味的な検索や要約を実現するためには，そのコンテンツへのメタ情報の付与（アノテーション）が不可欠である．例えば，MPEG-7 [18] のように動画コンテンツに対するアノテーションの規格が制定されている．また，動画に対する詳細なアノテーションを行う研究は様々なものが行われている [6, 29, 33, 24, 23, 22] が，人的コストが高く，作成に時間がかかる．

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を受け付け，容易に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いう観点から，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）が実現できると思われる．そのために，コンテンツを閲覧しつつ映像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ことができるシステムを提案する．

映像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，本論文では，大きく分けて以下の 3 種類があると考えている．

- 自動映像アノテーション
- 半自動映像アノテーション
- オンライン映像アノテーション

自動映像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[36]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 [56] など様々な研究が行われている．自動映像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．

半自動映像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 [6, 29, 33, 23]．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，意味情報を付与することができるのは大きな利点である．ただし，前章で述べたように，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 [48] 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映像区間の包含関係に基づいて自動継承する OVID モデル [25] 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手法である．

最後に，オンライン映像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．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 [53, 55] や，映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[40] などの研究はこれに当たると考えられるが，アノテーションとしての利用や，応用研究，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，オンライン映像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，閲覧者による映像アノテーションシステムを提案する．

一般に，映像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，各種映像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，映像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，映像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，閲覧者による映像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 (http://www.2ch.net/) 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 [28] では，複数のアノテータによる映像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事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加するごとに，逐次反映され，検索などの応用例の精度向上に結びつけることができるため，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．

4.2　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，閲覧者による映像アノテーションシステムである iVAS(intelligent Video

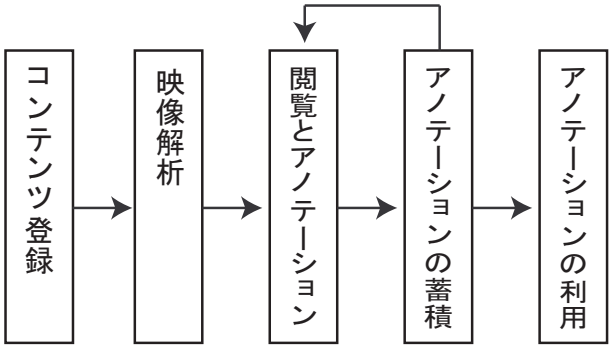

図 4.1: 処理の流れ

Annotation Server) を構築した．

具体的な処理の流れを図 4.1 に示す．まず，ユーザが映像コンテンツを Web インタフェースからアップロードすることによって，コンテンツの登録を行う．次に，投稿された映像コンテンツの解析を行い，具体的には色ヒストグラムの抽出とカット検出，および，それぞれのカットのサムネイル画像の生成を行う．ユーザは，専用の Web インタフェースを用いて映像コンテンツの閲覧とアノテーションの投稿が可能である．投稿されたアノテーションは専用のデータベースに蓄積され，検索などの応用システムに利用される．

4.2.1 システム構成

本システムの構成図を図 4.2 に示す．ユーザは，ネットワークからアクセス可能な任意の映像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を行うことができる仕組みである．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．映像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つつ，閲覧者がアノテーションを投稿する仕組みである．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，6 章で述べる各種のアノテーションを利用したサービスなどで

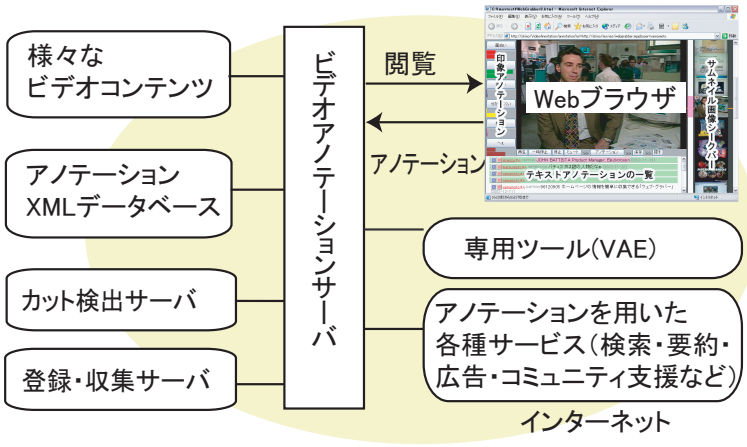

図 4.2: システム構成

利用される．アノテーションシステムとしては，2 章で述べたプロキシモデル [15] に相当する．

4.2.2　想定する映像コンテンツ

アノテーションが可能な映像コンテンツは，PC からアクセスできるデジタル映像コンテンツである．図 4.3 に示すように，インターネット上で公開されている Web 映像コンテンツだけでなく，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．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．

4.2.3　コンテンツ登録

映像コンテンツは，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．

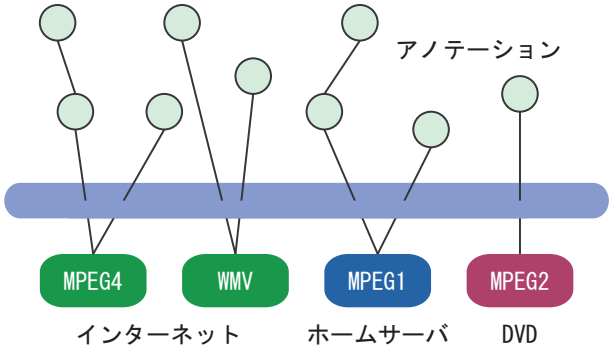

図 4.3: アノテーションを想定する映像コンテンツ

4.2.4 映像の解析

Web ブラウザを用いて映像へのアノテーションを行う場合，インタラクティブに映像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ので，解析を行いたいコンテンツに対してあらかじめ前処理を行っておく必要がある．そのために，投稿時にカット検出を行い，映像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，前章と同様に，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 [41] 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，スポーツの中継や個人で撮影したホームビデオや講演映像などはそもそもカットが多くなく，カットを基準とした映像シーンのセグメントは適切ではない．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．

また，カット検出の過程で得られる色ヒストグラム情報も XML データベースに蓄積する．

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 4.1 に示す．全体の適合率は 89.3 %，再現率は 91.7 %であった．

表 4.1: カット検出の精度．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たかった数をいう．ニュースの精度が悪いのは，花火大会の映像で誤検出が多かったためである．

	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

4.3　閲覧者による映像アノテーションインタフェース

閲覧者は，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（コメントアノテーション）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）及びコメン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 つを行うことができる．

4.3.1　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 4.4 のようになる．画面左に次に述べる印象アノテーション用のインタフェース，中央部上部に動画閲覧画面，画面中央下部にコメン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．なお，カット検出直後のフレーム画像は乱れていることが多いため，カットとして検出したフレームから 20 フレーム後をサムネイルとして選択している．

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映像のシークを直感的に支援する．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．

また，コメン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

図 4.4: アノテーション編集ページ

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．

4.3.2 コメントアノテーション

コメントアノテーションは，映像内の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．対話形式でアノテーションを行うことにより，カットの範囲や対象・種類などの情報を半ば強制的にアノテーションさせる利点がある．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・

表 4.2: 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
関連 URL	テキスト入力	任意
参照	XML による記述	任意
評価	○・×ボタン	任意

場所など），種類（名前・状況説明・補足情報・感想など）を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コメン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 4.5 に示す．結果は XML データベースに格納される．

投稿される内容はアノテーション ID，アノテーションを行ったユーザ名，E-mail などの個人情報，コンテンツ ID，日時などが自動的に記録される．具体的な XML は図 4.6 ようになっている．

なお，この対話的アノテーションは XML によって記述された設定に基づき生成されており，対話内容のカスタマイズが可能となっており，アノテーションを行う対象やアノテーション内容に応じて適切なインタフェースを提供可能となっている．

まとめると表 4.2 のような情報を取得することができる．

4.3.3 印象アノテーション

印象アノテーションとは，映像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．

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，正規分布を用いて印象情報をつけるとすると，各印象 I_k は次の式でパラメータを付与する．

$$I_k(t) = \sum_{t_i \in \text{印象 } I_k \text{に対応したクリックの時間集合}} \frac{1}{2\pi\sigma} \exp(-\frac{1}{2\sigma^2}(t - t_i)^2) \qquad (4.1)$$

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 4.7)．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，また何個必要かなどに関する詳細な検証は，コンテンツの種類なども考慮すべき点などから今後の課題とする．

4.4 アノテーションの解析

アノテーションの解析手法について提案する．

4.4.1 アノテーション信頼度

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．

$$e_k = s \cdot \frac{g_k - a \cdot b_k}{g_k + a \cdot b_k} + t \cdot c_k \tag{4.2}$$

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，

$$a = \frac{g_{all}}{b_{all}} \tag{4.3}$$

と表す．e_k は，$-1 < e_k < 1$ の値をとる．

(4.2) は機械的な評価と人間の評価を組み合わせた式であるが，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，

$$p = d(G+B)\frac{G - a \cdot B}{G + a \cdot B} \tag{4.4}$$

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

$$d(x) = 1 - \exp(-\tau \cdot x) \tag{4.5}$$

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，映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．

$$r_k = (1 - d(g_k + b_k)) \cdot p + d(g_k + b_k) \cdot e_k \tag{4.6}$$

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．

登録ユーザに対してはアノテータ信頼度の計算が可能であるが，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と思われる．

4.4.2 アノテーションに基づく印象距離

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．コンテンツの意味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．しかしながら，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，誰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少なく，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し，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することを目指す．

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．

$$Dc(P_j, P_k) = \int (I_{pj} - I_{nj}) \cdot (I_{pk} - I_{nk}) dt \tag{4.7}$$

ここで，I_{pj} は式 4.1 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．

現在は一つのコンテンツでしか考えていないが，複数のコンテンツ間での関係を表現できればより詳細なコミュニティ形成などに応用できると考えている．

4.5　実験と考察

iVAS システムを利用したアノテーション信頼度と印象アノテーションの評価を行うために以下の実験を行った．

映像処理評価用映像データベース [54] から表 4.3 で示す 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，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．得られたアノテーションは表 4.4 のとおりである．

コメン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，自由に書き込んでもらうグループ D，コメン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，いわゆる「荒らし」的アノテーションのことをいう．だれでも自由に情報を投稿することが可能な Web 上においては，このような荒らしの要素を無視できないためである．

また，この実験を行う前に事前アンケートとして，NHK ONLINE MEMBERS[1]の

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

表 4.3: 評価用映像コンテンツ.

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

表 4.4: 得られたアノテーション情報の数．テキストはコメントアノテーションの書き込み数，印象は印象アノテーションのマウスクリック数，○評価，×評価はそれぞれの評価ボタンを押した数.

	評価人数	テキスト	印象	○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

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.

4.5.1 アノテーション信頼度に関する実験と考察

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，○×評価は一般的な投票行為であり議論を挟む余地は少ない．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 4.4 より求め，図 4.8 に示した．なお今回の実験では，式 4.2 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 4.3 より $a = 4$ とした．また，閲覧者による評価数が 4 個で

式 4.5 の値がおよそ 0.5 になる $\tau = 0.2$ とした．

図 4.8 から分かるように，アノテータ信頼度の値は A グループ $>$ B グループ $>$ C グループ 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ため，アノテータ信頼度を求める意義があると考えられる．

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 個のコメントアノテーションに対する式 4.6 によって求めたアノテーション信頼度の値と，式 4.2 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コメン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言える．

4.5.2　印象距離に基づくコミュニティの視覚化に関する実験

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 4.7 を用いてそれぞれの印象距離を求め，印象距離がより近いユーザほど近接して可視化したものを図 4.9 に示す．一つの円が一つ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．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 4.9 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 4.9 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 4.7 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．

今回は実験していないが，サッカーなどのコミュニティが分かりやすいコンテンツの場合には，応援するチームごとにコミュニティが形成されたり，さらにそのコミュニティの中でも応援する選手ごとにコミュニティが視覚化されると考えられる．

印象ボタンに対する問題点としては，押したいボタンがない場合の対処方法があげられる．ボタンをより多く用意する，ボタンを追加できるようにする，ポジティブな感情の時に押すポジティブボタンとその逆のネガティブボタンの二つだけにするなどが考えられるが，どれがベストなのかは今後の課題としたい．

4.5.3 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．以下にその結果を示す．アンケートは，評価実験をしてもらった後にやってもらい，5 段階評価で集計をした．

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．

4.6 本章のまとめ

本章では，オンライン映像コンテンツに対して，不特定多数の閲覧者による，Web ブラウザを用いた映像アノテーションシステムを提案した．具体的には，映像シーン

表 4.5: 5 段階評価によるアンケート結果．数字が大きいほど良い．なお，コメントアノテーション，印象アノテーションはそれぞれのインタフェースの使いやすさについてのアンケートである．また，コメン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．

	1	2	3	4	5	平均
コメントアノテーション	0	3	7	12	2	3.50
印象アノテーション	0	3	8	11	8	3.80
正確に付与出来たか	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

に対して対話的に属性付きコメントを付与可能なコメントアノテーションと，さまざまな種類の印象を付与できるボタンアノテーションを提案した．これらは 2 章で述べたユーザコメント型アノテーションの仕組みであり，閲覧者にアノテーションを委ねることで人的コストが低い特徴がある．さらに，不特定多数のアノテーションで問題となるアノテーションの信頼度の計算方法を提案した．アノテータ信頼度を導入することによって，より正確なアノテーション信頼度の計算が可能になることを確認した．

前章で述べた専用のアノテーションツールを用いた半自動アノテーションシステムであるビデオアノテーションエディタ (VAE)[48] を用いたアノテーションと本研究で提唱する Web ブラウザによる閲覧者によるアノテーションシステム (iVAS) についての比較を表 4.6 に示す．

アノテーションの手軽さという点では電子掲示板感覚で利用できる iVAS の方が優れている．また，Web サーバアプリケーションであるため，複数人で単一のコンテンツに対してアノテーションを行ったり，アノテーションの共有という点でも，iVAS は優れている．しかしながら，原理的に iVAS は詳細な画像解析などをリアルタイムに行う事ができないので，詳細なアノテーションには VAE の方が優れている．どちらが優れた手法かという問題ではなく両者は補完関係にあり，コストをかけても詳細にアノテーションしたい場合は VAE を，コストをかけずにアノテーションをしたい場合は iVAS などがよいと考えられる．またこの二つを組み合わせ，iVAS などで収集したアノテーション情報を元に VAE を利用して詳細にアノテーションする仕組みを作

表 4.6: iVAS と VAE の比較

	VAE	iVAS
気軽にアノテーションできるか	△	◎
詳細なアノテーションが可能か	○	△
正確なアノテーションが可能か	○	△
人的コスト	×	○
複数人によるアノテーションが可能か	△	○
アノテーションの共有	△	○

るなど両者を統合するのが有効だと考えられる．

しかしながら，いくつか解決しなければならない問題点も残っている．

アノテーションに関するそもそもの問題点として，果たしてユーザが本システムを実際に使ってくれるのかという点があげられる．アンケート結果からは必ずしも本システムを積極的に使いたいという結果にはなっていない．アノテーションを付与するといった作業よりもコミュニケーションとしての iVAS の利用法の方が良いという意見もあった．コメントアノテーションを付与する場合，カテゴリを選ぶことやシーンの選択などの作業がまだ煩雑であった可能性もある．掲示板型コミュニケーションという目的には，それらの作業は特に意味を持たない．また，今回の実験では被験者を集めて行ったが，一般のユーザを対象としたより大規模な実験を行わなければ，より正確なデータの取得は困難である．また，映像を話題としたコミュニケーションには，本章で述べた掲示板型のコミュニケーションだけではなく，映像を話題としたブログを書くことも近年頻繁に行われている．このような知的活動を効果的に支援する仕組みも必要だと思われる．

次章では，本章の内容に基づき，大規模な実証実験に適した形に変更し，さらに引用型アノテーションの仕組みを追加した Synvie と呼ばれるシステムについて述べる．

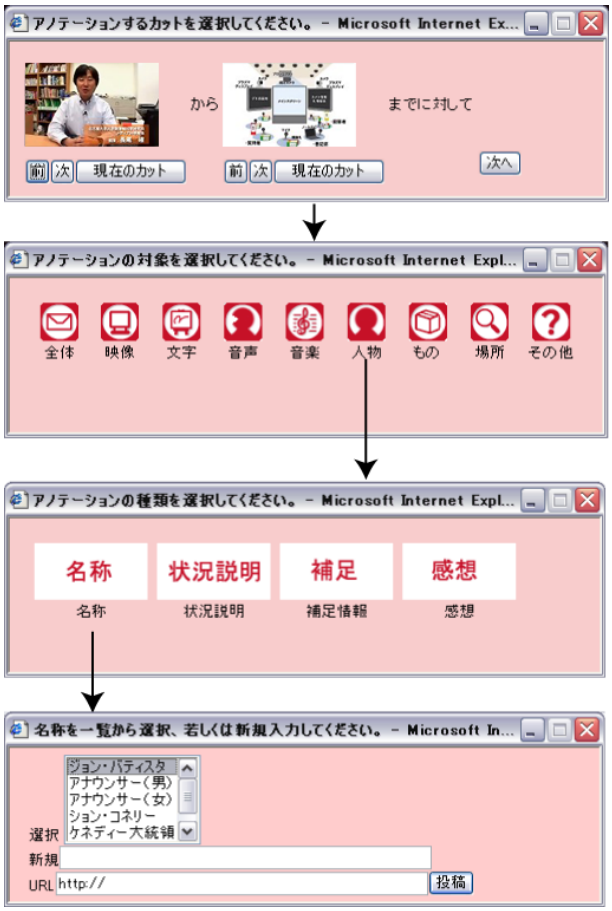

図 4.5: コメントアノテーションの例.「この人物の名称は○○だ」という内容を投稿している.

```
<?xml version="1.0" encoding="Shift_JIS" ?>
<annotation id="an1080830483-2" object="HUMAN" type="NAME" vid="va0000003">
  <timecode scut="7" ecut="8" stime="15.273" etime="35.358" />
  <position x="0.44124" y="0.31245" />
  <annotator handle="山本大介" id="yamamoto@nagao.nuie.nagoya-u.ac.jp" />
  <description>長尾 確</description>
  <url>http://www.hoge.com/foo/</url>
  <date>2004-03-09</date>
  <time>16:46:30</time>
  <evaluation bad="1" good="5" />
</annotation>
```

