

情報工学コース卒業研究報告

**オンラインアノテーションに基づく
ビデオシーン検索に関する研究**

増田 智樹

名古屋大学 工学部 電気電子・情報工学科

2007年2月

目次

第1章	はじめに	3
第2章	ビデオアノテーション	7
2.1	オフラインビデオアノテーション	8
2.1.1	オフラインビデオアノテーションとは	8
2.1.2	Video Annotation Editor	10
2.2	オンラインビデオアノテーション	11
2.2.1	オンラインビデオアノテーションとは	11
2.2.2	Synvie	13
第3章	シーンタグの作成	17
3.1	専用ツールを用いたタグ付け	17
3.2	オンラインビデオアノテーションからの自動抽出	19
3.3	オンラインタグ選択システムを用いた抽出	20
3.4	タグ選択実験	23
3.4.1	実験内容	23
3.4.2	実験結果	23
3.4.3	考察	25
第4章	ビデオシーン検索	31
4.1	タグを利用したビデオシーン検索システム	31
4.1.1	シーン検索インタフェース	32
4.1.2	検索アルゴリズム	35
4.2	シーン検索実験	37
4.2.1	実験内容	37
4.2.2	実験結果	38
4.2.3	考察	40
第5章	関連研究	43
5.1	オンラインビデオアノテーションに関する研究	43
5.1.1	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	43
5.1.2	番組実況チャットに基づく放送番組のビュー生成	43

5.1.3	ネットワーク簡易動画編集アプリケーション	44
5.1.4	分散環境下における複数人による動画編集	44
5.1.5	コミュニケーションを活発化させるインターネット街頭 TV	45
5.2	オフラインビデオアノテーションに関する研究	45
5.2.1	Informedia プロジェクト	45
5.2.2	Web 上の情報を利用したアノテーションの作成	46
5.2.3	ビデオアノテーションツール	46
第 6 章	まとめと今後の課題	49
6.1	まとめ	49
6.2	今後の課題	50
6.2.1	タグの付与されていないシーンの検索法の実現	50
6.2.2	シーンタグ作成法の改良	50
6.2.3	アノテーションをより大量に獲得するためのシステムの開発	50
6.2.4	ビデオコンテンツの新しい応用の考案と実現	51
	参考文献	51
	謝辞	55

概要

ビデオコンテンツの検索や要約などの応用のために、ビデオに対して、アノテーションと呼ばれる意味情報記述を作成する研究が行われている。従来、ビデオに対するアノテーションは、画像処理や音声処理などの機械処理や、人手によって詳細に付与するといった手法によって作成されてきた。これらの手法は一般にユーザサイドで利用する専用ツールで実行されるため、オフラインビデオアノテーションと呼ばれるが、アノテーションの質や作成コストなどのいくつかの問題点が存在する。一方、Web 上で行われている人間の自然なコミュニケーション活動から獲得されるビデオコンテンツに対する情報を、オンラインビデオアノテーションと呼ぶ。そこで、本研究では、オンラインビデオアノテーションから効率よく有用な情報を抽出する仕組みを提案する。

具体的には、オンラインビデオアノテーションに含まれるテキスト情報と時間情報からシーン検索に有用なタグを低コストで作成するためのオンラインタグ選択システムを開発した。このシステムは、オンラインビデオアノテーションから自動抽出されたタグがシーンに対して適切なものであるかをオンラインで Web ブラウザ上で複数の人間によって選択するシステムである。このシステムを用いた被験者実験を行うことで、シーンに対するタグの作成を行った。

次に、作成したタグの評価を行うために、タグを利用した新しい発想のビデオシーン検索システムを開発した。このシステムの特徴として、シーン検索の過程で、ビデオコンテンツに直接アクセスすることなく、シーン情報を Web ブラウザ上で閲覧できる点などがある。この検索システムを用いて、オフラインで人手で付与したタグと、タグ選択システムによって作成したタグ、オンラインビデオアノテーションから自動抽出したタグの 3 種類のタグを利用したシーン検索の被験者実験を行った。

一連の実験から、タグのコストパフォーマンス（検索効率とタグ作成コストの比率）の比較を行うことで、それぞれのタグ作成手法の比較を行った。その結果、オンラインタグ選択システムによるタグ作成手法が最も優れていることが確認できた。

このように、本研究を通して、オンラインビデオアノテーションの有用性を検証した。

第1章 はじめに

近年、インターネット技術の発達により、Web 上で膨大な数のビデオコンテンツが配信されるようになった。また、YouTube [25] などのビデオ共有サイトの出現により、誰でも手軽に Web 上にビデオコンテンツを公開することができるようになったため、一般人の作成したビデオも氾濫するようになった。これらの動きから Web 上のビデオコンテンツ閲覧者数も増加している。

このような流れは今後も続いていくと考えられ、ビデオコンテンツの量、閲覧者ともに増加していくと考えられる。そのため、ビデオコンテンツの検索や要約などの応用に対する要求が高まっていくと考えられる。

検索や要約などの応用を実現するために、ビデオに対してアノテーションと呼ばれる意味情報記述の作成を行う研究が行われている [1]。アノテーションとは、コンテンツに対して付与されるメタ情報であり、ビデオだけではなく音楽などのマルチメディアコンテンツに対してそれを付与する研究が行われている [2]。そのため、様々なコンテンツを関連付ける手段にもなりうる非常に可能性を秘めた研究分野である。

動画像処理、音声処理などの機械処理によってビデオに対するアノテーションを自動生成する研究が古くから行われている [12][13][14]。これらの手法は全自動でアノテーションを作成することができるという利点はあるが、得られる情報は、音声やテロップの情報、カット検出やオプティカルフローの情報などであり、ビデオに含まれる意味的な情報をすべて抽出することは不可能であると言える。また、特に一般的なビデオカメラを用いて制作されたビデオコンテンツは音声や画像などのノイズも多く含むため、機械処理の精度が上がらず、作成されるアノテーションの質も低いという問題がある。実際に、機械処理のみのアノテーションによって十分な精度で検索などの応用を実現した例はなく、限界があると考えられる。

機械処理による自動解析によって得られた情報に加え、アノテータと呼ばれる専門家がビデオに対して詳細なアノテーションを作成する研究が行われるようになった [4][15][16]。しかし、この手法の問題点はアノテーション作成のための人的コストが非常に高いということである。そのため、膨大な数のビデオコンテンツに対してそれを行うことは非現実的であると言える。さらに、一般人の作成したビデオに対してはビデオの制作者がアノテーションを詳細につけるということができないため、この手法は不向きであると言える。

そこで、筆者らは近年 Web 上で行われている自然なコミュニケーション活動から

得られる情報をオンラインビデオアノテーションとして利用する手法を提案する。

現在、ブログやSNS、電子掲示板などのコミュニケーションサービスが多く存在し、サービス利用者も増加傾向にある。これらのサービスを利用する一般人が発信する情報には、非常に意味のある情報が含まれ、時には社会に対して大きな影響を与えることもある。最近では国内最大の電子掲示板である2ちゃんねる [27] や、国内最大のSNSであるmixi [28] など書き込まれた情報による社会への影響が最も典型的な例であろう。この問題は社会現象とも言うことができ、一般人の発信する情報の影響の大きさを象徴している。

また、Wikipedia [26] は、Web上で誰でも編集可能な百科事典であるが、信頼性があり、また非常に広く深い情報を有する百科事典であるといえ、世界中で利用されている。辞書を編集するという活動も一種のコミュニケーション活動である。コミュニケーション活動から得られる情報を、大量に収集することで集合知と呼ばれる大きな利用価値のある情報となるという典型的な例である。クチコミサイトやQ & Aサイトなどもこの一例として挙げられる。

近年、このようなコミュニケーション活動が、ビデオコンテンツに対しても行われるようになった。ビデオコンテンツを中心としたコミュニケーション活動とは、例えば、ビデオに対して投稿されるコメントや、ビデオを引用したブログなどが挙げられる。今後もこのようなビデオコンテンツを取り巻くコミュニケーション活動が活発化していくと考えられる。

そこで、筆者らは、それらの情報から効率よく利用価値のある意味情報の抽出を行うというアプローチで研究を進めた。そして、それによって作成されたアノテーションを検証することで、オンラインビデオアノテーションの有用性の検証を行った。

そのためにはまず、オンラインビデオアノテーションを獲得する必要がある。現在多くのビデオ共有サービスが存在し、その中でもビデオコンテンツに対してコメントを投稿することができるサービスは存在するが、それらはビデオコンテンツ全体に対するコメントであることが多く、ビデオの内部シーンに対しての情報を抽出することが困難であり、ビデオアノテーションとしては不十分である。そのため、筆者の所属する研究室で開発されたオンラインビデオアノテーションシステム Synvie [3] の一般公開実験 [24] から獲得されたアノテーションデータを利用することとした。筆者も一般公開実験のプロジェクトメンバーとして、システムの運用やコンテンツの管理に携わった。Synvieの詳細は第2章で述べる。

オンラインビデオアノテーションは、集団の力が大きく影響するため各個人の負担すべきコストが極めて小さいと考えることができるという利点や、アノテーションがビデオの制作者に依存しないという利点もある。さらに、人間のコミュニケーション活動には機械処理では決して解析することのできない意味情報を含んでいる可能性がある。しかし、応用に利用可能な情報とそうでない情報が混在するという問題点もあるため、それらの情報を選別する必要がある。

本研究では、それをビデオシーン検索に利用することでオンラインビデオアノテーションの有用性を検証した。そのための仕組みを提案する。シーン検索以外にも、要約やその他の応用方法が存在する可能性もあるが、それは今後の研究課題である。

最近では、タグと呼ばれるマルチメディアコンテンツの検索や分類のために付与されるメタ情報の利用が広く行われている。タグを利用することで、複数のコンテンツをキーワードによって関連付けることが可能になる。タグは、検索や分類のために非常に有用なものである。タグを利用したサービスとしては写真共有サービスである Flickr [29] やソーシャルブックマークサービスである del.icio.us [30] などが世界的に利用されている。国内では、はてなブックマーク [31] などが存在する。

多くのビデオ共有サービスでも、ビデオコンテンツに対してタグが付与されるようになった。しかし、それらはビデオ全体に対して付与されるものであるため、ビデオコンテンツの検索には非常に有効な手法であるが、ビデオ内部の任意のシーンを検索したいという要求には応えることができない。

そこで、本研究では、ビデオシーンに対するタグ（以後シーンタグと呼ぶ）を作成することを考えた。つまり、ビデオの内部要素に対してキーワードを関連付けることで、それを基にシーン検索を可能にしようというアプローチである。

ビデオ共有サービスである Ask ビデオ [32] には、シーンタグの作成という機能がすでに存在する。具体的には、ビデオの視聴中に、任意のタイムコードに対してテキストを入力することができる。

しかし、入力するテキストには特に制約がないため、実際に作成されているシーンタグには、文章で表現されたテキストが多く含まれる。そのため、“タグ”という名称であるが、獲得できるアノテーションテキストは、Synvie で投稿されるコメントとあまり変わらない。そのため、Ask ビデオのシーンタグもそのままではキーワード単位でタグとして利用することはできない。また、Synvie のコメント投稿の方が、Ask ビデオのシーンタグ機能よりもコミュニケーション性が強いいため、より自然なアノテーションが作成される可能性があると考えられる。実際に、Ask ビデオは、ビデオコンテンツ数が 10000 を超え、大量のユーザが利用しているにも関わらず、作成されているシーンタグは非常に少量である。そのため、Synvie で獲得されるアノテーションからシーンタグの作成を行った。

オフラインビデオアノテーションとオンラインビデオアノテーションの比較を行うために、まずは、オフラインビデオアノテーションとして、専用ツールを開発し、それを用いてシーンタグの作成を行った。

次に、オンラインビデオアノテーションとして、Synvie から獲得される数種類のアノテーションから機械処理によってシーンタグを作成した。

しかし、機械処理のみで生成されたシーンタグの質には問題があると判断し、より質の高いタグを作成するために、生成されたシーンタグを基にさらにオンラインで人間の力を利用する仕組みを考えた。そのための仕組みとしてオンラインタ

グ選択システムを開発した。このシステムは、ビデオを視聴しながら、付与されているシーンタグがシーンに適切であるかどうかを視聴者が選択するシステムである。このシステムを利用した被験者実験を行うことで、新たなシーンタグの作成を行った。また、シーンタグを作成するだけでなく、今後より有効なシステムを開発するために、実験結果を様々な視点から分析し、考察を行った。

以上の専用ツールを用いた手法、オンラインビデオアノテーションから自動抽出する手法、オンラインタグシステムを用いた手法の3種類の手法によってシーンタグの作成を行った。さらに、そのタグをシーン検索に利用することでそれらの有用性の比較を行った。そのために、タグを利用した新しい発想のビデオシーン検索システムを開発した。この検索システムには、ビデオコンテンツに直接アクセスすることなくブラウザ上でシーンの情報が閲覧できるという、従来のビデオ検索システムにはない特徴がある。また、シーンタグの特性を活かしたインタフェースになっている。

このシーン検索システムに対して各手法によって作成されたシーンタグを利用し、それぞれの手法ごとの検索システムを用意し、シーン検索の被験者実験を行った。被験者に対して、検索を行うシーンの出題を画像と文章によって行った。被験者は検索によって回答シーンを決定し、それに費やされた時間は自動的に計測した。また、検索に費やされた時間以外に、検索クエリやビデオの視聴回数などもデータとして保存した。

タグ選択実験とシーン検索実験を通して、それぞれの手法によって作成されたシーンタグの作成コスト、シーン検索の精度、検索コストを比較することで、オンラインビデオアノテーションの有用性を検証した。

以下に本論文の構成を示す。第2章ではビデオアノテーションについてさらに詳細に述べる。第3章では、シーンタグ作成手法や、作成のために行った実験について述べる。第4章では、開発したシーン検索システムの詳細と、第3章で作成したシーンタグを利用したシーン検索実験について述べる。第5章では、関連研究について述べる。最後の第6章では、まとめと今後の課題について述べる。

第2章 ビデオアノテーション

ビデオコンテンツに対する内容情報記述をビデオアノテーションと呼ぶ。例えば、記述の規格としてXML [19] ベースのMPEG-7 [20] が存在する。

アノテーションはコンテンツの検索や要約などの応用のために作成されるものであるが、対象となるビデオコンテンツの種類や、応用などによって必要となるアノテーションが異なる。例えば、検索のためのアノテーションを作成する場合、ニュース映像とスポーツ映像では必要となる情報が異なる。ニュース映像はアナウンサーやナレーションの音声情報が最も重要であり、スポーツ映像では、実況や解説者の音声などに加え、選手の位置情報などの映像情報も非常に重要となる。ビデオコンテンツは非常に様々なものが存在し、ジャンル分けすることも困難であるコンテンツも存在する。また、最近では、一般人が容易に自身の作成したビデオコンテンツをWeb上で公開できるようになり、今後もそのような個人的なコンテンツがWeb上に氾濫していくと考えられる。それらのコンテンツは専門家が制作したコンテンツとは異なる特徴を持つ可能性も高いため、それに対応したアノテーションの作成手法も必要であると考えられる。

応用に関して述べると、ビデオシーン検索を実現する場合、検索可能な対象はアノテーションの質と量に依存する。ニュース映像では、アナウンサーの発話文や字幕、画面中のオブジェクトなどに関するアノテーションが必要であり、スポーツ映像では、選手の名前やチーム名、選手などの位置情報やプレイ内容、実況内容などが必要である。また、ビデオ要約を実現するには、検索のための情報に加え、さらに深い意味情報が必要であり、ビデオの任意の要素に関する意味的なアノテーションを必要とする。このほかの応用としては、コンテンツの推薦や、統合などが考えられ、いまだに考えられていないようなビデオコンテンツの応用も存在するであろう。このように、ビデオの利用方法によっても必要とされるアノテーションが異なる。

また、アノテーションに関する重要な要素に、作成にかかる人的コストという問題がある。あるビデオコンテンツに対して、十分に人手をかけて詳細なアノテーションを作成すれば、高度な検索や要約などの応用が比較的容易に実現できるであろう。しかし、応用が実現できればそれで良いのではなく、そのために費やされる人的コストも考慮しなければいけない。

つまり、アノテーションを作成する際には、コンテンツの種類、コンテンツの利用目的、その精度、人的コストといった複数の要素を考慮する必要がある。それ

らのバランスをうまくとり、応用を十分な精度で実現できるようなアノテーション作成手法が求められている。このようにビデオアノテーションに関する研究は非常に重要かつ可能性のある分野である。

これまでに、様々な手法によってビデオアノテーションを作成する研究が行われてきた。しかし、現在に至るまで、上記のバランスを十分に実用的なレベルで達成した手法は発見されていない。

ビデオアノテーション作成の手法は、大きく分けてオフラインビデオアノテーションとオンラインビデオアノテーションの2種類が存在する。この章では、それぞれの特徴と、代表的な研究例について述べる。また、本研究で利用したオンラインビデオアノテーションシステム Synvie について特に詳細に述べる。

2.1 オフラインビデオアノテーション

2.1.1 オフラインビデオアノテーションとは

オフラインビデオアノテーションは、まず画像処理や音声処理などの機械処理によってコンテンツの自動解析を行い、その後、機械処理によって抽出された情報の修正や、現状の技術では抽出が困難な情報を人手で付与するアノテーション作成手法である。

まずは、機械処理について述べる。

画像処理には、オブジェクトの検出や、カット検出、オプティカルフローの検出などがある。オブジェクトの検出とは、映像中の人物や、その顔の検出、テロップなどのテキスト情報の検出などのことである。またカット検出は、画像情報のRGB値の変化の大きさなどから、シーンの切れ目を検出する解析処理である。オプティカルフローからは、映像中のズームやパンなどの情報を得ることができる。

音声処理は、音声を認識してテキストとして書き出すという解析処理や、音の大きさの変化から情報を得るといった処理も存在する。

これらの機械処理は、一度その解析システムが完成してしまえば、あとはビデオコンテンツに対して解析システムを実行するだけでアノテーションが自動で生成されるため、アノテーション作成のコストが極めて小さいという非常に大きな利点がある。機械処理によってビデオの意味内容情報をすべて抽出することが可能であれば、人間の作業を必要としない。しかし、機械処理のみのアノテーションによって一般のビデオコンテンツの十分な精度の検索や要約などを可能にした例はない。

その理由として、まず、解析の精度の問題がある。画像処理や音声処理の解析の精度は、ビデオコンテンツに大きく依存する。ビデオの撮影環境や、機材などが機械処理の際に大きく影響するのである。例えば、室内のスタジオで収録されたビデオコンテンツなどは、解析の精度が高いかもしれないが、野外で撮影され

たようなものは、周囲の雑音を多く含み、特に音声認識の精度が低くなる。また、撮影する際の機材によっても、映像の質や、音声の質や大きさが変化し、それに伴い機械処理の精度も変化してしまう。また、専門家ではなく一般人の制作したビデオコンテンツに対しては機械処理の精度が低くなるという可能性が十分に考えられる。

画像処理や音声処理の研究は非常に古くから研究されているにも関わらず、現在でもこのような問題が残っているということを考えると、現状の技術からの革新的な精度の向上は非常に困難であると考えられる。

もう1つの理由としては、機械処理から抽出できる情報には現状の技術では限りがあるため、応用を実現するために十分なアノテーションの生成が困難であるという点である。機械処理で得られる情報は上で述べたような情報のみであり、ビデオの内部のすべての意味内容情報を十分に抽出ことはできない。

以上のことをまとめると、機械処理の利点は次の通りである。

- アノテーション作成にかかる人的コストが極めて小さい。

また、問題点は、次の通りである。

- 機械処理の精度が、ビデオコンテンツに依存する。
- 作成されるアノテーションが音声、テロップ、人、カット検出などの限られた情報である。

次に、機械処理による自動解析によって得られた情報を人手によって編集することでアノテーションを作成するという手法について述べる。この手法は、自動で抽出された情報を基にアノテーションを作成するので、半自動アノテーションということもできる。

この手法は、専用のツールを利用してアノテータと呼ばれる専門家によって、ビデオに対して詳細な意味情報を記述していくという手法である。アノテータは、機械処理によって抽出された情報を修正する。例えば、音声の誤認識によって生成された間違ったテキスト情報の修正や、それらのテキストとビデオの時間情報との対応の修正を行ったりする。また、機械処理では抽出できなかった情報を新たに追加するというを行う。

この手法は、機械と人間がそれぞれの得意な分野を手分けしてアノテーションを作成しようというアプローチである。また、人間の手によってアノテーションの作成を行うため、詳細かつ意味的な情報を付与することができ、コンテンツの応用に合わせたアノテーションの作成も可能である。