図 4.6: コメントアノテーションの XML

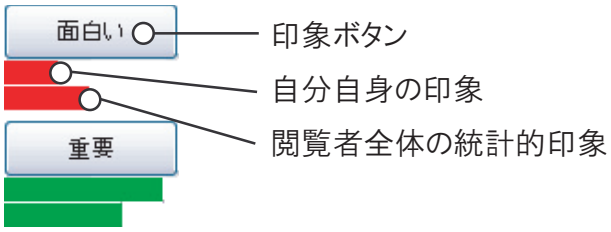

図 4.7: 印象アノテーションのインタフェース．印象ボタンの下に 2 段の棒グラフがあり，ユーザの現在の印象情報の表示と，閲覧者全員の統計的な印象情報を表示する．

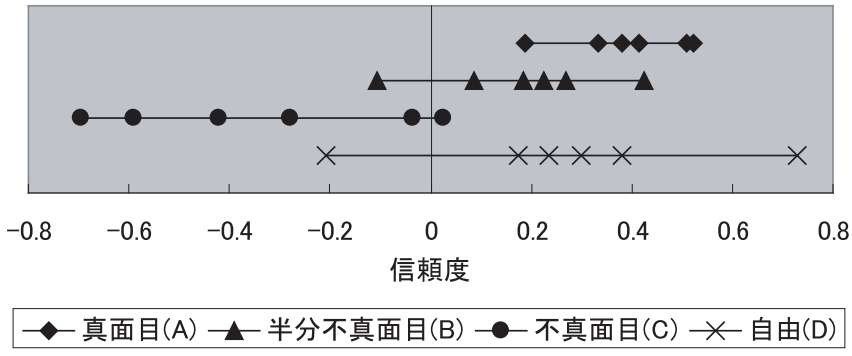

図 4.8: アノテータ信頼度．数値が大きいほど信頼できるアノテータである．

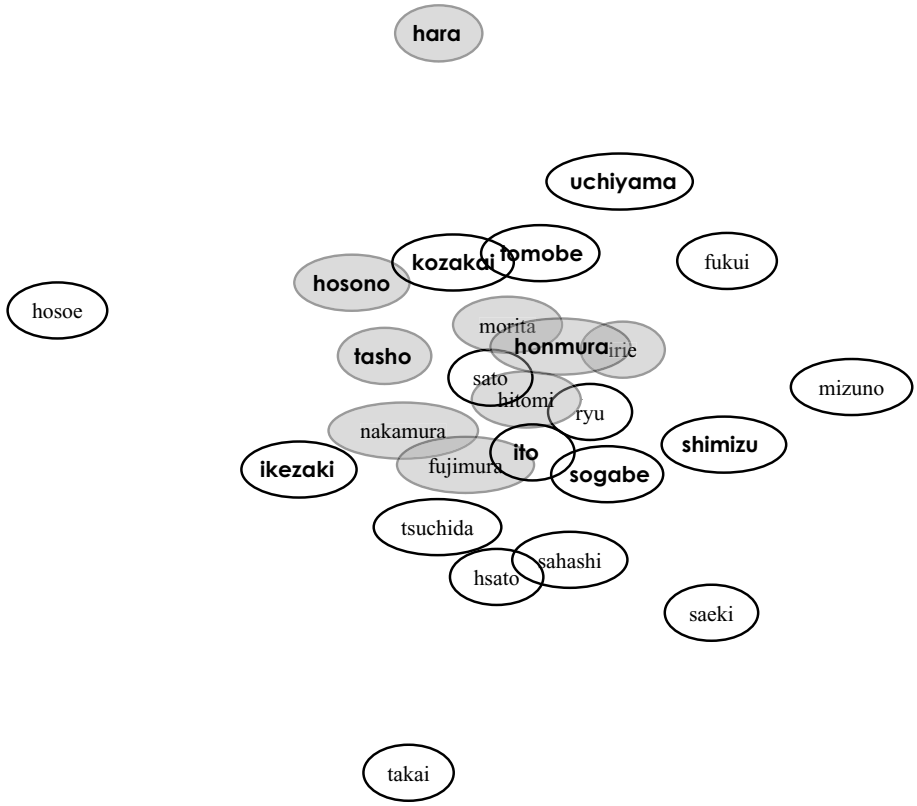

図 4.9: 印象距離．一つの円が一人のアノテータを意味し，距離が近いアノテータ同士ほど興味が近い．背景が灰色の円はアンケートでファッション番組に興味があると答えた人を示し，太字の名前は歴史・紀行に関する番組に興味がある事を示す．

第5章　Webコミュニティ活動に基づく映像アノテーション

本章では，映像を話題としたWebコミュニティから映像に関する意味情報をアノテーションとして獲得する仕組みの提案とその実証実験を行う．対象とするコミュニティは，映像シーンに対してユーザがコメントを付与する掲示板型コミュニティと，映像シーンを引用したブログエントリーを執筆するブログ型コミュニティである．これらのコミュニティ活動を支援するための具体的なシステムを作成し，また，これらのコミュニティ活動によって作成されたユーザコメントと映像の内部要素を関連付けることによってアノテーションの獲得を目指す．これらを実現する具体的なシステムとしてSynvieを開発した．2章で述べたアノテーションモデルとしては，ユーザコメント型アノテーションと引用型アノテーションの仕組みを実装した．

これら2つのモデルに基づく映像アノテーションシステムは存在しない．映像を引用する，映像の部分にコメントを書く，既存のブログシステムと連携するなどといった1つ1つのシステムが手探り状態で開発しなければならず，また，映像と同期して情報を更新するためにはAJAXと呼ばれる新しいプログラミング手法を取り入れなければならないなど実装上の問題も大きい．Weblogとの連携を考慮した映像共有基盤の提案も行う．また，映像や音楽などのマルチメディアコンテンツに対するアノテーションに関する大規模な実証実験は存在せず，本論文のアイデアを実証するためには，実験によって得られたデータに基づき検証する必要性があった．そこで，本システムを用いた実証実験を行った．2006年7月から一般公開し，284名の登録ユーザと，223個の映像コンテンツと，3534個のユーザコメントに基づくアノテーションを含む19182個のアノテーションの取得に成功した．さらに，実証実験によって収集されたデータの解析を行い，アノテーションの特性を定量的に分析した．

前章の閲覧者によるアノテーションとの差分は，実証実験に特化した仕組みにするために，アノテーションの機能を一部絞り込み，映像コミュニティを意識した映像配

信基盤の提案を行った．また，ブログ等への映像シーンの引用の仕組みと具体的なインタフェースの提案，実証実験に基づくデータ収集とその評価なども行っている．具体的には，アノテーションの仕組みとユーザグループの違いによる，アノテーションの品質や量に関する評価を行った．これらの取得されたアノテーションの定量的な分析を行うことによって，我々が提案した引用型アノテーションの優位性を示す．

5.1　はじめに

近年，インターネットの発達とともに，映像・音楽などのマルチメディアコンテンツが Web 上で頻繁に配信・共有されている．専門家が作成したコンテンツだけではなく，一般ユーザが撮影・作成したコンテンツも爆発的に増加しており，それらのコンテンツをいかに効率よく配信・管理・検索するかといった問題が顕在化している．その一方で，ブログや SNS，Wiki などの登場により個人や Web コミュニティからの情報発信が一般化し，影響力も増している．

映像コンテンツの内容検索や要約などの応用を実現するためには，映像シーンに対応するメタ情報（アノテーション）の取得が必要 [23] である．とりわけ，映像シーンの内容に関連したキーワードの抽出や，そのシーンの重要度の推定が有効である．映像シーンに関連したアノテーションの取得に関する従来手法としては，映像認識や音声認識などの自動解析技術を利用する自動アノテーション方式 [36] や，専任の作業者が専用のツールを用いてアノテーションを作成する半自動アノテーション方式 [6, 33] などがある．しかしながら，とりわけ個人が作成したコンテンツの場合，手ぶれ・ピンぼけ・雑音・不明瞭な声などといった撮影者の技能の問題や，カメラ付き携帯電話やデジカメといった撮影機器の性能問題から映像や音声の品質のばらつきが大きく自動解析は限定的にしか利用できない．また，専任の作業者による半自動アノテーションを行うためには，視聴者が限定され，費用対効果の問題から，すべての映像コンテンツに対するアノテーションを施すことは困難である．

そこで，映像コンテンツとブログや SNS などといった Web ユーザコミュニティとを効果的に融合させる仕組みを提案し，それらのコミュニティにおけるユーザの自然な知的活動からコンテンツに関する知識をアノテーション [24] として獲得・蓄積・解析することを目的としている．具体的には，2 つのコミュニケーション手段を提供する．1 つ目は，映像コンテンツの任意のシーンに対して，コンテンツの内容に対する

感想や評価などの情報の関連付けを支援する掲示板型コミュニケーションの仕組みであり，2 つ目は，任意の映像シーンを引用したブログエントリの生成を支援するブログ型コミュニケーションの仕組みである．これらの仕組みを作成することによって，ユーザ同士の映像を題材としたコミュニケーションを支援する．さらには，コンテンツの内容とこれらのコミュニケーションとを詳細に結び付けることによって，コンテンツに付随する様々な情報をアノテーションとして獲得する．このような方式ならば，映像の質やアノテーションコストに左右されず，上述した自動・半自動アノテーションの問題を回避できる．

映像シーンへのアノテーションの仕組み，映像シーン単位でのコンテンツの引用に基づくブログエントリからのアノテーション取得方法の提案，コミュニケーションに特化した具体的なインタフェースの提案，および，それらの仕組みを実装した Synvie[51, 50] というシステムを開発した．さらに，Synvie の公開実験に基づく分析・評価を行い，コミュニケーションから得られるアノテーションを用いたアプリケーション作成のための指針を提示する．

5.2 Web コミュニティのための映像共有基盤

一般に映像コンテンツは意味内容を考慮したうえで柔軟に扱うことは困難である．映像コンテンツを話題とした Web ユーザコミュニティからアノテーションを効率よく取得するためには，機械や人間にとっても扱いやすい枠組みを提供することが望ましい．しかしながら，現状では映像コンテンツを異なるサイト間で横断的に扱うためのプラットホームが存在しない．そこで，HTML コンテンツの管理・配信・機械的処理などで一定の成果をあげているブログの仕組みを参考にして，映像コンテンツの配信とアノテーションの枠組みについて考察する．

我々が開発した Synvie のシステム構成図を図 5.1 に示す．ユーザは任意の映像コンテンツを一般的な Web ブラウザを用いて閲覧することができる．ユーザはこのブラウザインタフェースを用いて映像の任意の部分にコメントを投稿し，アノテーションデータベースに蓄積される．また，本システムは映像の任意のシーンを引用したブログエントリーの執筆を支援するツールを提供し，ブログエントリーと映像コンテンツの関係をアノテーションとして蓄積する．他のユーザがブログエントリーを閲覧すると本システムの映像コンテンツに導かれる仕組みである．収集されたアノテーショ

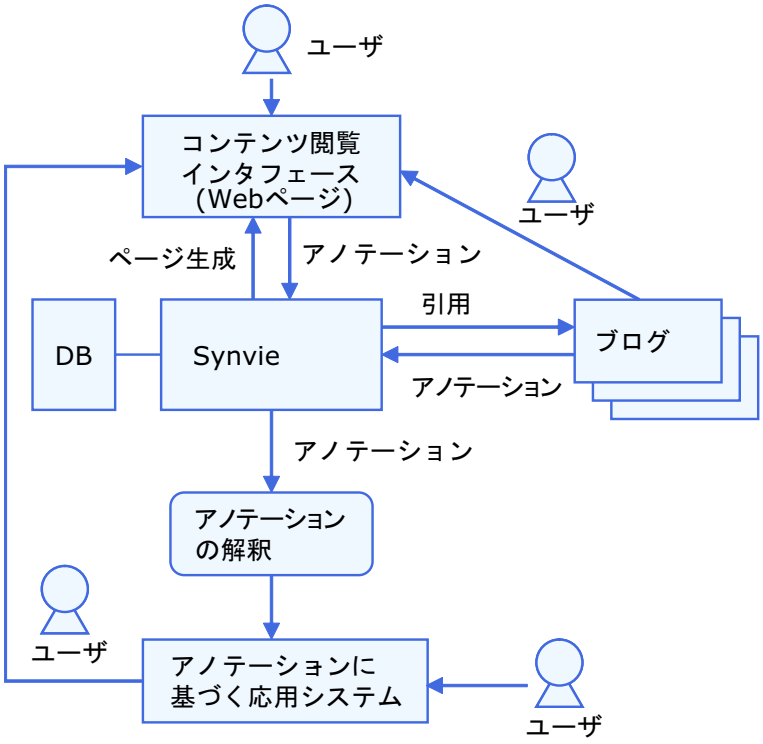

図 5.1: Synvie のシステム構成図

ンは映像シーン検索などのシステムに適用可能である．それらのサービスを利用したユーザもまた，本システムに導かれ，より多くのユーザと多くのアノテーションを獲得できる仕組みである．

5.2.1　ブログに学ぶ

ブログでは，エントリごとに，Permalink[1]，Trackback[4] などの仕組みを実装することによって異なるサイトにまたがるエントリ間のリンクや引用を可能にしている．また，XML Feed[3, 16] の仕組みを利用することによって，コンテンツの情報を機械が理解可能な形で積極的に配信している．さらに，エントリに対してコメント投稿機能を用意することによってユーザからのフィードバックを取得可能である．これらの仕組みを実装することによって，ブログコミュニティは急激な発展をとげることが可能になり，RSS リーダやブログ検索などの様々な応用を生み出してきた．Parker [27]

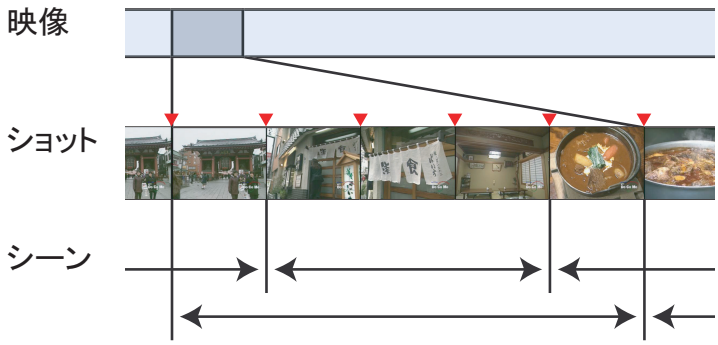

図 5.2: 映像のシーンとショットの定義

によると，ブログの仕組みを映像コンテンツに適用することによって，ビデオブログ検索やビデオブログ配信などといった高度なアプリケーションが実現できると述べている．我々は，さらにこれらの仕組みを映像コンテンツのシーンに対して適用することによって，映像シーン単位でのアノテーションや引用を実現する．これにより，既存のブログと親和性が高い，高度なアプリケーションが実現できるのではないかと考えた．

5.2.2 映像シーンとショットの定義

2.1.1 章でも述べたが，映像シーンとショットの関係について改めて述べる．本研究では，図 5.2 で示すように，映像は複数のショットからなるリストであると定義する．ショットは，一般に映像のカット（切れ目）から次のカットまでの時間範囲を示すが，必ずしもカットが意味的な内容の切れ目であるとは限らないので，長いショットは一定時間間隔に分割してもよいこととする．本システムでは間隔を 2 秒とした．また映像を Web 上でより扱いやすくするために，それぞれのショットの内容を表すサムネイル画像をあらかじめ用意する．シーンとは，複数の連続するショットからなり，意味的につながりを持っているものと定義する．1 つのショットが複数のシーンに属することも許す．

5.2.3　映像シーンに対する Permalink

映像の任意のシーンに対してアノテーションなどの処理を施すためには，それらのシーンに対して固有の Permalink を記述できる必要がある．そこで，本研究では梶ら [45] によって提案されている Element Pointer の仕組みを採用した．Element Pointer は任意のコンテンツの部分要素に対して URI を関連付ける仕組みであり，それぞれのコンテンツの URI が一意であることが保証されている．

映像コンテンツ全体に対する Permalink は以下のように，固有の ID を用いた URI を記述する．

```
http://[server]/[content ID]
```

また，任意のシーンに対する Permalink は，以下のように固有の ID とその時間区間を記述する．複数の時間区間に対する Permalink を記述する場合は，コンマで区切って複数記述する．

```
http://server/[content_id]#epointer(
urn:aps:timeline(begin,end),
urn:aps:timeline(begin,end), ...)
```

これらの仕組みにより，映像の任意の時間区間に対して，固有の Permalink を記述することができる．

5.2.4　ブログエントリーの内部要素に対する Permalink

映像シーンの場合と同様に，Weblog エントリーのパラグラフに対しても Permalink を付与する．われわれのシステムによって生成される Weblog エントリーは，その全てのパラグラフ ($< p >$ タグで囲まれるテキスト) に対して固有の ID が付与されている．そこで，http://server/blog.html に存在する Weblog エントリーの id=”id1” のパラグラフに対しては，

```
http://server/blog.html#epointer(urn:weblog:para(id1))
```

と記述する．この仕組みにより，Weblog エントリーの任意のパラグラフに対して一意の Permalink を記述することができる．

本論文では，映像コンテンツと Weblog エントリーのみに対して Permalink を付与したが，画像や音楽等のコンテンツの要素に対しても Permalink を記述することができる．このような仕組みにより，任意のコンテンツを要素単位で一意に指し示すことができる．

5.2.5 アノテーションの記述形式

一般にアノテーションは，コンテンツの属性情報や構造情報・意味情報など，検索や要約等の応用を目的としたメタ情報である。

本研究では、コンテンツにまつわるコミュニケーションや，コンテンツを引用した Weblog 記事（その作成過程における編集履歴も含む）などの，主にオンライン上での収集可能なアノテーションについて扱う．

本研究で扱うコンテンツは全て Web 上に存在するものとし，アノテーションデータは RDF[21] の枠組みを用いて記述する．

本研究で収集するアノテーションは，アノテーションの対象となるコンテンツ及びその要素を示す Permalink と，アノテーションデータ本体の組からなる．アノテーションデータには，その内容（テキストなど）に加えて，誰が，いつ，どのツールを用いて作成したかという情報も含まれる．ユーザ情報はそのユーザの情報を示す foaf[5] リソースへのリンクを記述することにより，複数のサーバや異なるシステム間でユーザ情報を共有する．日時情報は Dublin Core モジュールを利用している．また，それぞれのアノテーションはそれ自身が Permalink をもつことにより，アノテーションに対するアノテーションの作成が可能になる．

これにより，Web 全体を対象とした，コンテンツ及びその要素とアノテーションから成る，従来のハイパーテキストに比べて，より意味的なネットワークが構築される．

具体的には，以下のような形式となる．

```
<rdf:Description ref:about="[Scene URI]">
  <synvie:text>User Comment.</synie:text>
  <synvie:generator rdf:resource="[Tool URI]"/>
```

```
  <synvie:creator rdf:resource="[Foaf URI]"/>
  <dc:date>[Date]</dc:date>
</rdf:Description>
```

なお，[Tool URI] とは，現時点では，アノテーションを付与するために利用したツールやアプリケーションを一意に特定するための ID として利用している．取得されるアノテーションの種類や品質はツールに依存するため，URI の指し先には，アノテーションの解釈や応用に役立つ，そのツールの特性を記述したツールメタデータが存在する．

5.2.6　引用の記述形式

次に，引用に基づくアノテーションを定義する．これは，二つのリソース間の関係を記述するものである．基本的な記述形式は通常のアノテーションと同一であるが，コンテンツの引用元と引用先の Permalink を記述する必要がある．

具体的には，以下のような形式となる．

```
<rdf:Description rdf:about="[Scene URI]">
  <synvie:reffer rdf:resource="[Paragraph URI]"/>
  <synvie:generator rdf:resource="[Tool URI]"/>
  <synvie:creator rdf:resource="[Foaf URI]"/>
  <dc:date>[Date]</dc:date>
<rdf:Description>
```

引用元のシーン情報は，[Scene URI] に記述する．引用先の情報は，$< synvie : reffer >$ モジュールに記述する．本論文の場合は，引用先は Weblog エントリーの任意のパラグラフに該当するので，そのパラグラフへの URI を記述する．他の情報は通常のアノテーションと同一である．

5.3　映像シーンに対するユーザアノテーション

アノテーションには，従来からあるコンテンツの属性情報や構造情報・意味情報など，検索や要約などの応用を目的とした主次的なアノテーションのほかに，副次的な

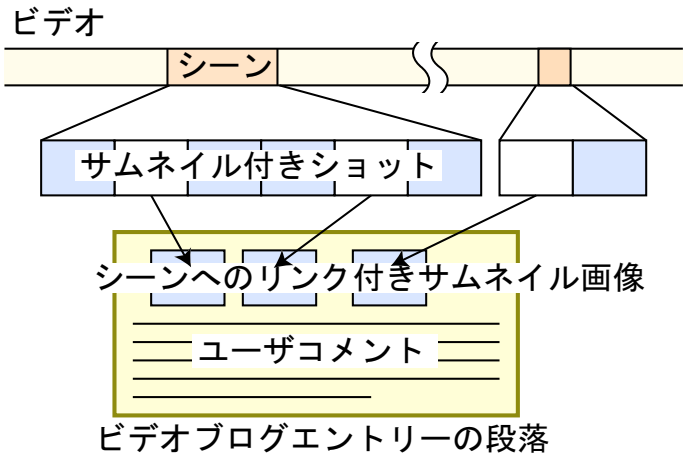

図 5.3: ビデオシーン引用のモデル. ユーザは，サムネイル画像を選択することによって，任意の映像シーンを引用したビデオブログパラグラフを作成可能である．ビデオブログパラグラフは，サムネイル画像・ビデオシーンへのリンク及びユーザコメントからなる．ビデオブログエントリーはこれらのパラグラフの集合である．

アノテーションが存在すると考えている．副次的なアノテーションでは，コンテンツに付随するユーザの自発的なコミュニケーションや，コンテンツを話題としたブログエントリの作成などのコミュニティ活動の副産物としてアノテーションの獲得を目指す．

本システムでは，様々な種類の映像コンテンツに対してアノテーションを付与することを想定している．そのため，コンテンツの種類やユーザの目的によってアノテーションインタフェースを使い分けることが有効であり，いくつかの具体的なインタフェースについて説明する．

5.3.1 映像シーンへのコメントアノテーション

ユーザがコンテンツの任意のシーンに対して容易にコメントの付与などのアノテーションを可能にする仕組みが必要である．そのために，我々が以前の研究で作成したオンラインビデオアノテーションシステム iVAS[49] で述べたテキストアノテーションの仕組みの一部を利用する．

ユーザは，ネットワークからアクセス可能な任意の映像コンテンツに対して，Web ブラウザを用いてアノテーションの投稿および共有を行う．本研究では，シーンに対

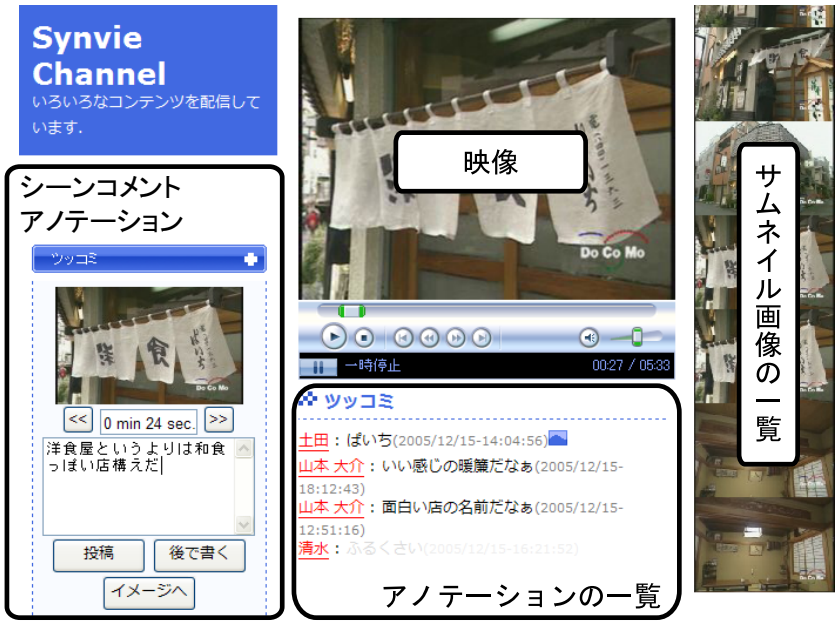

図 5.4: シーンコメントアノテーション．ユーザは現在再生中の映像付近の任意のショットに対してコメントを付与可能である．また，現在の映像に同期したアノテーションを表示可能である．

してコメントを記述することをシーンコメントアノテーションと呼ぶ．図 5.4 に示すように，映像の現在再生中のショットに対してコメントを付与できる簡便なインタフェースであり，映像の閲覧を継続したままアノテーションを付与可能である．

これにより，ユーザは映像コンテンツに対して，電子掲示板感覚で他のユーザとコミュニケーションを図ることが可能になると同時に，関連情報を提示したい，感動を共有したいなどという欲求を満たすことが可能になる．想定するアノテーションの内容としては，映像シーンに関連した有用情報やＵＲＬ，感想などで，比較的短いコメントである．アノテーションとして再利用が高い文章が記述されることは想定しておらず，このアノテーションをきっかけとした，次章で述べるシーン引用に基づくブログ執筆を促すことを考えている．

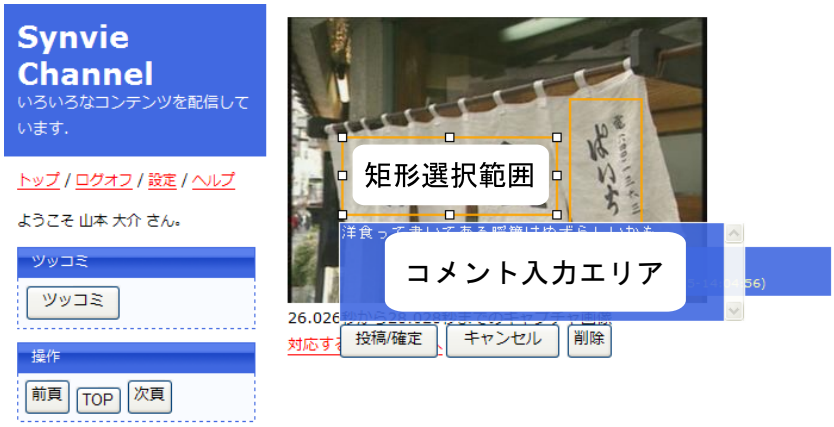

図 5.5: シーン領域コメントアノテーション

5.3.2 映像シーン領域へのコメントアノテーション

シーン領域コメントアノテーションとは，図 5.5 のように，任意の映像シーンの任意の矩形範囲に対してコメントを付与するためのインタフェースである．対象となるシーンの静止画像に対して，マウスで矩形範囲を選択した後にコメントを付与する．これにより，映像の任意のショットの矩形領域を対象としたアノテーションの付与が可能になる．このインタフェースは，映像の閲覧を一時的に停止する代わりに，より詳細で対象が明確なアノテーションを付与可能である．

これは，映像の特定領域に対してのみコメントを記述したいときに有用なインタフェースである．想定されるアノテーションの内容としては，映像上の登場人物やオブジェクトの名称の記述，テロップの書き下し，見落としがちな部分についての注釈などが考えられる．シーンコメントアノテーションよりは説明的な記述が想定される．

5.3.3 映像シーンへのボタンアノテーション

次に，映像シーンに対するより簡便なアノテーションとして，２種類のボタン押下によるアノテーションを提案する．１つは，映像に対するマーキングとしての機能であり，任意のシーンに対して “チェック” を行う仕組みである．これは，次章で述べる映像シーンの引用の手がかりとして用いられ，他のユーザとの共有は行わない．２つ目は，iVAS[49] において提案されたシーンボタンアノテーションである．シーンボタ

図 5.6: シーンボタンアノテーション

ンアノテーションでは，映像の任意の時間に対してマウスを用いてあらかじめ用意された閲覧者の主観的な印象を表すボタンを押すことによって統計的に評価する仕組みである．本システムでは，nice と boo の 2 種類のボタンを用意した．インタフェースを図 5.6 に示す．

本アノテーションでは，ユーザにとって興味深いシーンに対してより多くのボタンが押下されることを期待している．具体的には，面白いシーンや有用なシーン，映像的表現が面白いシーンや批判が集中しやすいシーンに対して多くのボタンが押下されると考えている．

5.3.4　コンテンツへのアノテーション

映像シーンに対するアノテーションだけでなく，YouTube などの従来の動画共有サイトで一般的に行われている，コンテンツ全体に対するコメント投稿の機能も実装した．これによって取得されるアノテーションを，コンテンツコメントアノテーションと呼ぶ．また，タイトル情報などあらかじめコンテンツに埋め込まれているメタデータも，コンテンツの内容を示す重要な情報でありアノテーションとして扱う．

想定されるアノテーションとしては，コンテンツ全体に対するコメントや感想・評価などである．

5.4 映像シーンの引用とそれに基づくアノテーション

一般的にユーザがコンテンツを閲覧し，そのコンテンツが有益で面白いと感じた場合，自身のブログ上でそのコンテンツへの URL を付与した紹介記事の執筆を行うことがしばしば見受けられる．これは，金銭的な見返りを期待しないユーザの自然で自発的な行動である．これらの記事の中にはコンテンツの内容について詳細に記述している記事も存在する．それらの記事の内容と映像コンテンツとを詳細に関連付けることができれば，コンテンツの要素に対するアノテーションとしてとらえることが可能になる．Synvie では特に個人のブログエントリへの引用を支援する仕組みを提供し，その仕組みを利用したユーザの詳細な編集履歴を蓄積することによって，ブログエントリの文章構造と映像のシーン構造とを関連付けたアノテーションの抽出を可能にする仕組みを提案した．映像コンテンツを引用したブログエントリの集合を本論文ではビデオブログと呼ぶ．想定される利用方法としては，ビデオコンテンツの紹介を目的とした記事の記述があげられる．

ビデオの任意のシーンの内容を表すサムネイル画像，そのシーンへのリンクおよびそのシーンに対応するユーザコメントからなる段落をシーン引用パラグラフと呼び，ビデオブログエントリは１つ以上のシーン引用パラグラフから構成される．シーン引用パラグラフの書式を統一することで，アノテーションの解析を行いやすくする意図がある．

5.4.1 引用シーンの選択

ユーザはコンテンツを閲覧する際，自身にとって興味のあるシーンに対してシーンコメントアノテーションやシーンボタンアノテーションなどの何らかのアノテーションを施す．しかしながら，それらのアノテーションは会話的なコメントである，コメント情報が含まれていないなど，必ずしもアノテーションとして優れているとはいえない．そこで，システムはこれらのアノテーションを施したシーンをビデオブログエントリの執筆のための引用シーン候補としてユーザに提示し，ユーザにこれらの候補をもとにしたビデオブログエントリの執筆を促す．これにより，シーンアノテーションの投稿履歴から，段階的にユーザへより説明的な記述が期待できるブログ執筆を促し，より再利用性の高いアノテーションの取得を目指す仕組みである．

図 5.7: 連続シーン引用アノテーションインタフェース

5.4.2　ビデオブログエントリの編集

ユーザが，ブログなどで通常のエントリを書くのと同様に，一般的な Web ブラウザを用いてビデオブログエントリの編集が可能になる仕組みを提案する．

本研究では，２つの編集インタフェースを提案する．１つ目は，連続する映像シーンを引用するのに適した編集インタフェース (図 5.7) である．これは，引用シーンをショット単位で時間的に展開させることで引用シーンの時間範囲をともなう修正・変更が可能であり，より正確にシーンを選択することが可能なインタフェースである．具体的には，シーン伸縮ボタンを押して引用シーンを時間的に前後に伸縮させることによって，正確に引用シーンを提示・選択可能であり，対応するコメントの編集も可能である．これは，シーンの流れやストーリを対象としたビデオブログエントリを記述するのに適したインタフェースであると同時に，より詳細なアノテーションを施すためのツールでもある．連続する映像シーンとブログエントリ上の対応するパラグラフ上のコメントとを関連付けることを連続シーン引用アノテーションと呼ぶ．

２つ目は，複数の非連続な映像シーンを引用するのに適した編集インタフェース (図 5.8) である．過去にユーザが施したシーンコメントアノテーションやシーンボタンア

図 5.8: 非連続シーン引用アノテーションインタフェース

ノテーションに対応するショットが右側のストックに保持されており，その中から任意のショットをドラッグアンドドロップ形式で複数選択し，その複数のショットに対してコメントを付与することが可能なインタフェースである．これは，複数の連続しないショットに対してコメントを記述することに適したインタフェースであり，シーンやストーリよりも特定のオブジェクト（たとえば特定の人物など）を対象としたビデオブログエントリを記述するのに適したインタフェースである．また，映像シーン検索機能と併用することで，他のコンテンツのシーンの引用も可能である．これによって取得されるアノテーションを非連続シーン引用アノテーションと呼ぶ．

ユーザはこの 2 つのインタフェースを使い分けながらビデオブログエントリを作成可能である．

ビデオブログエントリは HTML 文書として表現され，任意のブログサイトに投稿可能であると同時に，アノテーションデータベースに蓄積される．

5.4.3　複数映像コンテンツの映像シーンの引用

次に，複数の映像コンテンツの任意のシーンを直接引用するメカニズムを提案する．そのためには，我々はユーザが任意の映像コンテンツの任意のシーンに容易にアクセスでき，これらのシーンをブログエントリーなどへ引用できるインタフェースを提案する必要がある．そこで，我々は図 5.9 に示す新しいインタフェースを導入して解決した．このインタフェースは，サムネイルシークバーエリアとブログ編集エリアから成る．ブログ編集エリアは前節で述べた非連続シーン引用インタフェースと同様であり，ユーザはサムネイル画像をドラッグアンドドロップすることによって編集することが可能である．