しかし、アノテーション作成にかかる人的コストが大きな問題点である。確かに、無限に詳細なアノテーションの作成が可能であるが、詳細にすればするほど、その人的コストが大きくなる。当然のことであるが、長時間のビデオコンテンツ

になると、その分コストが大きくなる。そのため、この手法では、コストとアノテーションの質、量のバランスをうまく考えなければならない。さらに、Web上に氾濫する大量のビデオコンテンツに対してこの手法を適用するということは非現実的であると言える。

もう1つの問題点として、詳細なアノテーションを作成するためには、アノテータがビデオの内容に関する知識を有している必要があるという点である。もし、コンテンツに関する深い知識を待たない人がアノテーションの作成を行った場合、客観的に読み取れる情報しか付与することができない。そのため、様々な種類のビデオコンテンツに対してそれぞれの専門家でない1人の人間によってアノテーションを作成する場合、この手法の利点が半減する。

以上のことをまとめると、人手によってアノテーションを作成する手法の利点は次の通りである。

- 詳細なアノテーションの作成が可能である。
- 応用に合わせたアノテーションの作成が可能である。

また、問題点は次の通りである。

- アノテーション作成にかかる人的コストが非常に大きく、詳細に作成すればするほど大きくなる。
- 詳細なアノテーションを作成する場合、ビデオの内容に関する専門知識が必要である。

2.1.2 Video Annotation Editor

オフラインアノテーションの研究例として、長尾らが開発した Video Annotation Editor [4] について述べる。

Video Annotation Editor は、MPEG コンテンツに対し XML 形式のアノテーションデータを付与するツールである (図 2.1)。音声認識や、カット検出・オブジェクトトラッキングなどを行うことができ、それらの機械処理から生成されるアノテーションを人間の手で編集することで、ビデオに対する詳細なアノテーションを作成することができる。

音声認識と画像認識をひとつの画面でひとつの時間軸を元にして処理できるのが特徴となっている。



図 2.1: Video Annotation Editor

2.2 オンラインビデオアノテーション

2.2.1 オンラインビデオアノテーションとは

オンラインビデオアノテーションとは、オンラインで不特定多数の人間によって行われるビデオコンテンツを中心としたコミュニケーション活動から得られる情報をビデオに対するアノテーションとして利用する手法である。

コミュニケーション活動とは、ビデオコンテンツに対して投稿されるコメントや、ビデオコンテンツを引用したブログなどのテキストコンテンツの作成と発信のことである。

それらのテキスト情報には、ビデオの内容に関する情報が含まれていると考えられ、また、テキスト情報とビデオコンテンツとのリンク付けが自動で行われるため、アノテーションとして利用することができる。

この手法は、ビデオアノテーションの手法としては最も新しい分野である。筆者の研究室では、任意のビデオシーンに対してコメントの投稿や、ブログへの引用ができるオンラインビデオアノテーションシステム Synvie [3] を開発した。本研究では Synvie の公開実験によって得られた情報をオンラインビデオアノテーションとして利用した。Synvie の詳細な内容については次節にて述べる。この他にオンラインビデオアノテーションを獲得するシステムとして SceneNavi [5] などが存在する。

この手法は、不特定多数の人間の知識を少しずつ顕在化し、それを大量に収集することで、大きな利用価値のあるアノテーションとして利用しようというアプローチである。

現状ではオンラインビデオアノテーションに関する研究例や実験例が少なく、ア

ノテーションの詳細な分析を行った例も見られない。この手法の特徴に関してはまだ確定したことを述べることはできないが、現状で考えられる特徴を述べる。

この手法の大きな利点は、インターネット上で行われる自然な活動から情報を獲得するため、その情報を発信している各個人は、アノテーションを作成しているという意識が特にないという点である。このことは、アノテーション作成ために個人が負担するコストが非常に小さいということである。システムを利用する人が増えれば増えるほど、アノテーションの量が増加するが、各個人が負担するコストは増加しない。

さらに、もう1つの利点としては、コンテンツの解析からは決して抽出できないような情報を獲得できる可能性があるという点である。例えば、ビデオを引用したブログには、そのビデオ自体の内容だけではなく、それに関連した情報も記述される可能性が高い。ドラマを引用したブログの例を挙げると、そのドラマの出演者が以前に出演したドラマや、テレビ番組などについて述べられる可能性がある。そのことで、引用されたドラマとその他のドラマやテレビ番組との間に関係付けを行うことができる。このことはブログだけでなくコメントに関しても同様の可能性がある。そのような情報はオフラインでアノテータの手によって付与することも可能であるかもしれないが、そのためには、ビデオの内容やそれに関連する情報について非常に詳細な知識が必要となり、1人の人間で付与することは困難であると考えられる。

しかし、問題点もいくつか考えられる。まず、この手法によって獲得できる情報のすべてがビデオコンテンツの応用に利用可能である可能性は非常に低いということである。ブログやコメントに含まれるすべてのテキストが、ビデオについて、またはビデオに関連する情報について語られたものであるとは限らず、全く関係のない内容も含まれるということが十分に考えられる。すべての情報をアノテーションとして付与した場合、おそらくアノテーションの質は低くなると考えられる。そのため、オンラインビデオアノテーションから得られる情報の中から、ビデオの応用に利用価値のある情報を抽出する手法が必要になると考えられる。

もう1つの問題点は、大量のアノテーションを獲得するためには、大人数の人間がシステムを利用する必要があるということである。各個人が発信する情報量は小さいため、大人数からでないと大量のアノテーションを獲得できない。少人数から得られる情報からアノテーションを作成しても、オンラインアノテーションの利点があまり発揮されず、利用価値が低いものになってしまうと考えられる。

多くの人間を集めるには、システムを自然に利用させる動機付けを与えなければならない。そのためには、まず、システム自体の使いやすさや、おもしろさ、話題性などが必要である。また、ビデオを共有するシステムである場合は、ビデオコンテンツ自体のおもしろさも非常に重要な要素であるが、著作権や肖像権の問題が絡んでくるため運用には注意が必要である。そのため、ビデオコンテンツ自体は共有せずに、ビデオを中心としたコミュニケーションを行うシステムによっ

てアノテーションを獲得するという手法も存在する。

この手法が用いられたサービスとしてニコニコ動画 [33] がある。このサービスでは、ビデオコンテンツ自体が投稿されるのではなく、YouTube などのビデオ共有サイトにおけるコンテンツの URL が投稿される。そのビデオを閲覧しながら、任意のシーンにコメントを投稿することができ、投稿されたコメントはビデオにオーバーレイした形で表示されることで、ユーザ間で共有される。そのため、このサービスで共有しているのは、ビデオコンテンツではなく、投稿されたコメントである。現在非常に話題性のあるサービスであり、すでに2万個以上のビデオコンテンツに対して300万件以上のコメントが投稿されている。しかし、ユーザの多くが2ちゃんねるのユーザであると考えられ、2ちゃんねるで頻繁に書き込まれるようなアスキーアートや表現によるコメントが多く、ビデオには関係のないコメントが大半を占める。そのため、それらのコメントがビデオコンテンツの応用に利用可能であるかどうかを判別するにはデータの分析が必要である。このサービスは、アノテーションの獲得や利用を目的としているわけではないと考えられるが、新しいオンラインビデオアノテーションの獲得手法であると言えるため、非常に興味深いサービスである。

以上のことをまとめると、オンラインビデオアノテーションの利点は次の通りである。

- 集団の力を利用してアノテーションを作成するため、個人の負担するコストが非常に小さい。
- 機械処理では解析できないような意味情報を含む可能性がある。

また、問題点は次の通りである。

- ビデオに全く関係のない情報をアノテーションとして付与してしまう可能性がある。
- 十分な量のアノテーションを獲得するためのユーザへの動機付けが必要である。

2.2.2 Synvie

Synvie は、マルチメディアコンテンツ配信に、コメントの投稿やブログへの引用などのオンラインアノテーションの仕組みを取り入れることによって、マルチメディアコンテンツを話題の中心としたコミュニケーションを支援するシステムである [3]。さらに、コンテンツ閲覧者による日常活動から、コンテンツに関する知識をアノテーションとして獲得・配信するシステムである。

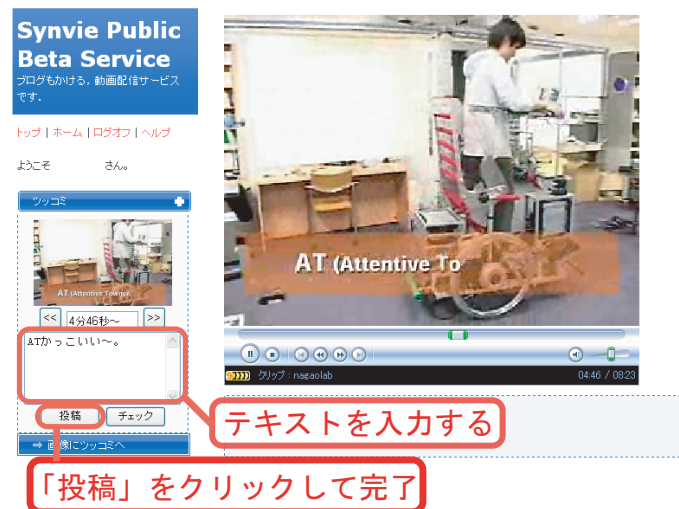


図 2.2: シーンテキストアノテーション

具体的には、Synvie は、ビデオの任意のシーンに対してコメントの投稿、ブログへの引用をすることができる。投稿されたコメントや、コンテンツを引用したブログからビデオに対するアノテーションを獲得することができる。Synvie で獲得できるビデオアノテーションは、具体的に次のものがある。

- シーンテキストアノテーション：
任意のタイムコードに対する自然言語によるコメント
- シーン画像テキストアノテーション：
任意のタイムコードの画像情報に対する任意の領域区間への自然言語によるコメント
- ブログ引用によるアノテーション（シーン引用アノテーション）：
任意のビデオシーンを引用したビデオブログ
- コンテンツテキストアノテーション：
ビデオ全体に対する自然言語によるコメント
- シーンボタンアノテーション：
「Nice」、「Boo」、「check」の3種類のボタンのクリックによるアノテーション

シーンテキストアノテーション、シーン画像テキストアノテーション、シーンボタンアノテーションは、一般にビデオの閲覧時に行われる。シーンテキストアノテーションの例を図 2.2 に示す。



図 2.3: ビデオシーンを引用したブログ

シーンボタンアノテーションの「Nice」と「Boo」のボタンは、シーンに対する印象をアノテーションするためのものであり、「check」ボタンは、ブログに引用したいと思ったシーンに対してクリックされる。