サムネイルシークバーエリアでは，ユーザが引用したい映像コンテンツに対応する複数のサムネイルシークバーが層状に表示されている．サムネイルシークバーは，左上に表示される小さなビデオウィンドウと，水平方向及び垂直方向に伸びる，現在のメディア時間を基準とした連続シーンに対応するサムネイル画像から成る．水平シークバーでは，ユーザは左右にドラッグすることによって映像をシークし，下方向にドラッグすることによってその映像シーンを引用することができる．垂直シークバーでは，より広い間隔でサムネイル画像が表示され，ユーザはそれを上下にドラッグすることによって，おおまかにビデオをシークすることが可能になる．ストリーミングビデオはタイムラグ無しにシークすることは出来ないので，ユーザはいくつかのシーンを見落としてしまうかもしれないが，全ての映像シーンをサムネイル画像によって表示しているのでユーザはこの仕組みを使うことによってこれらの問題を回避可能である．

5.5　アノテーションの解析

本システムでは，コメントアノテーションやシーン引用アノテーションを，なるべく情報劣化がない形式で蓄積する．そのため，本研究で意味するところのアノテーションはユーザコメントの列挙にすぎず，それ自身が機械によって理解可能な情報とは限らない．つまり，本研究によって取得されたアノテーションを用いたアプリケーションを構築するためには，アノテーションを解析し，機械が理解可能な情報に変換する必要がある．そこで，本章では 3 つの視点からアノテーションを解析する手法を提案する．1 つは，アノテーションのテキスト情報からコンテンツの意味内容を表す情報

図 5.9: 複数コンテンツの任意のシーンの引用インタフェース

の抽出を行う仕組みであり，具体的には，映像コンテンツ全体およびシーンの内容を表現するキーワード（一般にタグと呼ばれる）の抽出を目指す．2 つ目は，アノテーションや映像シーンの各々の重要性の計算手法の提案であり，3 つ目は，各々のアノテーション間やシーン間の関連性についての考察である．これらは，アノテーションに基づく応用を実現するために重要な情報である．

5.5.1 タグの抽出

アノテーションとして付与されたテキストからコンテンツやシーンの内容を表現するキーワードの抽出を行う．コンテンツと対応付けられたキーワードをタグと呼ぶ．特に，コンテンツ全体の内容を表現するタグをコンテンツタグといい，シーンの内容を表現するタグをシーンタグと呼ぶ．コンテンツタグ・シーンタグともに以下の手法

によって抽出する．まず，それぞれの自由コメントを形態素解析器茶筌 [46] を用いて形態素に分割する．それぞれの形態素から，名詞・動詞・形容詞・形容動詞・未知語を抽出する．ただし，代名詞や非自立名詞・非自立動詞は除外し，未知語は固有名詞として扱った．さらに一般的に不要語と判断可能な形態素（たとえば，する，ある，なる，できる，いる，など）も除外した．それぞれの形態素の基本形をタグとする．

5.5.2　アノテーションとシーンの重み

アノテーションやシーンの重みの計算手法を議論する．ここでいうアノテーションの重みとは，そのアノテーションが対象となる映像シーンの内容をどれだけ的確に，かつ，信頼性が高く表現しているかを示す指標であり，シーンの重みとは，そのシーンがその映像の中でどれだけ重要なシーンであるかを示す指標である．本来，重要なシーンとは状況や嗜好・目的に応じて変化する [45] ものである．しかしながら，PageRank [26] のように状況や目的を考慮しない重み付けによる検索システムであっても一定の成果をあげており，本論文では PageRank の概念，つまり，より参照されるシーンほど重要であるという指標に基づいて重要度を算出する．

具体的には，アノテーションの重みは，アノテーション対象の粒度（つまり映像シーンの長さ），アノテータの信頼性，アノテーションタイプの信頼性から推定する．つまり，信頼できる人がより正確にアノテーションを作成できるツールを用いて，より粒度の細かい対象（コンテンツよりもシーン，長いシーンよりも短いシーン）に対するアノテーションを付与した場合に，より大きい重みを与える．本来ならばアノテーションの意味内容を加味したアノテーションの重み付けをすることが望ましいが，本論文では意味内容を考慮したテキスト解析は一般に困難であるため見送っている．

また，映像シーンの重みは，より多くの，よりアノテーションの重みが大きいアノテーションから参照されているシーンほど重要であると仮定し，それぞれのシーンを参照するアノテーションの重みの合計がその映像シーンの重みであるとする．

具体的なアルゴリズムの提案と妥当性の検証は，十分なデータが不足している，コンテンツの種類やコミュニティに依存しやすいため検証が困難などの理由から，今後の課題とし，本章ではアノテーションとシーンの重みの計算手法のコンセプトのみを提示する．

5.5.3 アノテーション構造の活用

映像シーンに対するコメントアノテーションは，図 5.10 のように，対応する映像シーンとコメントとを「シーンコメントアノテーション」というラベルの付いたグラフで表現される．コメントは映像シーンに関する情報を含んでいる場合が多く，映像シーンに対するアノテーションとして利用可能である．その一方，ビデオブログエントリは，図 5.11 のように，引用した映像シーンとブログエントリのパラグラフとを「シーン引用」というラベルの付いたグラフで表現され，他のシーンやコンテンツ，ブログエントリとの何らかの関連性の抽出が期待できる．

具体的には，連続シーン引用アノテーションによって選択された連続するショットからなる引用シーンでは，それに対応するコメント内容という観点に基づきシーンの連続性があると見なすことができる．また，非連続シーン引用アノテーションを用いて選択されたショットの集合は，対応するコメントの意味内容という観点に基づいて，シーンの関連性があると考えられる．さらに，1 つのビデオブログエントリで複数のコンテンツを同時に引用した場合，そのビデオブログエントリの内容に基づいて，これらのコンテンツの意味的な関連性があるととらえることが可能になる．複数のコンテンツを引用したビデオブログエントリの例としては，CG アニメーション「ノラネコピッピ 1 話」とその元になった実写映像である「ノラネコピッピのモデルになった猫♪」を同時に引用し比較する記事などである．

本システムにより，Web と映像コンテンツの垣根を越えた引用に基づく詳細なネットワークを形成する．これによりブログネットワークと映像コンテンツを統合することが可能になる．ブログと映像コンテンツの統合されたネットワークでは，コンテンツを扱う粒度がコンテンツ/エントリ単位から映像シーン/パラグラフ単位へとより詳細になり，コンテンツに関連するコミュニティが共有サイト内から Web 全体に拡大されている．さらに，コンテンツ間のリンクをナビゲーションのための 1 方向的な Hyperlink から引用に基づく意味的な双方向リンクへと拡張させることができる．これにより，我々の提案する仕組みはコンテンツに付随する様々な知識を抽出するためのフレームワークとして機能し，それによって収集されるデータは検索やコンテンツ推薦などの様々な応用のための基礎的データとして利用されることが期待できる．

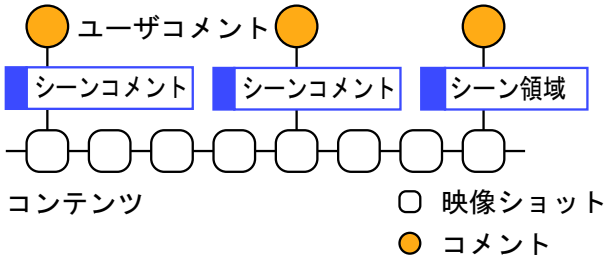

図 5.10: 映像シーンへのアノテーションのモデル

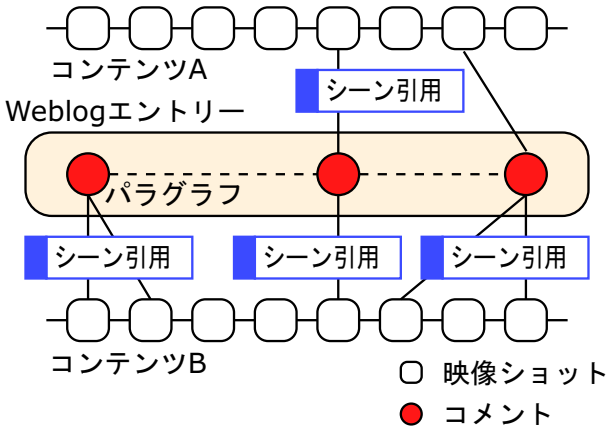

図 5.11: 映像シーン引用に基づくアノテーションのモデル

5.6 実証実験と考察

我々が提案したコミュニケーションを目的としたアノテーションから，検索などの応用に有用な情報がどれくらい取得可能であるかを検証するために，本論文で提案したSynvie[1]の公開実験を行った．2006年7月1日から公開を開始し，2007年11月30日までに，284名の登録ユーザと，223個の映像コンテンツと，3534個のユーザコメントに基づくアノテーションを含む19182個のアノテーションの取得に成功した．そのうち，2006年10月22日までの期間において収集されたデータに基づき評価を行う．この期間に，登録ユーザ数97人，投稿コンテンツ94個，1コンテンツあたりの平均メディア時間は321.5秒，総閲覧数は7,318回に達した．収集されたアノテーションは，表5.1に示すように計4,768個であった．

現在投稿されているコンテンツは，教育・研究に関するものが12件，アート・アニメーションに関するものが16件，ペット動物に関するものが8件，旅行・観光に関するものが19件，エンターテイメントに関するものが15件，時事ネタが6件，グルメ関係が2件，乗り物関係が9件，スポーツ関係が7件，コメディ関係が1件，科学・テクノロジ関係が2件，講演映像関係が2件，音楽関係が2件，人間関係が1件，その他が4件となっている．全てのコンテンツは著作権上問題がない，素人映像が中心であり，必ずしも面白い映像コンテンツばかりではないが，クレイアニメや3DCGアニメーション，ミニドラマなどといった手の込んだ作品も少なからずある．

コンテンツコメントアノテーションがYouTubeなどの従来システムで実用化されているアノテーション，シーンコメントアノテーションがiVASなどの従来システムによって取得されるアノテーションととらえ，本論文ではこれらに加えてシーン引用アノテーションを提案している．これらのアノテーションタイプの違いによるアノテーションの質と量を比較することによって，シーン引用アノテーションの有用性を示す．

5.6.1 タグに基づく分析

タグの評価を行うために，あらかじめすべてのタグ候補に関して，そのタグが対応するコンテンツやシーンの内容を直接表現しているかどうかに基づき，筆者がタグの分類を手作業で行った結果を表5.2に示す．有効であると判断されたタグを有効タグ

[1] http://video.nagao.nuie.nagoya-u.ac.jp/

表 5.1: 公開実験によって取得されたアノテーション

対象単位	行為	型	アノテーションタイプ	取得数
コンテンツ	投稿	文	コンテンツコメント	40
シーン	投稿	ボタン	シーンボタン	3412
		文	シーンコメント	795
			シーン領域コメント	187
	引用		連続シーン引用	283
			非連続シーン引用	51

表 5.2: アノテーションタイプごとの有効タグ率と有効タグ精度

アノテーション	形態素数 (平均)	有効タグ数 (平均)	有効タグ精度
シーンコメント	7.19	1.51	58.8%
シーン領域コメント	7.77	2.17	60.9%
連続シーン引用	25.8	5.96	60.0%
非連続シーン引用	23.0	4.74	53.4%
コンテンツコメント	15.3	0.85	11.1%

と呼び，1つのアノテーションに含まれる有効なタグで重複のないタグの数を平均有効タグ数という．シーンコメントアノテーションとシーン引用アノテーションはどちらも映像シーンに対するアノテーションであるが，前者の平均有効タグ数は1.51であるのに対して後者は5.96と3倍以上多い．どちらも1つの映像シーンを話題としたコメントであるため，シーン引用アノテーションの方が，より詳細な話題について記述していることが推定される．シーンコメントアノテーションやコンテンツコメントアノテーションなどよりも，ブログ上で記述されるアノテーションの方がより多くのタグが含まれている傾向がある．次に，機械的に抽出されたシーンタグのうち，どれくらいのタグがそのシーンの内容を的確に表現しているかどうかを示す割合として有効タグ精度という割合で評価する．これは，1つのアノテーションに機械的に除去できない，ノイズとなるタグがどれだけ含まれていないかを表す．すべての形態素をタグとした場合の有効タグ精度は平均20%前後であるが，前述したタグの絞り込み手法を用いると表5.2で示すように60%前後まで向上する．この数値は決して高いとはいえない．しかしながら，有効タグは対応するシーンやコンテンツに直接関連しているかどうかという基準で選別したために，有効タグ率には映像から派生した話題に関連しているタグは反映されていない．無作為に記述されたアノテーションでない限りは，そのコンテンツを閲覧して記述したという観点から何らかの関連性はあり，派生的に関連したタグも含めれば有効タグ率はこれよりも大きくなる可能性が高い．アノテーションタイプ別の傾向として，シーンコメントアノテーションやシーン引用アノテーションなど，対象単位が映像シーンとなるものの有効タグ率が高い．これらは，映像シーンというより粒度の細かい対象について議論しているため，コメントの内容が映像シーンの内容に影響を受けやすいためであるからと考えられる．

タグの分類を手作業で行うのには多大のコストがかかることが懸念される．我々は２種類の方式でこの問題を解決することを考えている．1つは，構文解析や意図解析などといったより高度な言語処理技術を用いる手法である．人的コストがかからないという利点がある一方，Synvieで取得されるような自由コメントに対してこれらの問題を適用することは非常に困難である．２つ目は，増田ら[52]が提案した，ユーザらによって協調的にタグを選別する手法であり，費用対効果の観点から有用性が確認されている．

5.6.2　アノテーションの主観的分類

収集されたアノテーションを評価するために，それぞれのアノテーションのコメント内容に対して，以下のとおり，アノテーションの意味に基づく分類を行った．

A 主にシーンの内容を説明・解説するコメント．

B 主にシーンに対する直接的な感想や意見などからなるコメントで，シーンに関連するキーワードが含まれるもの．

C 主に，シーンの内容から派生した話題に関するコメント．

D 感嘆符のみ，形容詞のみなど単独では内容を理解できないもの．あるいは，撮影手法，映像の品質に対する感想など，シーンの内容とは関係のない話題からなるコメント．

さらに A，B，C のカテゴリに関して，コメントの文章としての正しさに基づき，

X コメントに主語・述語・目的語が存在するなど，十分に内容を表現している．

Y 十分に内容を表現しているとはいえない．

のサブカテゴリに分類した．なお，分類は 2 人の評価者によって同時に行い，異なる意見が出た場合には話し合いによる調整を行った．

$A-X$ のアノテーションの例としては，朝顔の展示に対して映像撮影者が自身のブログで「名古屋式盆養切込みづくりの朝顔です． 蔓を伸ばさずに盆栽仕立てにしていてとてもユニークです．100 年の歴史があるそうです.」と記述したコメントのように，シーンの内容を的確に表現しており言語解析などを行うことによって，より多くの知識の抽出が期待できる．$A-Y$ のアノテーションの例としては，Web アプリケーションのデモ映像で画像のアップロードを行っているシーンに対する「画像のアップロード」というコメントのように，シーンの内容を表現しているキーワードを含んでいるが，十分に内容を表現しきれていないものである．$B-X$ は「私にとっての朝顔は，こういう蔓を上へ上へと伸ばしていくタイプです.」のようにシーンに対しての感想や意見を述べているものであり，$B-Y$ は「どれだけお菓子使うんだよ！笑」のような表現であり，ともにシーンの内容に関するキーワードの抽出が期待できる．$C-X$ の

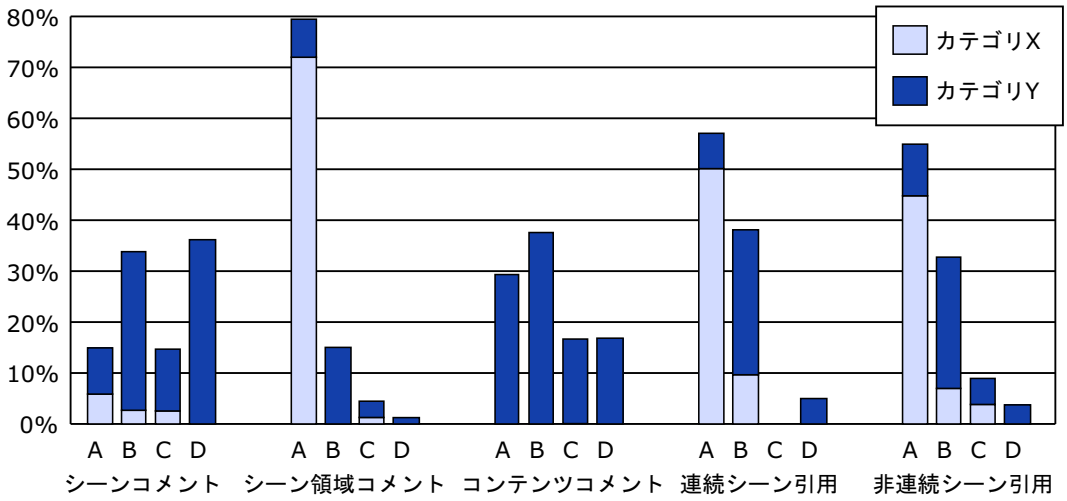

図 5.12: アノテーションタイプごとのアノテーションの質の比較

例としては映像中に表示される URL のキャプションに対して「リサイクルトナー専門店のようです． 著作権フリーの CG, 音楽を製作されているみたいです.」．$C-Y$ の例としては「長尾先生といえば，アノテーションの研究」などである．これらは，関連する話題について記述しており，必ずしも映像シーンの内容を直接的に表現していないが，シーンに関する補助的な情報としての利用が期待できる．D の例としては「すごっ！」や「キター」など，単独では意味をなさないコメントや,「なんでこの回だけ映像がぶれてるのでしょう？ウィンドウズメディアエンコーダーという無料ソフトで，ノンインターレス化できるので是非.」など映像の品質に関する話題などが含まれる．

アノテーションタイプごとにカテゴリ分けし，集計したものを図 5.12 に示す．

カテゴリ $A-X$ に該当する，コンテンツの内容を説明・解説するためのブログエントリは，コンテンツ投稿者自身によって執筆される事例が多く含まれた．これは，自分の投稿したコンテンツを広くいろいろな人に見てもらいたいがためであると推察される．

5.6.3 考察

まず，アノテーションの量の観点から考察する．ここで，アノテーションの量は，そのアノテーションを付与する手軽さや扱いやすさに関係していると仮定する．表 5.1

で示すように，従来型のコンテンツ全体に対するコンテンツコメントアノテーションよりもシーンコメントアノテーションの方が投稿数が多いため，シーンコメントアノテーションはより手軽なアノテーションであったと推察できる．一見すると，コンテンツ全体に対するアノテーションの方が，シーンを選択する手間がない分手軽であるように感じられるが，シーンに対するアノテーションの方が，注目対象を限定しているため，他の閲覧者と話題を共有しやすく比較的短いコメントで内容を記述できる，些細な問題や話題でもコメントを投稿しやすいなどの理由から，より手軽に投稿可能であるためだと推察できる．

次に，アノテーションの質の観点から考察する．厳密な質の定義は応用に依存するが，ここでは，コメント内容の品質が高くシーンの内容を的確に表現し，引用シーンに関連するキーワードなどを含んでいるものとする．具体的には，$A > B > C > D$ の順で質が高く，また，サブカテゴリ X の方が Y よりも質が高いものとする．ただし，カテゴリ C に属するアノテーションは直接的にシーンに関係しているとはいえないコメントであっても，シーンから派生した情報であるため無関係とはいえない．むしろ，MPEG-7 などの通常のアノテーションからでは得ることが困難な重要な情報が隠れている可能性があり，決して無視することはできないと考えている．

これらの観点からみると，図 5.12 で示すように，コメントアノテーションに比べて，シーン引用アノテーションの質の方が高い．特に，シーンコメントアノテーションにおいてサブカテゴリ X に属する割合は 11%なのに対し，シーン引用アノテーションは 59%になり，より正確な文章が記述されていること，また，シーンコメントアノテーションにおいてカテゴリ D に属する割合が 36%も存在しているのに対して，シーン引用アノテーションの場合は 4.8%であるなど，無関係なコメントや “荒し” と呼ばれるコメントが少ないなどの点で，シーン引用に基づくアノテーションの方がより質が高い傾向があるといえる．

つまり，アノテーションの質や量はアノテーションタイプに依存する．これは，閲覧者が映像を見ているという前提が成り立ち，その場限りのコミュニケーションを目的としたシーンコメントアノテーションよりも，映像コンテンツを閲覧しているとは限らない不特定多数に向けたブログエントリの執筆を目的としたシーン引用アノテーションの方がより丁寧な文章を記述する傾向があり，より質の高い情報を記述していると捉えることができる．また，掲示板よりもブログの方が一般的により良い文章が書かれている現状を反映した結果ともいえる．一見面倒で操作が多いアノテーション

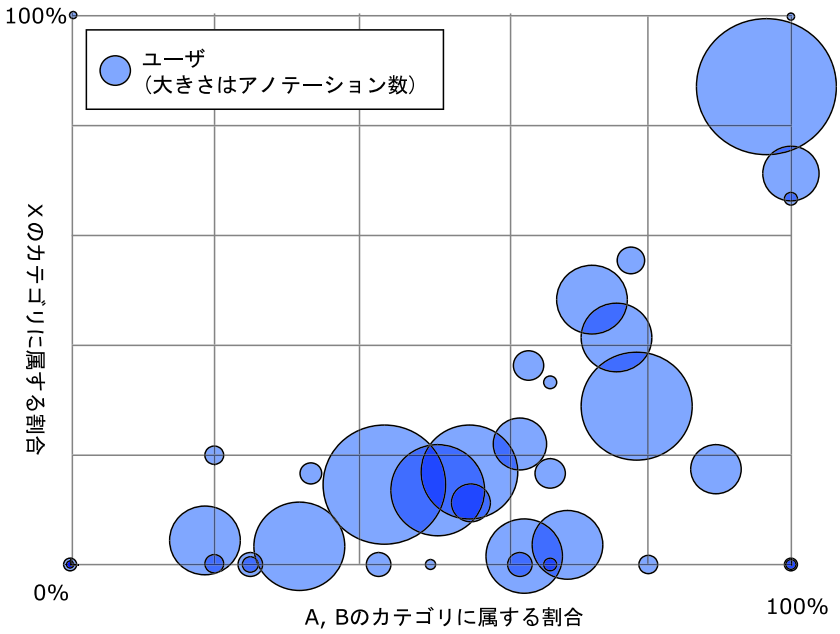

図 5.13: アノテーションを施した数および品質に基づくユーザの分布．1 つの円が一人のアノテータにあたり，円の大きさが投稿したアノテーションの数にあたる．右上に行くほど質の高いアノテーションを施したユーザである．

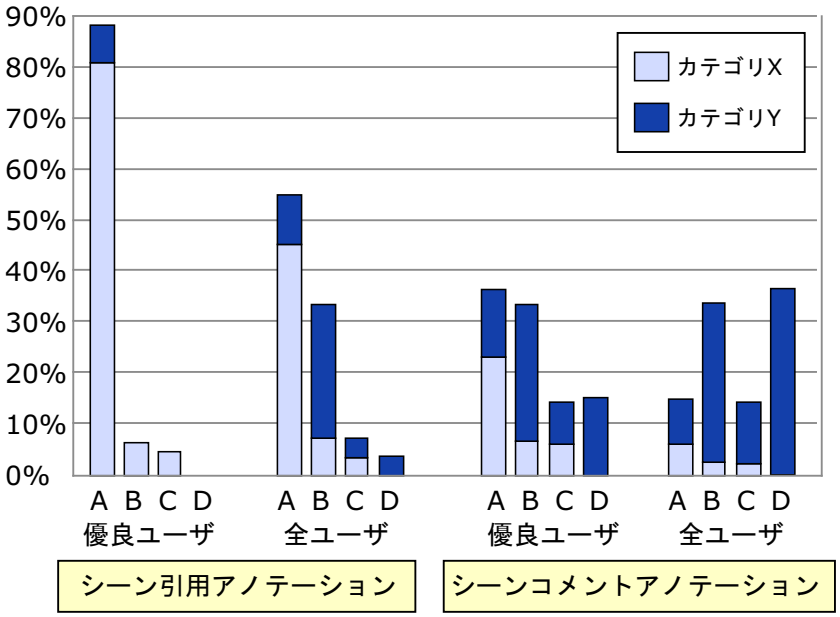

図 5.14: アノテーションタイプおよびユーザごとのアノテーションの質の比較．優良ユーザとは，サブカテゴリ X に属するアノテーションを施した割合が多い人，上位 30%を示す．

も，ブログを書くなどといった人間の自然な日常活動の一部として取り込むことができれば，十分な質と量をともなうアノテーションの取得が可能になることが分かる．

次に，アノテーションと人との関連性を考察する．図 5.13 に示すように，良いアノテーションを施す人もいれば，そうでない人もいる．つまり，人に応じてアノテーションの質や量は異なり，ばらつきがある．そこで，サブカテゴリ X のアノテーションを付与した数の割合が多い上位 30%のユーザ，つまり良いアノテーションを投稿する割合が多い人を優良ユーザと定義する．図 5.14 に示すように，優良ユーザがシーン引用アノテーション方式を用いて施したアノテーションのうち 80%が一番質の高いカテゴリである $A - X$ に分類される．これは，シーン引用アノテーション全体の平均の 45%や，質の高いアノテーションを施した上位 30%の人がシーンコメントアノテーションを用いて付与した平均 22%よりも圧倒的に多い．

つまり，アノテーションの量と質は人にもアノテーションタイプにも依存する．逆にいえば，人やアノテーションタイプが，アノテーションの質の推定パラメータの 1

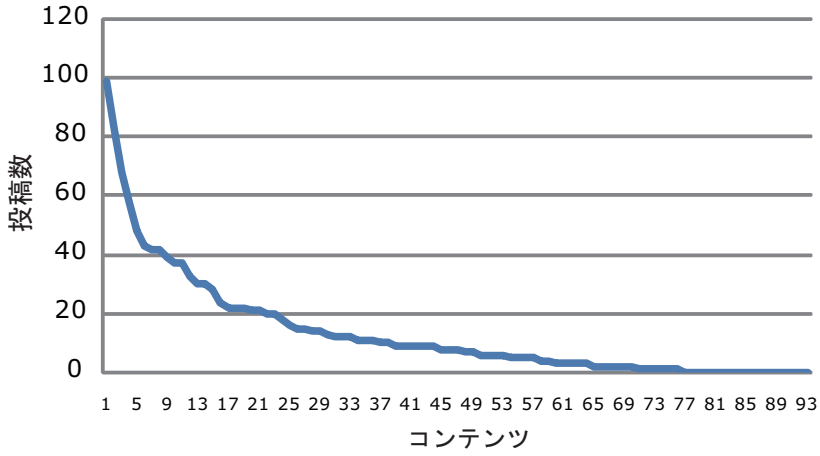

図 5.15: アノテーションの投稿数とコンテンツの関係

つとして利用することが可能になる．具体的なアルゴリズムは現状ではデータが十分揃っていないため今後の課題としたいが，学習アルゴリズムを用いてパラメータを動的に決定することを検討している．

次にコンテンツの内容とアノテーションの量や質の関係について述べる．縦軸をアノテーション数，横軸をコンテンツとしてアノテーション数順に並べると図 5.15 のようになり，また，94 のコンテンツのうち，上位 12 個でおよそ半分のアノテーションの投稿数を占めている．この結果から面白いコンテンツほどより多くのアノテーションが投稿されることが分かる．次に，アノテーションの多いコンテンツとそれ以外のコンテンツでのアノテーションの質について議論する．アノテーションの多い上位 12 件のコンテンツを人気ありとし，それ以外を人気なしとした場合の，それぞれのアノテーションタイプ別の割合を図 5.16 に示す．ブログ引用の場合は人気ありのコンテンツの方がより質の高いアノテーションの傾向が認められるが，有意な差とまではいえない．そのため，アノテーションの質はアノテーションの投稿数に依存しているとは認められない．

まとめると，YouTube などで実用化されているコンテンツコメントアノテーションよりも，我々が iVAS で提案してきたシーンコメントアノテーションの方がより多くのアノテーションを収集することが可能であり，また，シーンコメントアノテーションよりも，本論文で提案したシーン引用アノテーションの方がより質の高いアノテー

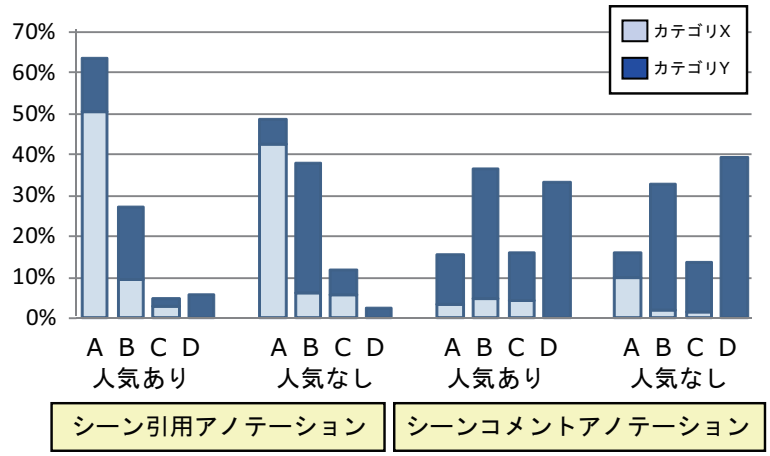

図 5.16: アノテーションの投稿数とアノテーションの質の関係

ションの収集が期待できる．シーンコメント / シーン引用アノテーションを併用することによって，バリエーションに富んだ質・量とも高いアノテーションの収集が可能になる．これにより，多くのアノテーションが集まった場合は，評価が高いユーザが施したシーン引用アノテーションを重視し，あまり集まらなかった場合は，シーンコメントアノテーションの情報も活用するなど，場合によって使い分けることが可能になる．

5.7　本章のまとめ

本章では，映像シーンへのアノテーション，映像シーン単位でのコンテンツの引用に基づくブログエントリからのアノテーション取得方法の提案，コミュニケーションに特化した具体的なインタフェースの提案と公開実験に基づく評価を行った．これにより，それぞれのアノテーションタイプによって得られるアノテーションの傾向をアノテーションの量と質の観点から分析を行い，それぞれのアノテーションに特有の傾向が見られることが分かった．特に，関連するブログエントリから情報を抽出することが質の高いアノテーションを抽出する手助けになることが示せたことが有用であると考えている．これは，シーンコメントアノテーションが掲示板文化を引き継いでいるのに対して，シーン引用アノテーションはブログ文化を引き継いでいることを反映

していると考えられる．また，これらのアノテーションは，2 つの観点により映像の構造・意味情報も抽出可能である．1 つは，コンテンツを引用することによってそれぞれのショット間の意味的な関係の抽出が期待できる．もう 1 つは，引用によって複数のコンテンツ間の意味的な関係の抽出が期待できる．

今後の課題として，タグ選別の自動化に関する問題や，アノテーションに基づく他のアプリケーションの開発が挙げられる．アプリケーションの例としては，ビデオ推薦システムやビデオスキミングシステムを想定している．我々が提案するビデオ推薦システムとは，映像と同期して関連性のある他のコンテンツのサムネイル画像とキーワード，およびその根拠となるビデオブログエントリを表示し，関連するビデオの推薦を行うシステムである．本システムでは，複数コンテンツを同時引用したビデオブログエントリの内容に基づく，統計情報に頼らない詳細なコンテンツ推薦を実現している．ビデオスキミング [43] とは，映像の重要なシーンのみを通常の速さで再生し，それ以外のシーンを早送りで再生する仕組みであり，映像の内容を短時間で把握するのに適している．具体的には，映像シーンの重要度を基にして，映像シーンの選別を行うことを考えている．

次章では，本章の仕組みによって獲得されたアノテーションに基づく応用システムについて述べる．

第6章　映像アノテーションに基づく応用

映像アノテーションに基づく応用について議論する．MPEG-7 関連の技術に代表される専門家によるアノテーションは，いかに良質なアノテーションを低コストで作成するかということが問題の本質であり，応用を強く意識したアノテーション方式であるため比較的容易に応用システムの開発が可能である．実際，我々も 3 章において映像シーン検索システムを容易に開発することが可能であった．しかしながら，Synvie に代表される Web コミュニティ活動によるアノテーションにおいては，ユーザは応用を意識せずにアノテーションを作成するため，関連性の少ないアノテーションや誤ったアノテーションを取得する可能性が高く，アノテーションの質のばらつきの問題が発生する．そのため，応用システムを作成するのには何らかの工夫が必要になるし，そもそもそれらのアノテーションが応用システムにとって有用であるのかという問題が残る．

そこで，Synvie によって獲得されたアノテーションの有用性を確認と応用システムの提案のために，我々はアノテーションに基づくいくつかのシステムを開発した．具体的には，５章の実験によって取得されたアノテーションに基づく応用の例として，２種類のビデオシーン検索システム，ビデオ推薦システム，ビデオスキミングシステムを提案する．従来のシステムには，信号処理技術に基づくものや半自動アノテーション技術に基づくもの等が存在するが，信号処理技術の精度やコスト等で問題がある．本章で提案する応用は，ユーザから収集したアノテーションテキストや，アノテーションとシーン間の関連性や重みに基づいて実現しており，信号処理技術の精度や人的コストの問題が発生しないという利点がある．

6.1　アノテーションに基づく映像シーン検索

映像シーン検索とは，テキストキーワードを用いて映像をコンテンツ単位ではなく部分要素であるシーン単位で検索しようとする仕組みである．本手法の特徴は，アノ

テーションから抽出されたタグを検索することによって，間接的にビデオシーンを検索しようとしている点があげられる．

具体的な検索プロセスは以下のとおりである．ユーザは，目的のシーンを検索するために，1 つないし複数の検索クエリをタグ形式で入力する．それらのタグと一致するタグを含むアノテーションの検索を行い，一致したアノテーションをコンテンツごとに列挙する．一致したアノテーションに対応するシーンを検索結果候補とする．1 つのコンテンツ内に多くのアノテーションが一致した場合は，検索結果候補が膨大かつ時間軸上で細切れになる危険がある．そこで，対象となるシーンが連続する，あるいは時間的に近い場合は類似するシーンである可能性が高いと考え検索結果候補を統合する．逆に，一致したアノテーションの数が少なく，また，分散しており，検索結果のシーンを特定できない場合はコンテンツ全体を検索結果候補とする．検索結果候補内に属するアノテーションの重みの合計が，その検索結果候補の重みとする．このような仕組みにより，アノテーションが多数存在する場合にも，アノテーションが少量しか存在しない場合にも，ある程度対応可能になる．検索結果候補の重みに基づき，検索結果候補のランク付けを行う．