コンテンツテキストアノテーションは、ビデオ全体に対して投稿されるコメントであり、Synvie 以外の多くのビデオ共有サービスでも行われているものである。

ビデオをブログに貼り付けることで作成するビデオブログは多く存在するが、シーンに対する情報を抽出することが困難である。一方、Synvie で作成されるビデオブログはビデオシーンへのリンクをブログに組み込むため、そこからビデオシーンに対するアノテーションを獲得することができる。Synvie では、ビデオの閲覧時に行ったシーンテキストアノテーションやシーンボタンアノテーションの情報を基に、そのシーンを引用したビデオブログを作成するインターフェースが用意されている。Synvie を用いてビデオシーンを引用したブログの例を図 2.3 に示す。

アノテーションの獲得方法や、インターフェースの改善などについて、現在でも研究が進められている。具体的には、複数のビデオコンテンツを引用したビデオブログを作成するための仕組みや、新しいブログ作成インターフェースの開発などが行われている。また、コンテンツ制作者による自身のコンテンツに対するアノテーションを差別化することによって、より信頼性の高いアノテーションを獲得するための仕組みの開発が進められている。

第3章 シーンタグの作成

シーンタグの作成とは、ビデオの任意のタイムコードに対してキーワードを関連付けることである。つまり、キーワードをビデオコンテンツのタイムコードに対応させてデータベースに登録を行うことで、シーンタグを作成した。データベースには PostgreSQL ver8.1 を利用した。シーンタグとして作成されるキーワードは、名詞、動詞、形容詞であり、助詞や助動詞などは含まない。

Synvie に登録されたビデオコンテンツの内 27 個のビデオに対してシーンタグを作成した。ビデオコンテンツには、一般の Synvie 利用者が制作したものと、研究室内で制作されたものの両方が含まれる。利用したビデオのビデオ時間は平均で約 349 秒、最長で 768 秒、最短で 76 秒である。また、ビデオの種類には教育、物語、エンターテインメントなど様々なものが含まれる。タイトルの例を挙げると、「長尾研究室紹介ビデオ」や、「あなたの愛車拝見」、「天空の楽園アンコールワット」などがある。

次章のシーン検索実験によって、オンラインアノテーションの有用性を検証するために、オフラインビデオアノテーションとオンラインビデオアノテーションの両方の手法によってシーンタグの作成を行った。まずは、オフラインビデオアノテーションとして、専用のアノテーションツールを用いてシーンタグを作成した。次に、Synvie で得られたアノテーションからシーンタグの自動抽出を行った。さらにもう 1 つの手法として、自動抽出によって作成されたシーンタグを、オンラインで Web ブラウザ上で人間によって選別するシステムを開発し、それによってシーンタグを作成した。

3.1 専用ツールを用いたタグ付け

まず、ビデオを視聴しながら容易にシーンタグの作成を行うことができるツールを作成した（図 3.1）。

このツールは、まず、ビデオを視聴中にテキストボックスに自動で読み込まれる時間情報を基に、任意の開始時間、終了時間を指定することができる。そして、時間を指定したら、キーワードの入力フィールドに、「,」区切りで任意の数のキーワードを入力し、登録ボタンをクリックすることで、情報がデータベースに保存される。



図 3.1: 専用アノテーションツール

また、タグ付けの開始時にスタートボタンを、終了時にエンドボタンをクリックすることで、タグ付けに費やされた時間がデータベースに保存される。

この時点では、タグを保存しているテーブルの品詞と読みのカラムが空であるが、Cabocha [17] を利用することで、それらが自動で登録される。

時間情報の読み込みやタグの登録には、Ajax (Asynchronous JavaScript and XML) と呼ばれる技術を利用しているので、ページを再読み込みすることなく、スムーズにタグ付けを行うことができる。

このツールを利用して、対象としたすべてのビデオコンテンツに対して筆者がシーンタグの付与を行った。

アノテータである筆者は、ビデオの制作者ではなく、各ビデオの内容に関する詳細な知識を持たないため、映像や音声から客観的に得られる情報のみをシーンタグとして付与した。また、ビデオシーンに関連する情報の知識を保持していた場合も、それぞれのコンテンツに対して公平にシーンタグを付与するために、それらの知識は利用しなかった。基本的に、次の2つの観点で客観的に得られる情報に基づいて、シーンタグを付与した。

- イベント情報

例：ディスカッションマイニング、趣旨説明、夜

- 人や物体とその様子、テロップ、音声などのオブジェクト情報

例：議論札、コアラ、鳥、寝る、飛ぶ

これらの情報をできる限り詳細に、かつ網羅的に付与した。1コンテンツあたり約 1082 秒の時間区間分のシーンタグを付与した。

ここで付与されたシーンタグは、オフラインビデオアノテーションの一種と見なせる。

ここで、シーンタグの付与のために費やした時間を、シーンタグ作成のためのコストとする。

その時間は、1 コンテンツあたり約 1480 秒、最長で 3692 秒、最短で 582 秒であった。

3.2 オンラインビデオアノテーションからの自動抽出

オンラインビデオアノテーションとして、Synvie の一般公開実験によって得られたアノテーションデータを利用した。データは 2006 年 7 月 1 日から 11 月 30 日までに獲得されたものである。この期間に登録されたビデオコンテンツ数は 94 個、ユーザ数は 97 人である。

Synvie から得ることができるアノテーションデータの詳細については前章にて述べたが、そのうち、シーンテキストアノテーションとコンテンツテキストアノテーション、シーン画像テキストアノテーション、ブログ引用によるアノテーションの 4 種類のアノテーションをオンラインビデオアノテーションとして利用した。

それぞれのアノテーションデータからは、テキスト情報とそれに対応する時間情報を得ることができる。

アノテーションデータからシーンタグを作成するまでの処理手順は以下の通りである。

1. Cabocha を利用してテキスト情報の形態素解析を行う。
2. 事前に作成した不要語辞書を用いることで、形態素解析の際に生成されてしまうカナ 1 文字の語句や、「する」「なる」などの一般的な語を除去する。
3. 固有名詞、名詞、動詞、形容詞、未知語を抽出する。
4. 時間情報に対応させてデータベースに保存する。

形態素解析の際、Cabocha の辞書にはシステム辞書に登録されていない専門用語である「Web2.0」などの単語は、ユーザ辞書に登録することで、抽出されるようにした。

これらの処理はすべて機械処理によって一括で行うことができる。また、アノテーションデータは人間の自然なコミュニケーション活動から得られるものであるため、シーンタグ作成のために個人が負担するコストは無視できるレベルである。

この手法によって、27 個のビデオコンテンツに対して合計 4136 個、平均で約 153 個、最多で 516 個、最少で 12 個のシーンタグが作成され、1 コンテンツあたり約 1703 秒の時間区間分のシーンタグが作成された。

3.3 オンラインタグ選択システムを用いた抽出

Synvie によって得られるアノテーションテキストにはビデオに関連のない情報も含まれるということが容易に推測される。ブログやシーンテキストアノテーションに含まれるすべてのテキストがビデオの内容について述べているとは限らず、ビデオ自体には全く関係のないテキストが含まれると十分に考えられる。そのため、前節で作成されたシーンタグにはノイズが含まれている可能性が非常に高い。実際にそのシーンタグを人間の目で見ると明らかにタグとして不適切であるものが発見できた。例えば「これら」や「ところ」などの、どのようなシーンにもタグとしては利用できないであろうと考えられるようなものや、単語としては意味を持ってもシーンには対応していないと考えられるものが存在した。また、形態素解析の際に、細かく解析しすぎて意味を持たなくなってしまった名詞などが含まれていた。

そのため、生成されたシーンタグの質は決して高いものであるとは言えず、利用価値のあるタグと、ノイズであるタグとを選別する必要があると考えた。その選別が正しく行われることで、より質の高いシーンタグが作成される。

この選別を機械処理によってすべて行うことができるということが最も理想的であるため、いくつかの方法で、それを試みた。

まずは、TF-IDF(Term Frequency-Inverse Document Frequency)[18] の原理によって、シーンタグの重み付けを試みたが、この手法は大量のドキュメント数が必要とするため、有効な選別を行うことはできなかった。

次に、Google Web API [22] を利用し、コンテンツに付与されたタグと、シーンタグの共起関係から重み付けを行った。Google Web API を利用するとプログラム内から Google にクエリを送信し、その検索結果などを得ることができる。ニュース映像の音声認識から得られたキーワードを用いて Web 上のニュース記事を検索することで、そのニュースに関する情報を抽出する研究でも Google Web API が利用されている [14]。しかし、この手法からも有効な選別を行うことはできなかった。この理由として、コンテンツに付与されているタグには、コンテンツの内部シーンに関係するものが詳細に含まれているわけではないため、シーンに関係するキーワードよりも、一般的なキーワードとの共起関係が強くなってしまったからであると考えられる。



図 3.2: オンラインタグ選択システム

以上の考察より、シーンタグの選別を機械処理で行うことは非常に難しい問題であると考え、人間の力を利用するようなシステムを開発し、それを実現した。

そこで、前節で述べた仕組みで作成されたシーンタグがそのシーンに対するタグとして適切であるかを人手によってオンラインで Web ブラウザ上で選択することのできるシステムを開発した（図 3.2）。

人間によって選別されたタグの質は高いと推測されるが、ここで人的コストを多くかけてしまえば、オンラインビデオアノテーションを利用した効果が減少してしまう。そのため、なるべく個人の負担するコストが小さく、効果的にタグの選別を可能にするシステムを開発した。

このシステムは、オンラインで多数の人間が Web ブラウザを用いてタグの選別をするシステムである。

このタグ選択システムは、ビデオの視聴中に前節の仕組みで作成されたタグが付与されたタイムコードにくるとビデオが一時停止し、タグが適切であるかをユーザが選択するシステムである。

複雑な操作を必要としないため、誰でも簡単に利用方法が理解できるシステムである。

このシステムの利用手順は次の通りである。

1. ビデオコンテンツページを開く。
2. 自動でビデオが再生される。

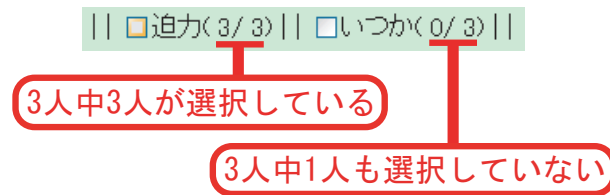


図 3.3: 履歴表示

3. タグの付与ポイントに来るとビデオが一時停止する。
4. 「このシーンに適切なタグを選択してください」というボタンをクリックする。
5. 適切であると判断したタグにチェックを入れて、「情報を登録」をクリックする。
6. 手順2に戻る（以後繰り返し）。