検索結果候補の内容を理解するために，シーンの内容を表現するサムネイル画像を提示することは有効である．サムネイル画像は，検索キーワードに合致したアノテーションに関連付けられているシーンに属するサムネイルを候補とする．ただし，サムネイル画像が一定個数以上存在する場合には，そのサムネイル画像が属する映像シーンの重みに基づいて絞る．ビデオシーン検索システムのインタフェースを図 6.1 に示す．

検索が成功する例としては，検索したいシーンに的確なキーワードを含むアノテーションが存在する場合である．逆に，検索が失敗する例としては，検索したいシーンに的確なキーワードが含まれないなど，アノテーションの量が不足している場合が考えられる．しかしながら，人気のあるシーンやコンテンツには，より多くのアノテーションが集まりやすく，また，人気のあるシーンほど検索ニーズが高い，このようなシーンやコンテンツには自然にアノテーションが増えていくことが考えられる．すなわち，ある程度の時間が経過すれば，この問題は解決される可能性が高い．また，同じ内容を異なるタグで表現している場合にも検索に失敗する．その場合は，シソーラスを用いて類義語や語彙の上位概念・下位概念の関係を考慮する必要がある．それでも，自動生成されたタグと検索のキーワードが必ずしも一致しないという欠点がある．映像コンテンツに関連付けられているタグを効果的に提示することによって，ユーザ

はあらかじめ用意されたタグを選択できる仕組みが必要になる．

図 6.1: ビデオシーン検索システム

6.2 タグクラウドの仕組みを用いた映像シーン検索

次に，タグクラウドの仕組みを用いた映像シーン検索の仕組みを提案する．まず，アノテーションからシーンの意味内容に対して有用なキーワード（シーンタグ）の抽出を試みる．抽出されたタグを用いて，タグクラウドに似た方式の新しいビデオシーン検索システムを開発した．自動抽出されたタグにはシーンの意味内容と無関係なタグが含まれている可能性があるので Web ブラウザを用いた専用のツールを用いて手作業でタグの分別を行った．その作業コストと検索効率の改善を実験によって比較することにより，アノテーションに対して手作業で後処理を加えることにより，専任の作業者が 0 の状態からアノテーションを関連付ける方式よりもコストパフォーマンスが高いことを示す．

6.2.1　システムの概要

我々はシーンタグの特性を最大限に利用したタグ集合に基づくビデオ検索システムを開発した．アノテーションから生成されるシーンタグには必ずしもそのタグの内容と対応する映像シーンの内容が一致しているとは限らないという根本的な問題がある．また，全てのビデオに十分なアノテーションが付与されていない場合，前に述べた文字列マッチングのみに基づくビデオシーン検索の仕組みが上手く機能しない．

これらの問題を解決するために，タグクラウドの仕組みをビデオシーン検索に応用した．図 6.2 に示すように，ユーザは全てのビデオコンテンツに含まれるシーンタグから生成されたタグクラウドの一覧から，検索したいシーンに関連したタグを選択する．これらのタグが含まれるシークバー付きのビデオ一覧が表示される．それぞれのシークバーには，その時間軸に対応するシーンタグとサムネイル画像が関連付けられており，ユーザはそれらのタグとサムネイル画像を閲覧することによって，ビデオを実際に閲覧することなくシーンの内容を把握できる．具体的なインタフェースは図 6.3 に示すように，ユーザがシークバーをドラッグすると，その時間に対応するサムネイル画像とタグの一覧が同期して表示される．さらに，ユーザが興味を持ったタグをクリックすると，それらのシーンタグの時間的分布がシークバー上に表示される．これらの仕組みによってユーザが目的のシーンを発見することを支援する．

この仕組みはビデオに直接アクセスする必要がないため軽量である．また，シーンに関連するタグやその分布，サムネイル画像からユーザが総合的に検索したいシーンを探索することが可能になるので，タグのイグザクトマッチに拠らない，曖昧さを許容するインタラクティブな検索が可能であるという利点もある．

6.2.2　タグの選別

本ビデオシーン検索システムはシーンタグの量と質に依存する．しかしながら，映像と関連性のないアノテーションが含まれていることは容易に想像できる．それゆえ，自動生成されるシーンタグには有効ではないタグが含まれているかもしれない．実際，我々はシーンと明らかに関連性のないタグを発見した．これらのタグを自動的に選別することは一般に困難であるため，実際の応用システムを開発するためには効率よくタグの選別を行うシステムを開発する必要がある．そこで，本研究では Web ユーザが適切なシーンタグを容易に選別できるタグ選択システムを開発した．

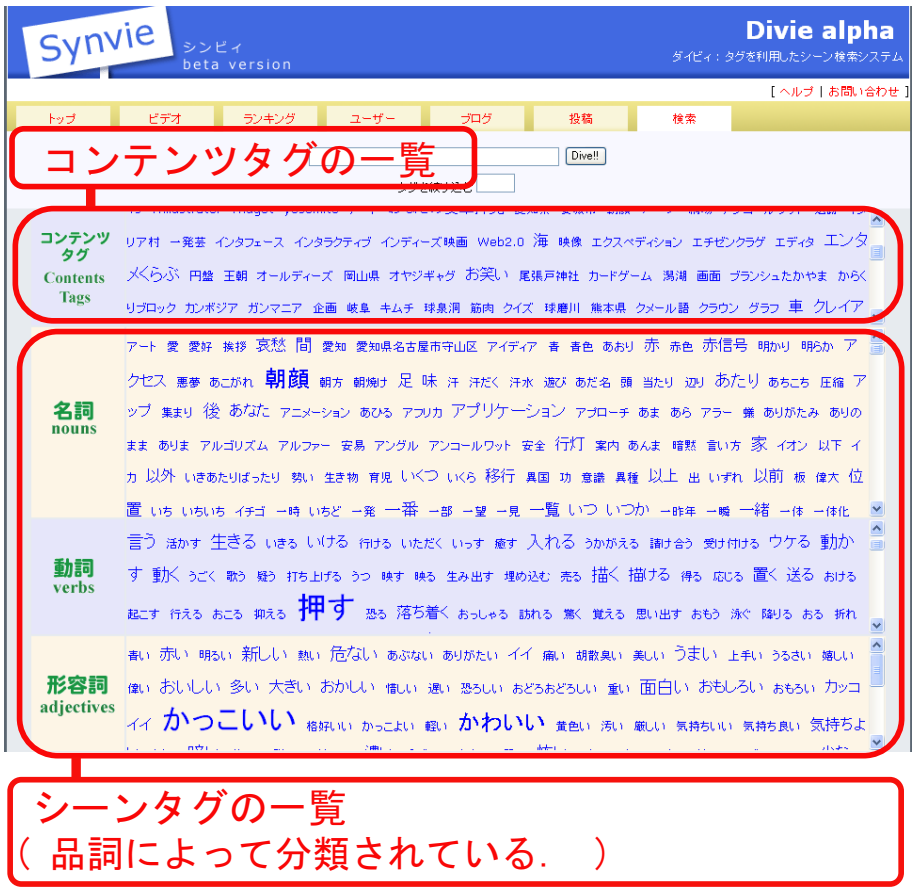

図 6.2: 映像シーン検索のためのタグクラウド

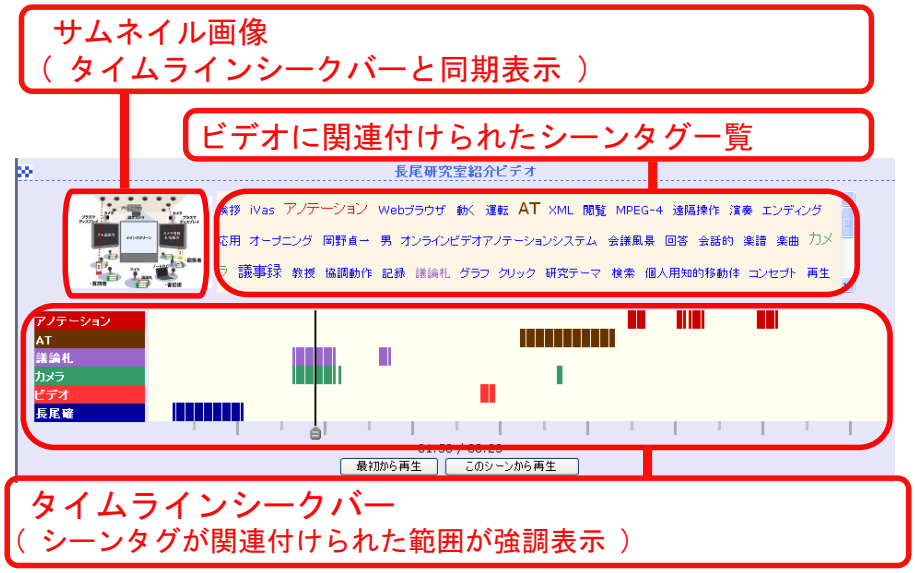

図 6.3: 映像シークバーとタグクラウドに基づくビデオ検索インタフェース

タグ選別システムを図 6.4 に示す．システムは自動的に抽出されたシーンタグを動画の閲覧中のユーザに提示する．ユーザは現在の映像シーンと関連性の高いタグを選択することによって，タグの選別を行うことが可能である．このタグを選別するのにかかる時間が，タグ選択にかかるコストである．なお，映像閲覧時だけではなく，先に述べたソーシャルブックマークをするときにも提示することを考えている．本システムは Web ベースのシステムであり，多くのユーザが参加することによって，一人当たりのタグ選択にかかるコストが下がることを期待する．本実験においては，平均 57%のタグが有用であると判断され選択された．

6.2.3　実験

ビデオシーン検索に関する被験者実験を行った．実験条件は以下のとおりである．

Synvie から獲得されたタグ集合に基づき，20 代〜40 代の男性 9 人が 9 つの検索問題に対してビデオシーン検索を行った．Synvie に投稿されたビデオコンテンツの内 27 個のビデオを対象とした．ビデオの平均時間長は 349 秒である．検索問題は，解答シーンの内容を記述した文章とそのシーンのサムネイル画像をぼかした画像からな

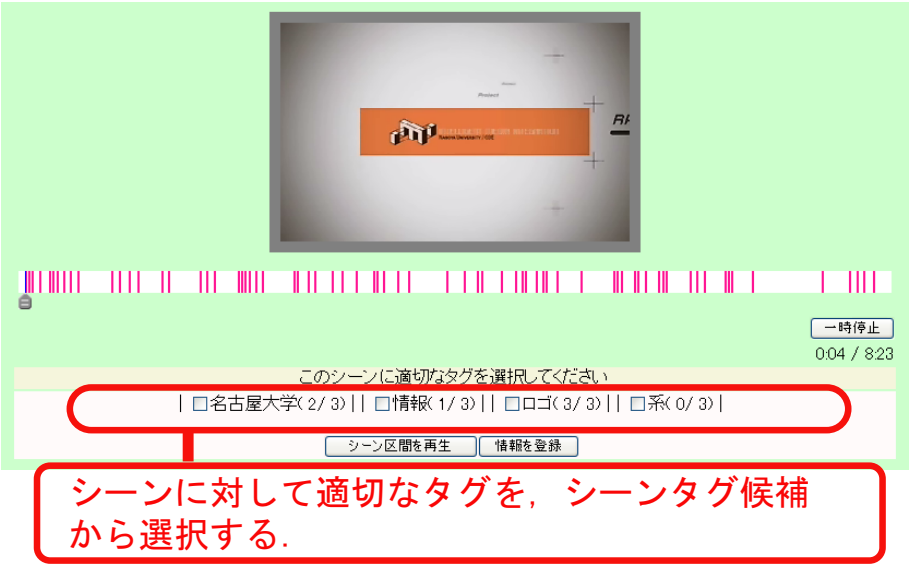

図 6.4: オンラインタグ選別システム

る．検索問題文に答えとなるシーンに含まれる特徴的なキーワードが含まれていると容易に答えが見つかる可能性があるため，なるべくそれらのキーワードが含まれないように留意した．不鮮明な記憶から目的のシーンを探し出すシチュエーションを想定している．問題文の例としては,「ある動物が親子で佇んでいるシーン」,「ある人がスノーボードでコースの外にクラッシュしているシーン」などがあげられる．正解シーンを見つけるまでの時間が検索コストとする．今回はすべての実験において正解シーンを見つけだすことができた．これは検索対象となったビデオコンテンツ数が少なく，探索的に正解となるシーンを探し出すことが可能なインタフェースであったため可能であったと推察できる．実験に用いたタグ集合としては，A: タグ選別前のタグ集合，B: タグ選別後のタグ集合，C: 比較検討用に出題者が人手で過不足無く全てのシーンにシーンタグを付与した理想的なタグ集合である．

6.2.4 議論

まず，タグ選別手法の適用による検索効率の改善について考察する．表 6.1 で示すように，タグ集合 A における平均検索時間 169 秒に比べてタグ集合 B の平均検索時間

表 6.1: タグ集合毎のコストパフォーマンス

タグ集合	作成コスト	検索時間	C/P
A	0	169	—
B	314	145	3.48
C	1480	118	7.18

145 秒となり，17 %の時間が短縮された．これによりタグ選択手法の適用は検索効率改善に有意であることを示した．しかしながら，理想的なタグ集合 C に基づく検索には敵わない．このままでは我々が提案した手法が必ずしも優位であるとはいえない．

次に，タグ集合 B と C について比較する．どちらも人手による明示的なコストをかけているため，コストパフォーマンスを比較する．本実験におけるコストパフォーマンスとは，タグ選別や作成にかかった平均時間とそれによる検索効率の平均改善時間の比を 100 倍した値である．これらはタグ作成や選別にかかわるインタフェースに大きく依存するので参考的な値ではあるものの，同様なインタフェースを用意することによって対処した．表 6.1 で示すように，タグ集合 C のコストパフォーマンスは 3.48 であるのに対して，タグ集合 B のコストパフォーマンスは 7.18 である．これによって，コストパフォーマンスの観点からも，タグ選択の仕組みが有意であることが分った．

これらの結果により，Synvie から抽出されたタグが検索に利用可能であることを示した．しかしながら，実際に適用可能かどうかはどれだけアノテーションが集まるかに大きく依存するので断定的な評価は今後の課題としたい．また，タグ選別が検索効率の改善に役立つことがわかった．しかしながらすべての観点からタグ選択が優れているとはいえず，対象となるコンテンツや検索頻度等を考慮して使い分けることが重要である．

6.3　映像の構造情報を用いた応用

Synvie では，映像コンテンツの内容と関連付けられたタグが抽出できるだけではなく，シーン引用やアノテーションに基づく構造的情報の取得が可能である．例えば，1 つのブログエントリーで複数の映像コンテンツを引用した場合，それらの映像コンテ

ンツには何らかの関連性があるはずである．また，1つのブログエントリー内の1つの段落で同時に複数の映像シーンを同時引用した場合にも，それらの映像シーンは何らかの関連性があるはずである．このような引用の共起関係に基づく映像コンテンツの関連性の抽出が可能になる．さらには，多くのユーザによって引用，あるいはアノテーションされているシーンほどより重要であるといった重要度の計算も可能になる．このような，映像の構造的情報を用いた応用例をいくつか提案する．

6.3.1 ビデオシーン推薦システム

ビデオシーン推薦システムとは，現在閲覧している映像の内容に基づき，関連する映像コンテンツを効果的に提示する仕組みである．具体的には，映像と同期して関連性のある他のコンテンツのサムネイル画像とキーワード，及びその根拠となるブログエントリーを表示する，図6.5のシステムである．従来からある閲覧履歴に基づくソーシャルフィルタリング[32]を用いたコンテンツ推薦システムでは，映像のシーン単位で推薦を行うことは困難であること，サンプル数が少ない場合には必ずしも精度が良くないこと，さらに推薦となる根拠を統計的にしか示すことができないなどの欠点がある．本システムでは，ユーザが複数コンテンツを引用したビデオブログエントリーを執筆した情報を用いて，少ないサンプル数で詳細なコンテンツ推薦を実現可能である．

図 6.5: ビデオ推薦システム

6.3.2　ビデオスキミングシステム

さらに，図 6.6 のような，ビデオスキミングシステムを開発している．ビデオスキミングとは，映像の重要なシーンのみを通常の速さで再生し，それ以外のシーンを早送りで再生する仕組みである．これにより，映像の内容を短時間で把握するのに適している．具体的には，映像シーンの重みや意味的な関連性に基づき再生速度を決定する．映像シーンの関連性は，重要なシーンに関連しているシーンは重要であるという考えに基づき関連する映像シーンの重みを伝播させる場合や，関連しているシーンは内容の重複があるという考えに基づき映像シーンの重みを減衰させる場合など，ユーザの目的や短縮したい時間によって使い分ける．

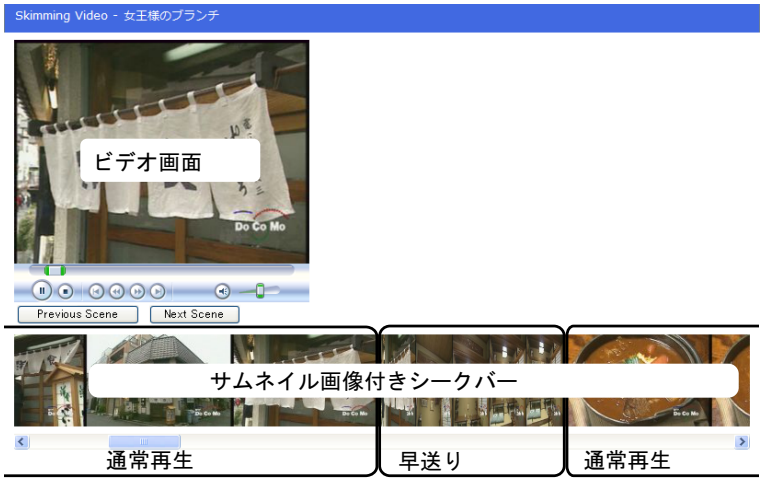

図 6.6: ビデオスキミングシステム

6.4　本章のまとめ

本章では，Synvie によって獲得されたアノテーションに基づくいくつかの応用について提案した．Synvie で獲得されたアノテーションは自由テキストであり，またそのテキスト品質も必ずしも良いとは言えないため，高度な自然言語解析処理を適用することは困難である．そこで，我々が開発した応用システムでは，比較的容易な自然言語解析技術で実現可能である，アノテーションからタグ集合を獲得し，それに基づく応

用システムを開発した．３章において有用なアノテーションとは構造的アノテーションと意味的アノテーションであると述べた．意味的アノテーションはタグ集合という形式で獲得すること，構造的アノテーションの獲得に関しては映像シーンの共起引用関係を利用することと参照度合いによるシーン重要度を提案したことによって獲得している．具体的な応用として，2 種類の映像シーン検索システムと映像の構造情報に基づいた応用システムを提案した．

しかしながら本章で述べた応用は，アノテーションに基づく応用の例であり，工夫次第で様々な応用に利用可能であると考えられる．そのためには，アノテーションをより体系的に解析するためのモデルを提案し，また，より高度な自然言語解析技術を用いる，あるいは，より情報量の多いアノテーション作成手段を提供するなどの工夫が必要になる．また，これらの応用の評価は，アノテーションの品質や量にきわめて依存するため単体での評価が難しい．そのためには，アノテーションに基づく応用の評価手法についても確立していく必要がある．これらの問題は今後の課題としたい．

次章では，本論文に関連する研究について述べる．

第7章　関連研究

本論文に関連する研究について述べる．具体的には，1) 映像アノテーションに関する研究，2)Web アノテーションに関する研究，3) 映像検索に関する研究及び 4) 映像共有サービスについて調査した．映像アノテーションに関する研究では，映像アノテーションに対する古典的な仕組みから近年標準化された MPEG-7 やその関連ツールについて言及する．Web アノテーションに関する研究では，映像やテキスト，画像や音楽などのコンテンツにタグやコメントなどのアノテーションを関連付ける Web システムを紹介する．映像検索に関する研究においては，自動解析技術に基づく検索システムやグラフ構造を利用した検索システムにいて紹介する．最後に，映像共有サービスにおける代表的なサービスをいくつか紹介する．これらの仕組みと我々の研究との関連性について言及することによって，我々の研究の優位性について議論する．

7.1　映像アノテーションに関する研究

まず映像コンテンツに対するアノテーションに関する研究について述べる．

映像アノテーションに対する取り組みとして，古典的な例は，MIT(マサチューセッツ工科大学) のメディアラボで Daivs ら [6] によって開発された MediaStream があげられる．MediaStream は，アイコンの集合によって記述可能なビジュアル言語を用いて映像に意味内容を関連付ける．３０００以上ものアイコンとその組み合わせによって映像シーンにアノテーションを作成する．映像の内容検索などの高度利用を考慮した場合，映像に任意のキーワードやコメントをアノテーションとして関連付けるよりも，アイコンを利用して意味属性を関連付けた方が客観性が高い利点がある．この仕組みの欠点は，映像をアイコンで表現するために非常に多くのアイコンが必要となり，またアイコンの組み合わせで表現できない意味内容を記述することができないという点である．事実，開発の最終工程では，アイコンの数が６０００個にもなり，アイコンの選択に手間取ること，アイコンの意味する内容が分かりにくい，複数のアイコン

の組み合わせといった高度な作業が必要になるという欠点がある．あらかじめ用意された情報の組み合わせで記述し，さらにそれらのアイコンのグループ化を行っているという点においてオントロジー型アノテーションに分類されるが，明示的にオントロジーとの関連性については述べられていない．

XML や XML スキーマの登場により，映像に対する内容を XML 形式で記述しようとするアノテーション方式が登場した．XML を用いたアノテーションの仕組みとして，MPEG-7[18] がよく知られている．MPEG-7 では，おもに単体の映像コンテンツに対して専任の作業者が，映像シーン検索や要約などの応用を実現するための有用で信頼性の高い情報を記述するための枠組みである．記述形式を詳細に標準化してあるものの，仕様が細かすぎて実用化にはなかなか結びついていない．何らかのツールを用いて MPEG-7 の記述を支援する必要がある．また，不特定多数のユーザが Web 上で自由コメントを執筆することは想定されていないなど，近年の Web 映像共有サービスを前提としたアノテーションのために利用することはできないなどの欠点がある．

MPEG-7 の記述を支援するための仕組みとして，MovieTool[29] や VideoAnnEx[20] やビデオアノテーションエディタ [23] があげられる．MovieTool[29] はリコーが開発した MPEG-7 記述用の映像アノテーションツールである．映像コンテンツに対して階層構造表現など構造的アノテーション情報や意味的アノテーション情報の記述が可能である．また専用の GUI を用いたアノテーションツールと，対応する MPEG-7 をテキスト形式で直接編集できるモードがある．IBM が作った MPEG-7 対応オフラインビデオアノテーションツールである VideoAnnEx Annotation Tool[20] がある．これは，映像コンテンツに対して，シーンやイベントやオブジェクなどの意味属性を付与することができるオントロジー型アノテーションツールである．意味情報を木構造型のオントロジーとして管理しており，またオントロジーの新規追加が可能である．長尾らが開発したビデオアノテーションエディタ [23] は，音声認識技術や映像解析技術とそれらの解析結果を修正する機能を備えている．これらの仕組みは，単体の映像コンテンツに対してオフラインでアノテーションを付与するためのツールあり，複数ユーザによってアノテーションを作成することは考慮されていない．

これらのアノテーション技術は，不特定多数の Web ユーザによってアノテーションが作成されることは考慮されていない．しかしながら，映像に対する意味内容を記述するという点においては，深く考察されており，ある特定の応用やコンテンツを考慮した上では，特に有用であると考える．

7.2 Web アノテーションに関する研究

一方，Web コンテンツに対するアノテーションに関する研究について述べる．

Web サイトに対するアノテーションの例としては Annotea[19] が存在する．記述形式として RDF を採用し，任意の Web コンテンツに対する注釈を付与し，それらの注釈を共有することが可能な仕組みである．現状では，専用の Web ブラウザが必要な点などの欠点もあるが，Javascript や AJAX の機能を用いて，通常の Web ブラウザで利用可能なシステムもいくつか存在する．これらの仕組みは，あくまでも注釈を共有することによって，情報を共有することを目的としており，コンテンツの中身を検索することは考慮されていない．なぜならば，通常の Web サイトは注釈無しで，Google などの Web 検索システムによって高い精度でコンテンツの検索が可能になるためである．

画像を共有するためのサービスである Flickr[10] は画像アノテーションとしての機能を備えている．Flickr では，個人が撮影した写真を簡単に Web 上で共有できる仕組みであり，写真の部分に対して任意のユーザがコメントを記述できること，写真全体に対してタグを付与し，そのタグに基づいて写真の共有と分類が可能になること，既存の Weblog エントリーに写真を容易に引用可能である特徴がある．タグを用いた情報の分類手法をタクソノミーと呼び，Web2.0 と呼ばれる Web 技術の草分け的存在である．Flickr はタグ集合型アノテーションの仕組みであり，ユーザコメントはあくまでもコミュニケーションのために利用される．このような仕組みを映像に適用した仕組みとして YouTube などがあげられる．

また，画像にタグをゲーム感覚で付与する仕組みとして，Google Image Labeler[12] がある．これは，対戦型のオンラインゲームであり，対戦者が互いに１つの画像に対して連想するだろうタグを入力し，一致したタグの数に応じて得点が増えるゲームである．タグ付与にかかる１人あたりの人的コストが最小化できるばかりか，エンタテインメントとしての側面も持ち合わせた仕組みである．映像を話題としたコミュニケーションもエンタテインメントの一種であると考えれば，我々が提案する Synvie においても同様な効果が期待できる．

音楽に対するアノテーションの仕組みとして，梶らの研究 [45] が存在する．音楽の時間軸や楽譜に対してユーザがアノテーションを付与することができる仕組みである．楽譜の任意の要素に対してユーザの主観的な感想や評価を関連付けることが可能であ

り，検索や推薦などの応用システムについても実現している．映像ではなく音楽に対するアノテーションであるという違いのほかに，我々の仕組みは，引用の仕組みがあること，ユーザコメントを解析し意味情報の抽出を行っていること，公開実験を行っている点において有意である．

ユーザ間で動画に対するコメント情報を共有するツールとして，NTT の山田ら [40] が作成した SceneNAVI というのがある．これは，ビデオコンテンツの時間軸に沿ってオンラインでコメントするツールであり，コメントに対して返信や検索ができる．これは本来ビデオアノテーションを行うツールではないが，投稿されたコメント情報を利用したビデオコンテンツの検索なども考えているようである．また，本研究とは違い，単純なコメント情報のみしか投稿できない，引用ができないなどの違いがある．

また，映像と外部の Web サイトを関連付けてアノテーションを抽出する研究の例としては，Dowman ら [8] による，ニュース映像の音声認識結果と CNN の Web ニューステキストの内容を比較することによって自動的に該当するニュース記事を特定し，その記事から映像コンテンツに関連した情報の取得を試みる仕組みがあるが，ニューストピック単位での関連付けであるため粒度が荒く，映像コンテンツはニュース記事に限定され，また，音声認識や言語解析結果にきわめて依存したリンクであるため，そのリンク自体の精度や再現性も高くない．

このように Web アノテーション技術の多くは，コメントを共有することに価値を見出しており，本論文におけるアノテーションとしての再利用を前提としている仕組みは少ない．

7.3　映像検索に関する研究

次に映像検索に関する研究について述べる．

カーネギーメロン大学の Informedia プロジェクト [36] 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た．画像認識，音声認識，自然言語処理の技術を統合して１０００時間ものビデオ映像の自動索引付けを行ったシステムである．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，tr-idf 法 (term frequency/inverse document frequency) 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．構造的情報を映像処理技術によって，意味的情報を音声認識技術やクローズドキャプションの取得からコンテンツの意味情報を抽出している．

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．是津耕司ら [44] 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と思われるものに対して，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．

このように映像検索に関するアプローチも種々なものが存在する．

7.4 ビデオ共有サービスに関する研究

映像共有サービスについて述べる．YouTube などに代表される映像共有サービスとは，一般ユーザが映像コンテンツを Web から投稿し，その映像コンテンツを他のユーザが自由に閲覧することができる仕組みである．

YouTube[38] や Google Video[14] などすでにいくつか限定的ながら提供されている．これらのサービスでは，コンテンツ閲覧者がその映像に対してのコメント付与による掲示板型コミュニケーションや個人のブログへの埋め込みなどが日常的に行われている．これらを映像に対するアノテーションとしてとらえることは可能であるが，アノテーションの対象がコンテンツ単位であるなど粒度が荒く，映像のシーン検索などの応用に利用することは困難であり，限定的な応用にしか利用できない．

映像の内部要素にコメントを記述できる仕組みとして，ニコニコ動画 [39] があげられる．ニコニコ動画は，Synvie のシーンコメントアノテーション機能を参考にして開発された．特徴としては，映像の中身にコメントを投稿可能であり，投稿されたコメ

ントが画面上にオーバラップされて表示される仕組みである．従来の動画共有サービスとは違って，他者のコメントが映像と同期して閲覧できるため，他のユーザと臨場感が共有でき，エンターテイメント性が強いという特徴がある．事実，400 万人以上の会員を有し，7 億件以上ものコメントが投稿されているなど，利用率も非常に高い．また，これらのコメントがコンテンツ投稿者のやる気を引き出し，創作性のある新しいコンテンツが他の動画共有サービスと比べて多く投稿されている．その一方で，著作権を侵害したコンテンツが多く，創作性が認められるコンテンツであっても音楽や映像素材などは他者の著作権を侵害している場合が極めて多く，違法コンテンツによって成立しているサービスであるという側面もある．また，我々の研究とは違って，投稿コメントをアノテーションとして再利用を考慮していない，ブログへのシーン引用がサポートされていないという違いもある．それまで，映像の内部要素に対してコメントを付与するサービスは，我々のシステムしか公開されておらず，一般にどのくらい受け入れられるかは未知数であったが，ニコニコ動画の成功により我々の仕組みの有意性が裏付けられたという側面もある．

第8章　結論

本章では本論文で述べた研究についての総括を行う．まず，本論文の内容についてまとめる．次に，本論文でやり残した問題点を指摘する．最後に今後の研究の展望について述べる．

8.1　本論文のまとめ

本論文では，映像コンテンツに対するアノテーションという切り口で，さまざまなアプローチについて検討してきた．従来，とりわけ基礎研究の分野において，映像コンテンツの高度利用に対するアプローチは自動解析技術のみに基づく手法に傾倒してきた．映像解析技術は映像コンテンツの構造情報を理解することは可能であっても，映像コンテンツの意味を理解することは困難であり，将来にわたって信号処理技術のみでの実現は困難であると予想する．我々の提案手法は，人手によって意味内容を映像コンテンツと関連付けることを支援する仕組みである．いかにアノテーションのコストを低く，また，より正確なアノテーションを関連付ける仕組みを作るかということが問題になる．

2章において，アノテーションのモデルについて検討した．まず，映像コンテンツの任意の要素を Element Pointer という枠組みを用いて一意に特定する仕組みについて述べた．さらに，それらの要素とアノテーションを RDF の仕組みを用いて関連付ける仕組みについても述べた．ここまでは，様々なアノテーションに対して共通的に利用可能な仕組みであるが，アノテーションにどのような内容を記述するかによって，様々なアノテーション型があることを示した．具体的には，スキーマ型アノテーション，オントロジー型アノテーション，タグ集合型アノテーション，ユーザコメント型アノテーション，引用型アノテーションの5つの形式である．本論文では，特に，ユーザコメント型アノテーションと引用型アノテーションを提案した．

3 章において，オントロジーに基づく半自動映像アノテーションシステムについて述べた．このアノテーションは主にスキーマ型アノテーションとオントロジー型アノテーションの実装例である．具体的には，映像コンテンツに対する 2 種類のアノテーションの作成を支援する．1 つは，映像に対する構造アノテーションであり，信号処理技術とその解析結果の修正を支援する仕組みを備える．もう 1 つは，映像に対する意味的なアノテーションであり，映像に関するオントロジーと映像の要素を関連付けることによって行うオントロジー型アノテーションの仕組みを備える．機械にできることは極力機械によって，人間にしかできないことは人間によって，協調的にアノテーションを作成する仕組みである．この方式のアノテーションは，専門家が商用コンテンツに対して詳細にアノテーションを付与することに適している一方，アノテーション作成コストが比較的高いので Web で共有されている映像コンテンツに対して同様の仕組みを適用することは困難である．

4 章において，閲覧者によるオンライン映像アノテーションについて述べた．このアノテーションは主にユーザコメント型アノテーションの実装例である．具体的には，映像の閲覧者にアノテーションの参加を促す仕組みである．そのために，映像シーンを話題とした掲示板型コミュニケーションを支援するユーザコメント型アノテーションの仕組みを備える iVAS というシステムを開発した．また，不特定多数の閲覧者がアノテーションを作成する方式であるためアノテーションの信頼性が一様ではない問題があるため，アノテーション信頼度と呼ばれる評価手法を導入し，それに基づくアノテーションの選別手法を提案した．さらに，閲覧者のアノテーション履歴から閲覧者の嗜好を評価するための仕組みとして，印象距離と呼ばれる指標を提案し，これらの評価を行った．

5 章において，Web コミュニティ活動に基づく映像アノテーションについて述べた．このアノテーションは主にユーザコメント型アノテーションと引用型アノテーションの実装例である．4 章では，閲覧者によるアノテーション方式を提案したが，映像を話題としたコミュニティ活動は映像の閲覧時だけではなく，ブログなどでも映像を話題とした記事が書かれている点に着目し，それらの情報もアノテーションとして獲得した．具体的には，映像シーンを引用したブログエントリーの執筆を支援する，引用型アノテーションの仕組みを提案した．引用型アノテーションでは，映像に関する意味情報だけではなく，引用の共起関係に基づく構造情報の獲得が可能である．さらに，これらの仕組みを実装したシステムとして Synvie を開発し，比較的大規模な公開実験

を行った．2006 年 7 月から一般公開し，284 名の登録ユーザと，223 個の映像コンテンツと，3534 個のユーザコメントに基づくアノテーションを含む 19182 個のアノテーションの取得に成功した．これらの取得されたアノテーションの定量的な分析を行うことによって，引用型アノテーションの優位性を示した．

6 章において，アノテーションに基づく応用について述べた．具体的には，5 章の仕組みによって獲得されたアノテーションに基づく検索システムなどを試作した．具体的には，アノテーションのテキスト情報からタグ集合を生成することによって映像シーンに対する意味情報と，複数の映像シーンを同時引用することから得られる共起頻度やアノテーションの参照頻度を用いた構造情報を利用した．また，タグクラウドの仕組みを利用した映像シーン検索システムでは評価実験まで行った．これにより，アノテーションの有用性について確認した．