このシステムの特徴として、ビデオウィンドウが2つ用意されており、適切であるタグを選択する際（手順5）には、左のウィンドウに該当するシーンの先頭のサムネイルが表示され、右のウィンドウでは、該当するシーンの再生を自由に行うことができる。タグはタイムコードに付与されているものであるが、そのポイントのサムネイル情報ではなく、そこから始まる、もしくはその周辺のシーンに対する情報として利用するものである。そのため、タグが適切であるかどうかを選択する際にも周辺の情報も閲覧できる必要がある。そこで、2つのビデオウィンドウを用意し、周辺の映像も自由に閲覧できるようにした。

また、タイムラインバーが表示されており、シーンタグが付与されている箇所がハイライト表示される。タグ選択が完了している箇所の色はグレー、未完了の箇所はピンクでハイライト表示される。コンテンツに対してシーンタグが付与されている箇所を一目で見ることができ、また、すでにタグ選択が終了している箇所としていない箇所を見ることがもできる。

さらに、このシステムは1人の人間によって利用されるのではなく、オンラインで不特定多数の人間によって利用されることを前提としているため、タグ選択の履歴をタグとともに表示する（図 3.3）。履歴には、そのタグに対して、それが適切であるかどうかの選別をした人数と、そのうちの何人が適切であると判断したかが表示される。

また、内部処理として、次の2種類の情報がデータベースに保存される。

- 適切だと判断されたシーンタグの ID

- シーンの ID とタグを選択する際に費やされた時間（手順 4 から手順 5 が終了するまでの時間）、その間にビデオが再生された回数、ユーザ名

これらの情報を利用することで、シーンタグの選別を行い、また、そのために各個人が負担したコストを計算した。

このシステムを用いて被験者実験を行うことで、人間によって選別されたより質の高いシーンタグの作成を試みた。実験の詳細については次節で述べる。

3.4 タグ選択実験

3.4.1 実験内容

前節でも述べたが、本研究では 27 個のビデオコンテンツに対してシーンタグの作成を行ったため、この実験でも同じビデオコンテンツに対して被験者実験を行った。

27 個のビデオコンテンツのうち 18 コンテンツは各 3 人、残り 9 コンテンツは各 2 人の被験者によって、タグ選択実験を行った。被験者数は 11 人である。被験者の中にはコンテンツ制作者も含まれるが、自身で制作したコンテンツに対してはタグ選択を行わないようにした。

各コンテンツに対して、自由にシーンを選んで一部のタグ選択を行うのではなく、タグが付与されている全シーンのタグ選択を行うようにした。

3.4.2 実験結果

この実験の全体的な結果として、27 個のビデオコンテンツに対して合計 1493 個、平均で約 55 個のシーンタグが作成され、1 コンテンツあたり約 566 秒の時間区間分のシーンタグが作成された。タグ選択に費やされた総計時間は 1 コンテンツあたり約 656 秒であった。

以後、適切であると判断されたシーンタグを選択されたシーンタグと表現する。

各コンテンツに対して前節でオンラインビデオアノテーションから自動生成されたシーンタグのうち、選択されたシーンタグの割合は、平均すると 0.36 であり、その標本分散は 0.022 であった。

選択されたシーンタグの例として、「長尾研究室紹介ビデオ」という、研究室を紹介するビデオでは、“挨拶”、“哀愁”、“動く”、“映像”、“A T”などが挙げられ、「～天空の楽園～アンコールワット」というアンコールワットでの旅行の様子を撮影したビデオでは、“アンコールワット”、“エコー”、“音”、“階段”、“神”、“観光”などが挙げられる。

まず、選択されたシーンタグの数や割合を、いくつかの観点から比較した結果を示す。

前節で生成されたタグの数でコンテンツを分類した場合の、分類ごとの選択されたシーンタグの割合は表 3.1 のような結果となった。

表 3.1: 生成されたタグの数と選択されたタグの割合

3.2 で生成されたタグの数	～30	31～50	51～100	101～200	200～
コンテンツ数	3	5	4	7	8
選択された割合 (%)	60.3	36.2	34.3	36.8	34.7

コンテンツをその種類ごとに分類した場合の、分類ごとの選択されたシーンタグの割合は表 3.2 のような結果となった。ここで、“教育・研究”には、「長尾研究室紹介ビデオ」や、「知的情報共有プラットフォーム Soya」などの研究内容を紹介するビデオコンテンツが含まれる。“ノンフィクション”には、「最北歩き隊」、「天空の楽園アンコールワット」などの実写映像によるドキュメンタリや、自動車をナレーションとともに紹介する「あなたの愛車拝見」などが含まれる。“フィクション”には、3DCGで作られた物語「ノラネコピッピ」などが含まれる。“その他”には、「スノーボードムービー」などのストーリー性の少ないプロモーションビデオのようなビデオコンテンツが含まれる。

表 3.2: コンテンツの種類と選択されたタグの割合

コンテンツの種類	教育・研究	ノンフィクション	フィクション	その他
コンテンツ数	6	9	5	7
選択された割合 (%)	35.9	36.7	56.1	29.9

また、Synvieで獲得されるアノテーションの種類ごとにシーンタグを分類した場合の、分類ごとの選択されたシーンタグの数と割合は表 3.3 のような結果となった。

この実験は、18 コンテンツは各3人、9 コンテンツは各2人の被験者によってタグの選別を行った。そこで、各コンテンツに対する被験者数と、選択されたタグの数、タグの選択に費やされた時間の関係についての結果を示す。

まず、各コンテンツに対する被験者数と選択されたシーンタグの数は、表 3.4 のような結果となった。

また、選択されたシーンタグが、何人の被験者によって選択されたシーンタグであるかを示した結果が表 3.5 である。

表 3.3: アノテーションの種類と選択されたタグの数

種類	アノテーション数	生成されたタグ数	選択されたタグ数	割合 (%)
シーン ^{※1}	458	1028	443	43.1
画像 ^{※2}	135	455	272	59.8
ブログ	35 ^{※4}	2612	778	29.8
コンテンツ ^{※3}	15	41	0	0
合計	643	4136	1493	36.1

※1 シーンテキストアノテーション

※2 シーン画像テキストアノテーション

※3 コンテンツテキストアノテーション

※4 ビデオコンテンツを引用したブログエントリの数

表 3.4: 選択されたタグの数の変化

各コンテンツに対する被験者数	1	2	3
18 コンテンツ	558	689	779
9 コンテンツ	555	714	—
合計 (27 コンテンツ)	1113	1403	1493

次に、各コンテンツに対する被験者数の変化によるタグ選択に費やされた時間の変化についての結果を述べる。

全 27 コンテンツに対して、1 人目のタグ選択に費やされた時間は 1 コンテンツあたり 287.6 秒であり、2 人目は 238.4 秒であった。

3 人の被験者によって実験を行った 18 コンテンツに対しては、1 人目のタグ選択に費やされた時間は 1 コンテンツあたり 248.9 秒であり、2 人目は 179.0 秒、3 人目は 195.5 秒であった。

3.4.3 考察

この実験において、被験者によって選択されたシーンタグは、そのシーンに対して検索に用いるためのキーワードとして適切であると、人間によって判断されているので、1 人でも選択していればそれは信頼性のあるタグであると言える。そのため、選択された数が多いほど、オンラインビデオアノテーションには、多く

表 3.5: 選択した人数によって分類したシーンタグの数

選択した人数	1	2	3	合計
18 コンテンツ	222	194	363	779
9 コンテンツ	285	429	—	714
合計 (27 コンテンツ)	507	623	363	1403

の有用な情報が含まれていたことになり、また、その有用な情報を多く抽出できたということになるだろう。

実験結果から、各コンテンツに対して前節でオンラインビデオアノテーションから自動生成されたシーンタグのうち、平均すると 36.1 % が適切であるタグとして選択された。この値は、特に大きいとは言えないが、概ね妥当な値である。すなわち、有用であると人間が判断した情報が 4 割近く含まれていたことになり、オンラインビデオアノテーションには有用な情報が含まれているということを実証できたと思われる。しかし、約 3 分の 2 は、有用ではないという判断がされたという事実も大きい。そのため、やはりオンラインビデオアノテーションは、より適切な手法を用いて、より効率良く有用な情報を抽出しなければならない。

また、このタグ選択システムを利用したシーンタグ作成のためのコストを、タグ選択に費やされた時間として評価を行う。その総計時間は、1 コンテンツあたり約 656 秒であった。これは複数人によって行われたタグ選択時間の総計であるため、各個人が実際に負担しているコストはさらに小さい。もし、このコストを、各コンテンツに対して、最も時間のかかった人の時間とすると、その時間は、1 コンテンツあたり約 314 秒であり、前々節で専用ツールを用いたタグの作成に費やした平均時間の約 5 分の 1 であった。この 2 つの値の間の値がコストであると定義するべきであると考えられるが、適切な値の計算法を定義することができないため、シーンタグ作成のためのコストは 314 秒から 656 秒の間であるとした。結果として、前々節の専用ツールによるタグの作成に比べて、2 分の 1 から 5 分の 1 程度のコストでシーンタグを作成できたことになる。

次に、各コンテンツに対して選択されたシーンタグの割合から、いくつかの観点で考察を行った。選択されたシーンタグの割合が大きいほど、アノテーションに含まれる有用な情報の割合が大きい。

まず、全体的な結果として、コンテンツごとのシーンタグが選択された割合の標本分散は 0.022 と非常に低い値であったことから、コンテンツごとのばらつきが小さかったと言える。コンテンツごとの値を見ると、選択されたタグの割合が 25 % ~ 45 % のコンテンツが半数以上を占めた。

次に、コンテンツに付与されるアノテーションの量と、その中に含まれる有用な情報の割合の関係について考察を行った。

表 3.1 より、前節で生成されたシーンタグの数が 30 個以下であるコンテンツだけは、平均で 6 割近くのシーンタグが選択されているが、その他の分類の値は、すべて平均値の前後の値という結果であった。30 個以下であるコンテンツだけが値が大きくなった原因として、それらコンテンツに対するアノテーションの量は、全部合わせても非常に少ないため、有用な情報を含むアノテーションが少量でも大きな影響を与えたからであると考えられる。この結果から、前節で生成されたシーンタグの数は、選択されるシーンタグの割合にあまり影響を与えなかったということである。言い換えると、前節で生成されたシーンタグの数が多いコンテンツほど、選択されたシーンタグの数も多いということである。このことから、アノテーションの量が増えれば増えるほど、有用な情報も増えていくと考えられる。

次に、コンテンツの種類の影響について考察を行った。

表 3.2 より、“フィクション”以外の 3 種類のコンテンツに対するタグ選択に費やされた平均時間はどれも近い値を示している。また、“フィクション”の値が大きくなっている原因として、“フィクション”に分類されたコンテンツの中に、前節で生成されたアノテーションの数が 30 個以下であるコンテンツが含まれ、そのコンテンツに対する選択されたシーンタグの割合が高く、その 1 コンテンツが大きな影響を与えたからである。そのため、“フィクション”という分類に入るコンテンツ一般が有用なシーンタグが含まれている割合が高いというわけではないと考えられる。そのため、今回の実験では、選択されるシーンタグの割合がコンテンツの種類には、特に依存しなかったと言える。以上のことより、オンラインビデオアノテーションの中に有用な情報が含まれている割合は、コンテンツの種類にあまり影響を受けない可能性が高いと考えられる。オフラインビデオアノテーションでは、コンテンツによって、アノテーションを付与する手法を変える必要がある場合が多いということを考えると、この点は、オンラインビデオアノテーションの優位性の 1 つであると考えられる。

次に、アノテーションの種類について考察を行った。

表 3.3 より、選択されたタグの割合は、シーン画像テキストアノテーション、シーンテキストアノテーション、ブログ引用によるアノテーション、コンテンツテキストアノテーションの順で大きくなった。シーン画像テキストアノテーションは、6 割近くのシーンタグが適切であると判断されていることから、有用な情報を得るための有効なオンラインビデオアノテーションの手法であると考えられる。この割合が大きくなった理由として、シーン画像テキストアノテーションは、ビデオの任意のシーンに対する任意の矩形範囲を選択して、それに関連付けて投稿されるコメントであるため、画像中のオブジェクト情報などを直接述べることが多いという点が考えられる。しかし、シーン画像テキストアノテーションは、ユーザにとって比較的手間のかかる活動であるため、量が集まりにくいという問題点がある。今回の一般公開実験でもシーンテキストアノテーションの半分以上の量であった。

ブログへの引用によるアノテーションの選択されたシーンタグの割合は29.8%であり、シーンテキストアノテーションの43.1%よりも小さい値であった。この値だけを見ると、ブログへの引用によるアノテーションに含まれる有用な情報量は小さく思えるが、前節で生成されたタグの数ではなく、アノテーションの数を見ることで、ブログの有用性が確認できる。実験対象コンテンツを引用したブログの数は35個であり、その中から778個のシーンタグが選択されたので、1ブログあたり22.2個のシーンタグが作成されたことになる。一方、シーンテキストアノテーションのアノテーション数は458個であり、その中から443個のシーンタグが選択されたので、1アノテーションあたり0.97個しかシーンタグが作成されなかったということになる。このことから、1アノテーションあたりの情報量を考えた場合、ブログへの引用によるアノテーションには、有用な情報がシーンテキストアノテーションの20倍以上含まれていたことになる。そのため、他のアノテーションに比べて、特に有用な情報を抽出する仕組みが最も必要なアノテーションである。また、このアノテーションもシーン画像テキストアノテーション以上に収集することが困難なアノテーションであるという大きな問題点がある。そのため、アノテーションをより多く集める仕組みが必要である。それらの2つを実現する有効な仕組みが実現されれば、ブログへの引用によるアノテーションが、最も有効なオンラインビデオアノテーション獲得の手段であると考えられる。

コンテンツテキストアノテーションのシーンタグが全く選択されなかった理由として、コンテンツテキストアノテーションには、ナレーションに対して「いい声ですね。」などのコメントや、コンテンツ投稿者の「近日、後編載せます♪」などが含まれ、キーワードを取り出したときにコンテンツの情報を直接表現するようなものが得られなかったからだと考えられる。

次に、各コンテンツに対して複数人によってタグ選択を行うことの影響について考察する。

まず、選択されたタグの数と各コンテンツに対する被験者数の関係について述べる。表3.4より、各コンテンツに対する被験者数の増加とともに、選択されたシーンタグの数が増加していることがわかる。また、表3.5より、選択されたシーンタグのうち3分の1は、1人だけの被験者によって選択されたものであり、他の1人または2人の被験者には選択されていないものである。その理由として、まず、タグを見落とした、または、チェックをするのを忘れたということが考えられる。また、人の考え方の違いということも考えられる。つまり、ある人はあるタグがそのシーンを検索するために適切であるタグと考え、ある人は適切でないと考えたということもあり得るということである。どちらの理由であれ、人間によって選択されたタグは有用なものであると考え、1人だけの人間によってタグ選択を行うと、作成することが可能である有用なシーンタグが作成されないという可能性があると考えられる。そのため、タグの選別は複数人で行う必要があると言え、また、複数人で行うことで、様々な人の思考をシーンタグに反映することができ

と言える。

次に、被験者数の増加と、各被験者が負担するコストの関係について述べる。その関係について、各コンテンツに対する被験者の順番とタグ選択に費やされた時間から考察することができる。実験結果から、1人目のタグ選択に費やされた時間に比べ、2人目のタグ選択に費やされた時間は1コンテンツあたり50秒近く短縮されていた。この結果は、履歴表示（図3.3）によって前の被験者がどのタグを選択したかを閲覧しながらタグ選択を行うことができるため、一覧の中から適切なタグを素早く発見することができたと考えられる。2人目から3人目に関しては、費やされた時間に変化があまり見られなかった。このことから、タグ選択を複数人で行う場合、1人目のユーザに最も負担がかかるが、その他のユーザの負担は小さくなると言える。しかし、人数が増加すればするほどその負担が小さくなっていくわけではないと考えられる。

この実験によって、比較的低コストで、オンラインビデオアノテーションから自動抽出されたシーンタグの選別を人間によって行うことができた。それらのシーンタグの有用性については、シーン検索実験にて検証する。

この実験で明らかになった問題点の1つとして、形態素解析の際に細かく分割しすぎた語の取扱いの問題がある。例えば、アノテーションテキスト中の「研究科」という単語は「研究」＋「科」に分割されてしまうため、結局タグとして利用できなくなってしまった。この例は、「・・化」や、「・・語」、「・・性」などの単語にも共通の現象が起きてしまった。また、このことは上記のような場合以外にも起きる現象であり、最も顕著な例は「ブリティッシュパブ」という単語が、「ブリ」＋「ティッシュ」＋「パブ」に分割されたり、「エントランスマット」という単語が、「エン」＋「トランス」＋「マット」と分割されたりと、全く違う意味のものになってしまった。名詞の抽出の際に、連続する名詞列を結合するなどのいくつかの手法によって解決を試みたが、どの手法でも、それによって他のタグへ悪影響が出るということがあり、結局解決をすることができなかった。そのため、この問題の解決にも人間の力が必要であると考えられる。その具体例として、タグ選択システムに、タグを結合して登録する機能や、文節を表示して、その中からタグを取り出して選ぶような機能を実装するということが考えられる。

この他に、ビデオシーンを直接的に表現するような情報と、そのシーンに対する関連情報の区別についても考える必要がある。例えば、あるシーンに対する登場人物の名前と、その人物の出身地などの関連情報の区別である。これらの情報は、どちらも価値のある情報であると考えられるが、利用可能な応用も異なる可能性もある。また、関連情報が獲得できるということはオンラインビデオアノテーションの優位性でもあるため、それを利用するべきである。そのため、これらを効率よく区別して保存する仕組みも必要であると考えられる。しかし、システムの機能などが増えることで、システムの利用者の負担が増えるということも考慮しなければならない。そのため、そのバランスをうまくとるような仕組みを考え

る必要があり、この点も今後の課題の1つである。

その他の今後の課題として、タグ選択システムに、機械処理の情報も組み合わせることが考えられる。例えば、カット検出によって定義したシーン区間に対してタグを選択する仕組みが考えられる。それによって、有用であるタグの数が増える可能性がある。

今回の実験に基づいて、システムの改善点や、インタフェースの利便性など、今後の課題となる問題が多く発見された。これらの問題点を考慮し、より低コストで質の高いビデオアノテーションを作成するためのシステムの実用化に向けて今後も研究を進めていく予定である。

第4章 ビデオシーン検索

本研究では、タグを利用した新しい発想のビデオシーン検索システムを開発した。そして、そのシーン検索システムに前章の各手法で作成された3種類のシーンタグを適用し、被験者実験を行った。

4.1 タグを利用したビデオシーン検索システム

シーンタグの特徴を考慮し、それを有効に活かしたシーン検索システムを開発した。

オンラインビデオアノテーションを基に作成されるシーンタグには、コンテンツに対する網羅性が低いという大きな問題点がある。この問題は、アノテーションの量にも依存するが、コンテンツ全体を網羅するほどの量のシーンタグを作成することは非常に困難であると考えられる。

また、タグの本質的な特徴として、タグの付与されていない箇所は、検索にヒットしないという問題がある。しかし、付与されているシーンタグを小さなスペースに大量表示することが可能であり、検索の手助けとなるという利点もある。

以上のような特徴から、検索クエリに対して、各シーンをランク付けして表示するのではなく、まずコンテンツをランキング表示し、同時に付加情報を提示することで、その情報を基にシーン検索を実現するシステムを開発した。