最後に，7 章において，映像アノテーションに関する研究に関する関連研究について述べた．

Web ユーザやコミュニティによって映像コンテンツに対するアノテーションを作成するというアプローチが本論文の最大の新規性にあたる部分である．また，実際に実証実験まで行うことによって，本論文の有用性を検証した．それにより，Web コミュニティからアノテーションを獲得することは有用であることを示した．Weblog を書く，映像を閲覧しながら掲示板型のコミュニケーションをするなどといった Web コミュニティ活動は，ユーザによって自発的に行われる活動であり，実質的なコストを 0 として捉えることができる．そういった意味で，映像コンテンツの自動解析技術と同等に，人的なコストが掛からず意味情報の抽出が可能になる点が新しい．しかしながら，Web コミュニティから取得されたアノテーションのみで十分な精度が得られるとはいえない．将来的には，信号処理技術的なアプローチと，人間が介在したアノテーション的なアプローチとを融合させた仕組みを実現させることが重要である．これらの仕組みが実現することによって，コンテンツ作成者にとっても，コンテンツ閲覧者にとっても，コンテンツ権利者にとっても有益なシステムになることを期待する．

8.2 今後の課題

今後の課題について述べる．本研究においては，3 章においてオントロジーに基づく半自動映像アノテーションの仕組みについて述べた．3 章は応用目的を特に意識し

た仕組みであり，ある意味理想的なアノテーションの獲得ができている．3 章によって獲得されたアノテーションと同等なものをいかに，ユーザコミュニティ活動から獲得するかが本研究の目標である．3 章においては，意味アノテーションと構造アノテーションの観点から作成を行った．この観点から今後の課題を述べる．

意味アノテーションの観点から述べる．ユーザコメントから作成したタグ集合と 3 章で述べたオントロジー集合とをいかに関連付けるかという問題が残る．つまり，映像に関するオントロジーを効果的に作成する手法を提案し，それらのオントロジーとタグ集合とを関連付ける必要がある．映像に対するユーザコメントを大量に収集することによって，オントロジーの作成と関連づけが同時にできることが望ましいが，必要最小限の人間の手が介在しなければならないかもしれない．また，より高度な自然言語解析処理を用いて，タグ集合以外の意味情報の記述方式を検討する必要がある．文章に対する構文情報や意味情報を機械にとっても理解可能な記述形式として，GDA[42] が存在するが，ユーザコメントを GDA 形式で蓄積する手法も検討する必要がある．

構造アノテーションの観点から述べる．本研究においては，映像シーンに対するアノテーションや引用の参照度合いから映像シーンに対する重要度を計算しているが，具体的な計算手法の妥当性に対する評価を行っていない．また，映像シーンの意味内容を考慮した重要度の算出手法ではないので，なんらかの意味内容を検討する必要があるだろう．また，複数の映像シーンの同時引用などから映像シーンの共起頻度を算出し，コンテンツの類似度を計算する手法についての検討する必要がある．また，映像の内部要素に対する引用やアノテーションに基づいて，より詳細かつ広範囲な引用に基づくネットワークが作成されるが，それらの情報を用いた解析や応用についても検討していくと良い．

さらに，アノテーションシステムについて詳細に評価するためには，そのアノテーションシステムを用いて，コンテンツに対してどれだけ必要十分なアノテーションが付与されているかを評価するための基準を検討する必要がある．これらの評価基準を自動的に計算できれば，アノテーションが不足しているコンテンツやシーンに対して集中的にアノテーションを付与することが可能になるなど利点も大きい．

また，映像アノテーションのインタフェースの検討や，他のマルチメディアコンテンツに対して本論文で述べた手法を適用していくことについても検討していきたい．

8.3 今後の研究の展望

今後の研究の方向性については 2 つの方向性を考えている．

1 つは，映像コンテンツの種類を限定することによって，より詳細かつ高度なアノテーションの獲得手法を検討することである．講義コンテンツの映像配信と共有に関するシステムの開発と実証実験を検討している．講義コンテンツの配信に対する社会的要請は日に日に高まっており，また，それらのコンテンツを効率よく管理・蓄積・検索できる仕組みを提案する必要がある．具体的には，講義に関する意味情報をアノテーションとして蓄積する機能を備えた講義コンテンツ共有システムを開発する．講義映像配信に関連したシステムを開発するための基盤として機能するのと同時に，これらのシステムを利用したユーザの編集履歴を講義映像に対するアノテーションとして蓄積・管理する仕組みを備える．具体的な機能としては，コンテンツを投稿・管理する仕組み，映像と同期してスライドを表示する仕組み，映像やスライドの部分に対するアノテーションや引用の支援，それらを蓄積する仕組みを備える．また，具体的なアプリケーションとして，受講者が講義内容を引用したレポートの執筆支援ツール，講義内容と連携した質問掲示板システムを開発する．これらの取得されたアノテーションと，講義オントロジーを関連付け，より詳細なアノテーションの解析を実現したい．

もう 1 つは，Web コミュニティによる映像コンテンツの協調的作成支援の仕組みである．個人が映像コンテンツを作成する機会の増加に伴い，容易に質の高い映像コンテンツを作成する仕組みが必要になる．映像に関する素材の共有や，映像作成に関するノウハウの共有ができる仕組みを開発したい．また同時に，著作権管理の問題や容易に映像を編集できるインタフェースを提案したい．これによって，本研究で述べた映像を配信する技術だけではなく，映像を作成する側も支援することによって，映像作成段階で取得できる情報をアノテーションとして獲得する仕組みを考案したい．

本論文で提案した仕組みや今後の研究の方向性で述べた仕組みを実現し社会に還元することによって，映像コンテンツ産業や Web の発展に貢献できれば良いと考えている．

謝辞

本論文は，名古屋大学大学院情報科学研究科メディア科学専攻長尾研究室に在籍していた期間に行われた研究をまとめたものです．本研究を遂行するにあたり，多くの方々の御指導，御支援を頂きました．ここに心からの感謝の意を表します．

とりわけ，指導教員である長尾確教授には，自由な研究環境を提供して頂き，また日頃より，研究の心構えなど基礎的なことから研究全般に関する御指導と御鞭撻を頂きました．さらに，今後の研究活動を行っていく上でも有益となるような，幅広い知見に基づく，多くの貴重な御意見を頂きました．

大平茂輝助教には，日頃の研究に関する様々な相談に常に親身になって対応して頂き，ゼミなどを通じて，非常に多くの御助言と御支援を頂きました．

情報科学研究科メディア科学専攻の武田一哉教授と村瀬洋教授には，本論文の審査活動を通じて，幅広い角度からの的確な御助言と御示唆を頂き，本論文をまとめる上で大変有益なものになりました．

また，梶克彦さん，山根隼人さんを始めとする研究室の諸先輩方には，研究はもとより私生活の面においてもサポートして頂き，有意義な研究生活を送ることができました．研究室の歴代の秘書である，兼松英代さん，金子幸子さん，鈴木美苗さんには研究生活のための様々なサポートを頂きました．情報科学研究科博士過程の同期である，清水敏之さん，原直さんにも様々な点において感謝しています．また長尾研究室の皆様には研究を進めるにあたり有益な議論と様々な励ましを頂きました．

また，本研究を遂行するにあたり，独立行政法人情報処理推進機構 2005 年度上期未踏ソフトウェア創造事業による支援と，独立行政法人日本学術振興会 2006-2007 年度 特別研究員奨励費による支援を受けました．これにより資金的な面も含めて研究を円滑に遂行することが可能になりました．さらに，名古屋大学 21 世紀 COE において 2005 年度リサーチアシスタントに，独立行政法人日本学術振興会において 2006-2007 年度特別研究員に採用されました．これにより研究に専念することができました．

最後に，ここまで育てて頂いた両親に感謝致します.

参考文献

[1] Esma Aimeur, Gilles Brassard, and Sebastien Paquet. Using personal knowledge publishing to facilitate sharing across communities. In *Proceedings of the Twelfth International World Wide Web Conference (WWW2003)*, 2003.

[2] Apache xindice. http://xml.apache.org/xindice/.

[3] Gabe Beged-Dov, Dan Brickley, Rael Dornfest, Ian Davis, Leigh Dodds, Jonathan Eisenzopf, David Galbraith, R.V. Guha, Ken MacLeod, Eric Miller, Aaron Swartz, and Eric van der Vlist. *RDF Site Summary (RSS) 1.0*. RSS-DEV Working Group, http://web.resource.org/ rss/1.0/spec, 2001.

[4] Benjamin and Mena Trott. *mttrackback - TrackBack Technical Specification*. movabletype.org, http://www.movabletype.org/docs/ mttrackback.html, 2002.

[5] Dan Brickley and Libby Miller. *FOAF Vocabulary Specification*. http://xmlns.com/foaf/0.1/, 2004.

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

[7] del.icio.us. http://del.icio.us/.

[8] Mike Dowman, Valentin Tablan, Hamish Cunningham, and Borislav Popov. Web-assisted annotation, semantic indexing and search of television and radio news. In *Proceedings of the the 14th International World Wide Web Conference 2005 (WWW 2005)*, pp. 225–234, 2005.

[9] Dieter Fensel, James Hendler, Henry Lieberman, and Wolfgang Wahlster. *Spinning the Semantic Web: bringing the World Wide Web to its full potential*. The MIT Press, 2003.

[10] Flickr. http://www.flickr.com/.

[11] Gmail. http://mail.google.com/.

[12] Google Image Labeler. http://images.google.com/imagelabeler/.

[13] Google Maps. http://maps.google.com/.

[14] Google Video. http://video.google.com/.

[15] Toshio Hirotsu, Toshihiro Takada, Shigemi Aoyagi, Koji Sato, and Toshiharu 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.

[16] Paul Hoffman and Tim Bray. *Atom Publishing Format and Protocol (atompub)*. http://www. ietf.org/html.charters/atompub-charter. html, 2005.

[17] Dublin Core Metadata Initiative. Dublin core. http://dublincore.org/.

[18] ISO. *Information Technology - Multimedia Content Description Interface(MPEG-7), ISO/IEC 15938:2001*. International Organization for Standardization(ISO), 2001.

[19] José Kahan, Marja-Riitta Koivunen, Eric Prud'Hommeaux, and Ralph R. Swick. Annotea: an open rdf infrastructure for shared web annotations. *Computer Networks*, Vol. 39, No. 5, pp. 589–608, 2002.

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

[21] Frank Manola and Eric Miller. *RDF Primer, W3C Recommendation 10 February 2004*. W3C, http://www.w3.org/TR/rdf-primer/, 2004.

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

[23] 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.

[24] 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.

[25] Eitetsu Oomoto and Katsumi 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.

[26] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. In *Technical report, Stanford Digital Library Technologies Project*, 1998.

[27] Conrand Parker and Silvia Pfeiffer. Video blogging: Content to the max. *IEEE Multimedia*, Vol. 12, No. 2, pp. 4–8, 2005.

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

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

[30] A. W. Rivadeneira, Daniel M. Gruen, Michael J. Muller, and David R. Millen. Getting our head in the clouds: toward evaluation studies of tagclouds. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 995–998, 2007.

[31] SemanticWeb.org. Semantic web. http://www.semanticweb.org/.

[32] Upendra Shardanand and Patti Maes. Social information filtering: Algorithms for automating “word of mouth”. In *Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems*, Vol. 1, pp. 210–217, 1995.

[33] John R. Smith and Blaise Lugeon. A visual annotation tool for multimedia content description. In *Proceedings of the SPIE Photonics East, Internet Multimedia Management Systems*, pp. 49–59, 2000.

[34] Technorati. http://technorati.com/.

[35] W3C. Synchronized multimedia integration language (SMIL 2.0) specification. http://www.w3.org/TR/smil20, 2002.

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

[37] Wikipedia. http://www.wikipedia.org/.

[38] Youtube. http://www.youtube.com/.

[39] ニコニコ動画. http://www.nicovideo.jp.

[40] 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, 第 2001-GN-43 巻, pp. 37–42, 2001.

[41] 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.

[42] 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.

[43] 是津耕司, 上原邦明, 田中克己. 映像の意味的構造の発見. 情報処理学会論文誌, Vol. 41, No. 1, pp. 12–23, 2000.

[44] 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.

[45] 梶克彦, 長尾確. 楽曲に対する多様な解釈を扱う音楽アノテーションシステム. 情報処理学会論文誌, Vol. 48, No. 1, pp. 258–273, 2007.

[46] 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. http://chasen.aist-nara.ac.jp/, 2003.

[47] 神崎正英. セマンティック・ウェブのための RDF/OWL 入門. 森北出版, 2004.

[48] 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, 2003.

[49] 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.

[50] 山本大介, 増田智樹, 大平茂輝, 長尾確. Synvie:映像シーン引用に基づくアノテーションシステムの構築とその評価. インタラクション 2007, pp. 11–18, 2007.

[51] 山本大介, 清水敏之, 大平茂輝, 長尾確. Synvie:ブログの仕組みを利用したマルチメディアコンテンツ配信システム. 情報処理学会第 58 回グループウェアとネットワーク研究会, pp. 13–18, 2006.

[52] 増田智樹, 山本大介, 大平茂輝, 長尾確. オンラインアノテーションを利用したビデオシーン検索. 第 21 回人工知能学会全国大会 講演論文集, 2007.

[53] 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, 第 D-279 巻, 1996.

[54] 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, 第 PRMU2002-30 巻, 2002.

[55] 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.

[56] 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 79-4 巻, pp. 559–567, 1996.

研究業績

雑誌論文

- 山本 大介, 増田 智樹, 大平 茂輝, 長尾 確. 映像を話題としたコミュニティ活動支援に基づくアノテーションシステム, 情報処理学会論文誌, Vol.48, No.12, pp. 3624-3636, 2007.
- 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.

国際会議

- Daisuke YAMAMOTO, Shigeki OHIRA, Katashi NAGAO. Weblog-style Video Annotation and Syndication, The 1st International Conference on the Automated Production of Cross Media Content for Multi-channel Distribution (AXMEDIS), pp.235-238, 2005.12.
- Daisuke YAMAMOTO, Katashi NAGAO. iVAS: Web-based Video Annotation System and its Applications, The 3rd International Semantic Web Conference (ISWC2004), 2004.10.
- Katashi NAGAO, Katsuhiko KAJI, Daisuke YAMAMOTO, Hironori TOMOBE, Discussion Mining: Annotation-Based Knowledge Discovery from Real World Activities, The Fifth Pacific-Rim Conference on Multimedia (PCM 2004), pp.522-531, 2004.11.

国内シンポジウム

- 山本大介, 増田智樹, 大平茂輝, 長尾確. Synvie:映像シーンの引用に基づくアノテーションシステムの構築とその評価, 情報処理学会シンポジウム，インタラクション 2007, Vol.2007, No.4, pp.11-18 2007.3.

研究会

- 山本大介，清水敏之, 大平茂輝，長尾確. Synvie:ビデオブログコミュニティからの知識獲得とその応用,Web インテリジェンスとインタラクション:電子情報通信学会 第二種研究会資料 (WI2-2007-01), 2007.3.
- 山本大介，清水敏之, 大平茂輝，長尾確. Synvie:ブログの仕組みを利用したマルチメディアコンテンツ配信システム, 情報処理学会第 58 回グループウェアとネットワーク研究会, Vol.2006 No.9, 2006-DBS-138, pp 13-18, 2006.1.

その他の発表

- 山本大介, 増田智樹, 大平茂輝, 長尾確. Synvie: ビデオブログコミュニティから獲得されたアノテーションに基づく応用, 人工知能学会 第 21 回全国大会, 2007.4.
- 増田智樹, 山本大介, 大平茂輝, 長尾確. オンラインアノテーションを利用したビデオシーン検索, 人工知能学会 第 21 回全国大会, 2007.4.
- 山本大介, 増田智樹, 大平茂輝, 長尾確. Synvie:ブログの仕組みを利用したマルチメディアコンテンツ配信システム. 第 68 回情報処理学会全国大会, 2006.3.
- 山本大介, 大平茂輝, 長尾確. オンラインビデオコンテンツを中心としたコミュニティ支援システム. 第 67 回情報処理学会全国大会, 2005.3.
- 山本大介, 長尾確. 閲覧者によるオンラインビデオアノテーションとその応用. 情報科学技術フォーラム (FIT), 2004.9.
- 山本大介, 長尾確. Web ブラウザを用いた閲覧者による実時間ビデオアノテーション. 第 66 回情報処理学会全国大会, 2004.3.

- 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. 第 65 回情報処理学会全国大会, 2003.3.

受賞

- 2007 年 4 月. 人工知能学会第 21 回全国大会において優秀賞.（共著）
- 2007 年 3 月. 情報処理学会第 67 回全国大会において学生奨励賞.（共著）
- 2006 年 7 月. 情報処理学会 18 年度コンピュータサイエンス領域奨励賞.
- 2006 年 5 月. 情報処理推進機構 2005 年度上期 IPA 未踏ソフトウェア創造事業スーパークリエータ認定.
- 2006 年 1 月. 情報処理学会 DBS/GN/BCC 合同研究会において学生奨励賞.
- 2005 年 3 月. 情報処理学会第 67 回全国大会において学生奨励賞.
- 2004 年 3 月. 情報処理学会第 66 回全国大会において学生奨励賞.
- 2003 年 3 月. 情報処理学会第 65 回全国大会において学生奨励賞.

外部研究資金等

- 山本 大介. マルチメディアコンテンツへのアノテーションとその応用, 科学研究費補助金 (学振特別研究員 DC-2), 学術振興会, 2006.4-2008.3
- 山本 大介, 清水 敏之. マルチメディアコンテンツの配信とそのコミュニティ支援システムの開発, 2005 年度上期 IPA 未踏ソフトウェア創造事業, 情報処理推進機構, 2005.5-2006.2