ここで、付加情報として次のような情報を提示する必要がある。

- 各シーンの情報
- タグが付与されているタイムコード
- タグが付与されていない箇所の情報
- 付与されているシーンタグの一覧

これらの情報を Web ブラウザ上で効率よく提示するためのインタフェースを実現した。

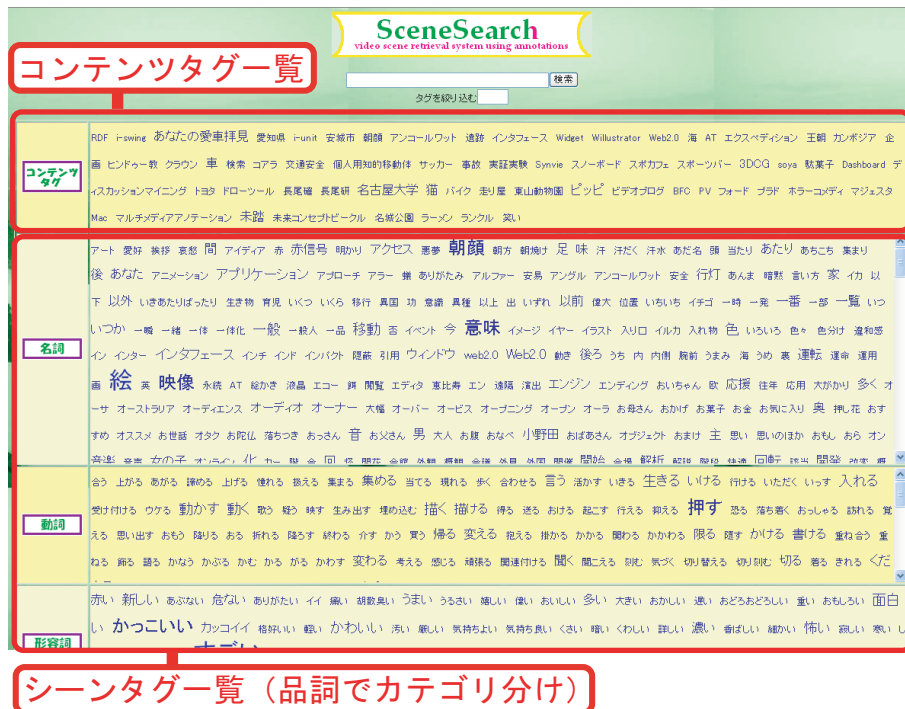


図 4.1: シーン検索システムトップページ

4.1.1 シーン検索インタフェース

図 4.1 がこのシーン検索システムのトップページである。

トップページには、コンテンツタグ、シーンタグのすべてが表示される。また、シーンタグは名詞、動詞、形容詞にカテゴリ分けされている。未知語は名詞として扱われる。さらに、それぞれのカテゴリの中では、タグが50音順にソートされて表示される。それぞれのシーンタグをクリックすることで、検索クエリ用のテキストフィールドに追加される。そのため、キーボードを使ったテキスト入力することなくタグを利用することができる。

品詞にカテゴリ分けし、50音順にソートすることで、利用したいタグを発見するための手助けになるが、それでも、大量のタグの中から必要とするタグを探すためには手間がかかる。本実験では検索対象のコンテンツが比較的少量のため、上の情報のみでも検索が不可能ではない。しかし、このシステムを大量のビデオコンテンツに対して利用した場合、付与されているタグは膨大な量になり、もはや、タグを利用することの利点が半減してしまうと考えられる。

そこで、タグをインクリメンタル検索できる仕組みを実装することで、この問題の解決を試みた。インクリメンタル検索とは、絞り込み検索とも言われ、検索キーワードを1文字入力するたびに、徐々に検索結果を絞り込んでいく検索方法

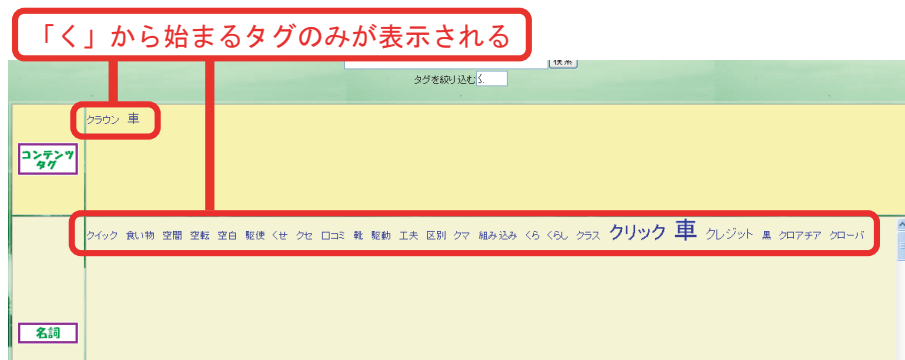


図 4.2: インクリメンタル検索

のことである。インクリメンタル検索を利用している例では Google Suggest [23] がある。

この検索システムでは、タグを絞り込むためのテキストフィールドにひらがなで文字を入力することで、その文字から始まるタグのみが表示され、そのほかのタグは見えない状態になる。例えば、「車」というタグが付与されているコンテンツが存在するかどうかを調べたい場合に、「く」と入力すると、読みが「く」から始まるタグのみが表示される (図 4.2)。それでもタグの候補が多すぎる場合は続けて「る」と入力することで、読みが「くる」から始まるタグのみが表示されるようになる。このような手順でタグの絞り込み検索を行い、もし検索クエリとして利用したいタグを発見した場合は、そのタグをクリックすることで、検索クエリのテキストフィールドにタグが追加され、タグを絞り込むためのテキストフィールドはリセットされる。この検索システムは付与されていないタグで検索してもコンテンツがヒットしない。そのため、存在しないタグで検索するという行為は無駄な行為になってしまう。それを防止して検索コストを少しでも下げるための効果もある。また、あらかじめ読み込んでおいたタグを JavaScript によってインクリメンタル検索するため、サーバへのアクセスを必要としない。そのため、高速でインクリメンタル検索を行うことができる。

これらの機能を実装することで、大量のタグの中から検索クエリとして利用したいタグを効率よく発見することができる。

次に、クエリに対する検索結果について述べる。この検索システムでは、検索クエリに基づき、コンテンツのランク付けを行い、その結果を表示する。ただし、検索にヒットしたコンテンツに対して、次の情報を同時に表示する。

- シーンのサムネイル画像
- タイムラインシークバー
- シーンタグの一覧

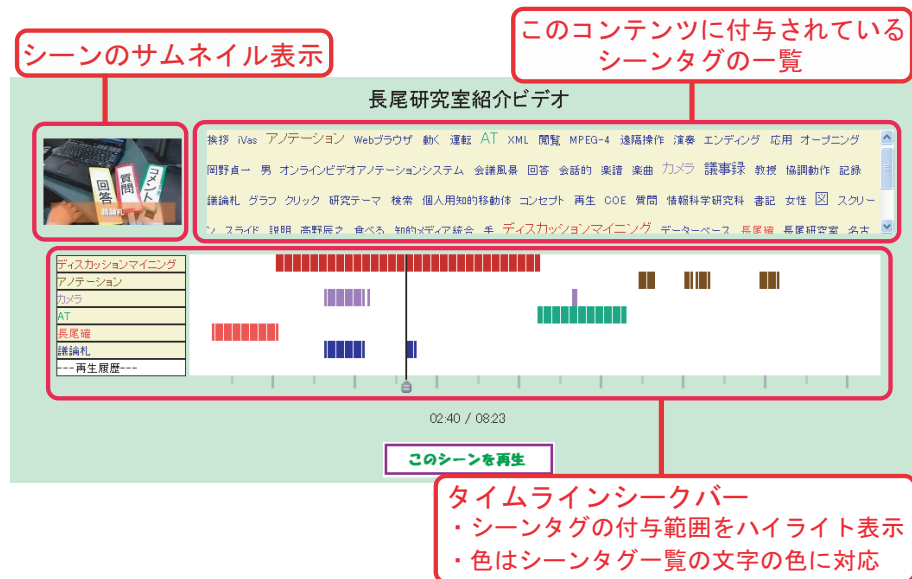


図 4.3: 検索結果ページ

検索結果ページの例を図 4.3 に示す。

検索にヒットした各コンテンツの左上部分には、シーンのサムネイル画像がシークバーのタイムコードに対応して表示される。Synvie には、各コンテンツに対して、ビデオだけでなく 2 秒間ごとのサムネイル画像が登録されているため、それを利用した。

また、中央部分に表示されるタイムラインシークバーには、検索クエリとして用いられたシーntagが付与されている箇所がハイライト表示される。シークバーは自由に操作することが可能であり、また、ハイライト部分をクリックすることで、シークバーがそのポイントにジャンプするという機能も有している。これらの機能によって、ビデオに直接アクセスすることなくシーンの画像情報を閲覧することが可能であり、シーntagが付与されている箇所、されていない箇所ともに容易に閲覧することができる。また、オンラインビデオアノテーションを基に作成されたシーntagを利用する場合は、アノテーション情報として、ブログに引用されたシーンや、シーンテキストアノテーションが行われたシーンをハイライト表示させた (図 4.4)。図 4.4 にある“ツッコミ”とはシーンテキストアノテーションのことである。

サムネイル画像の右部分には、そのコンテンツに付与されているすべてのシーntagの一覧が表示される。クエリとして用いられているシーntagの色は他のものと区別され、その色はシークバー上のハイライト表示の色に対応する。また、一覧表示内のシーntagをクリックすると、そのタグが検索クエリに追加され、そ

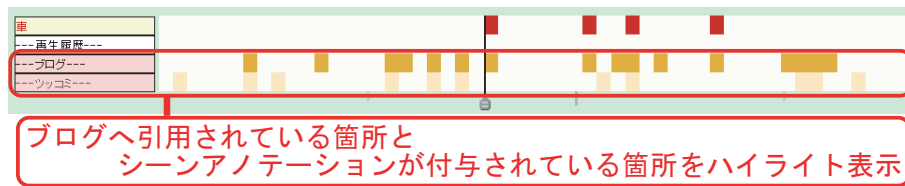


図 4.4: アノテーション情報の表示

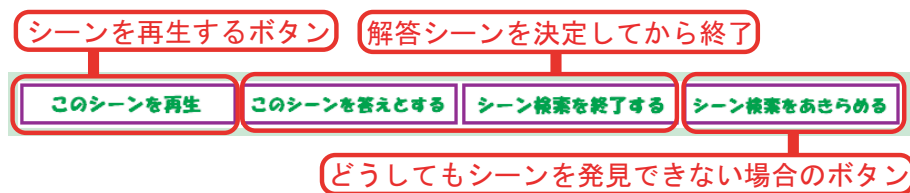


図 4.5: 実験用ボタン

のコンテンツのみを対象とした再検索が行われる。この操作によって、1クリックでシークバーへのシーンタグの追加が可能になり、また、1つのコンテンツに注目してシーンの絞込みを行うことができる。

また、「このシーンを再生」というボタンをクリックすることで、シークバーで選択したビデオの任意の時間からのビデオの再生が可能であるので、シークバーやサムネイル情報をもとに、任意のビデオシーンの視聴ができる。シーン検索実験では、被験者に出題に対する解答シーンを決定させるために、図 4.5 のボタンを追加した。ビデオシーンの視聴は Synvie のインターフェースを用いて行う (図 4.6)。

4.1.2 検索アルゴリズム

このシステムの検索は、タグによって検索が行われるということが前提となっているため、検索クエリと一致するタグが付与されているコンテンツが検索結果に加えられる。

ただし、検索対象であるタグはシーンタグだけではなく、コンテンツタグも存在する。コンテンツタグとは、コンテンツ全体に対するタグであり、直接入力されたものと自動生成されたものの2種類に分類される。直接入力されたコンテンツタグとは、コンテンツ投稿時に投稿者によって付与されるタグである。自動生成されたコンテンツタグとは、コンテンツのタイトル、サブタイトル、投稿時のコメントを形態素解析することで抽出される名詞（未知語を含む）である。

これらのタグと検索クエリとの一致に基づいてコンテンツが重み付けされる。重み付けの基準として、シーンタグ、直接入力されたコンテンツタグ、自動生成され

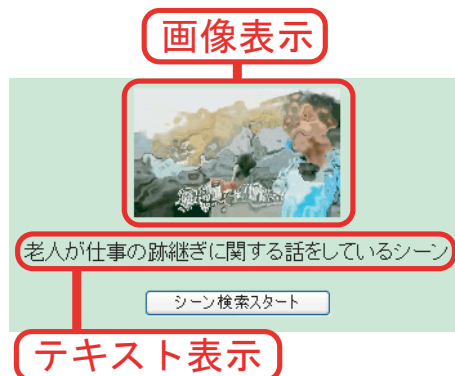


図 4.7: シーン検索実験出題例

4.2 シーン検索実験

開発したシーン検索システムに前章の各手法で作成された3種類のシーンタグを利用し、それぞれの手法ごとの検索システムを用意し、被験者実験を行った。

4.2.1 実験内容

検索対象として9シーンを設定し、被験者への出題を行った。シーンの出題文は「ある動物が親子で映っているシーン」や「スノーボードをしている人がコースの端に突っ込む前の、滑っているシーン」などであり、必ずしもシーンタグとして付与されている語句が出題文に含まれるのではなく、付与されているシーンタグをヒントにしてシーンを推定できるような出題文とした。また、出題に対する答えが唯一の時間区間であることを明確にするために、文章だけでなくそのシーン中のサムネイル画像をぼかした画像も提示した。図4.7に出題例を示す。

被験者は各出題に対する解答シーンの検索を行い、解答までに費やした時間を自動的に計測した。

被験者の数は9人である。各被験者は、各手法で作成されたタグを利用した検索を手法ごとに3シーンずつ計9シーンの検索を行った。その組み合わせによって、各シーンは、各手法で作成されたタグを利用して各3人ずつ計9人によって検索が行われる。そのため、タグの各作成手法を公平に比較することができる。被験者には、各シーンに対して、自身がどの手法によって作成されたタグを利用して検索を行うのかについては知らせずに実験を行った。

そのために、シーン検索には前節で開発したシステムを利用したが、シーン検索に費やされる時間の計測や、利用しているタグがどの手法で作成されたものであるかを被験者に隠すための、実験用のページを利用した。

実験用トップページには、検索が終了していないシーン番号とラジオボタンが表示される。

各被験者の実験の手順は次の通りである。

1. 検索を始めるシーン番号を選択し、「検索するシーンの説明を表示する」というボタンをクリックする。
2. 検索するシーンの説明が、画像と文章で表示される（図 4.7）。
3. 「検索を開始する」というボタンをクリックする。
4. シーン検索システムのページが別ウインドウで表示される（検索の開始）。
5. 解答とするシーンを決定する（自動で検索システムのウインドウが閉じる）。

手順3から手順5までの時間を機械的に自動で計測することで、シーン検索に費やされた時間の計測を行った。

また、手順4で開かれるシーン検索システムのページは、タグの作成手法ごとに3種類存在する。そのためには、「検索を開始する」というボタンに埋め込まれたリンクのURLを変化させる必要がある。そのURLは、ユーザ名とシーン番号の対応によって自動で決定される。それを決定するための情報はあらかじめデータベースに保存しておいた。これにより、被験者は利用しているタグの作成手法に関して特に意識することなくシーン検索を行うことができる。

4.2.2 実験結果

この実験によって、各出題に対して以下のデータを取得した。

- 解答として決定されたシーン
- 検索に費やされた時間
- 検索クエリ
- 視聴されたシーン

以上の各データをまとめた結果についてこれから述べる。

まず、9個の出題に対する9人の被験者の解答シーン、合計81個の解答シーンは、すべて各出題に対する適切なシーンと一致した。

次に、検索に費やされた時間を、シーンタグ作成手法ごとにまとめた結果を表4.1に示す。

表 4.1: タグ作成手法と平均検索時間

タグ作成手法	A:平均検索時間 (秒)	B:第 1クエリ決定時間 ^{※1} (秒)	A-B
オフラインでタグ付け	118.1	64.1	54.0
自動抽出	169.6	97.3	72.3
オンラインでタグ選択	145.4	90.1	55.3

※1 検索を開始してから最初のクエリを送信するまでの時間

各出題に対して、各タグ作成手法ごとの平均検索時間を比較した場合、その時間が最短であった回数（検索問題数）は、専用ツールで付与したタグ、タグ選択実験で作成されたタグ、自動生成されたタグの順で多く、その数は、5回、3回、1回であった。

検索クエリとして送信されたクエリの総数と、その中に含まれるキーワード数を表 4.2 に示す。

表 4.2: タグ作成手法と検索に用いられたクエリ

タグ作成手法	クエリ送信数	重複含まないクエリ数
オフラインでタグ付け	132	40
自動抽出	202	59
オンラインでタグ選択	156	64

検索に用いられたクエリには、シーンタグとコンテンツタグ、そのどちらでもないキーワードが含まれる。それぞれの数を表 4.3 に示す。

表 4.3: クエリの分類

タグ作成手法	シーンタグ (個)	コンテンツタグ (個)	その他 (個)
オフラインでタグ付け	34	2	4
自動抽出	49	4	6
オンラインでタグ選択	47	5	12

検索の途中におけるシーンの視聴回数についての結果を表 4.4 に示す。

表 4.4: ビデオの視聴回数

タグ作成手法	ビデオ視聴回数	解答と一致しない シーンの視聴回数
オフラインでタグ付け	33	2
自動抽出	34	5
オンラインでタグ選択	30	3

4.2.3 考察

この実験から、各手法によって作成されたシーンタグを利用した検索のそれぞれの検索精度と検索コストを推定することができる。検索精度とは、検索対象となるシーンを発見できるかどうかであり、検索コストとは、シーン検索に費やされた時間や手間などである。また、前章で、それぞれのシーンタグ作成のコストを計測しているので、それらを組み合わせることで、検索精度と検索コストの2つの観点で各タグ作成手法を比較をする。

まず、検索精度の比較を行う。しかし、今回の実験では、すべての被験者がすべての出題に対して正しいシーンを発見することができたため、この観点ではそれぞれのシーンタグの比較をすることはできなかった。

次に、検索コストの比較を行う。

表 4.1 より、専用ツールによって付与したタグ、タグ選択システムによって作成したタグ、オンラインアノテーションから自動抽出したタグの順で短時間でのシーン検索が可能であった。また、それぞれの差は、27.3 秒、24.2 秒であり、ほぼ等間隔で検索に費やされた時間に差があった。また、検索を開始してから最初のクエリが送信されるまでの時間も、同じ順で短時間で行われた。最初のクエリの決定までの時間が短いほど、検索に利用したいタグを容易に発見できたことになる。表の“A-B”の欄は、平均検索時間から、最初のクエリが送信されるまでの時間を引いた値であるため、コンテンツの検索結果から対象のシーンを発見するまでに費やされた時間である可能性が高いと考えられる。この時間には、検索結果ページで、シークバーのスライダーを使ってシーンを探したり、シーンタグによってシーンの絞込みを行っていたりしていたと考えられる。よって、この時間が短いほど、シーン検索までの手間が少なかったと考えられるため、この時間も検索コストを推定するための1つの要素であると言える。その時間は、オンラインアノテーションから自動抽出したタグの場合が最も長く、他の2つの手法はほぼ変わらず、それらには大きな差が見られた。

次に、検索に用いられたクエリに関して述べる。クエリの送信数が少ないほど、検索コストが小さいと言える。表 4.2 より、クエリの送信数は、オンラインアノ

テーションから自動抽出したタグの場合が最も多く、その差は大きかった。また、クエリ中の重複を含まないクエリ数は、専用ツールによって付与したタグの場合だけが、特に少ないという結果になった。クエリの種類に関して、用いられたシーンタグや、コンテンツタグの数が、自動抽出したタグの場合と、タグ選択システムによって作成したタグの場合でほぼ同じになった理由として、タグ選択システムによって作成したタグは自動抽出したタグの部分集合であるからであると考えられる。この表の“その他”というのは、タグとして付与されていないキーワードである。その例として、自動抽出によるタグと、タグ選択によるタグによる検索の両方で、「老人が仕事の跡継ぎに関する話をしているシーン」を検索する際に「老人」や「おじいさん」というキーワードが送信されていた。実際には、それらのタグは付与されておらず、「おばあさん」というタグによってシーンが検索された。タグのみによって検索を行うことを前提に開発した検索システムであったにも関わらずこのようなことが起きたということは、システムのインタフェースに問題があったからであると考えられる。

次に、検索の過程におけるビデオ視聴に関して述べる。表4.4より、ビデオの視聴回数に大きな差はなく、また、解答と一致しないシーンの視聴回数は、どの手法に関しても少なすぎるため、この結果からはそれぞれの手法を比較できるような結果を得ることはできなかった。この要因として、検索対象としたシーンが出題として難易度が低かったためであると考えられる。

以上の結果から、検索コストは、専用ツールによって付与したタグ、タグ選択システムによって作成したタグ、オンラインアノテーションから自動抽出したタグの順で低いと考えられる。また、その中でもオンラインアノテーションから自動抽出したタグを利用した検索コストは特に高く、やはり、タグ作成のコストを払ってでも検索コストを下げる必要がある。そのため、専用ツールによって付与したタグとタグ選択システムによって作成したタグのコストパフォーマンス、すなわち検索コストの自動生成されたタグとの差とタグ作成コストの比率の比較を行った。

ここで、次の式によってタグのコストパフォーマンスを計算して比較を行った。

$$\text{タグのコストパフォーマンス} = \frac{(\text{自動抽出による平均検索時間}) - (\text{平均検索時間})}{(\text{シーンタグの作成に費やされた時間})} \quad (4.1)$$

この式から、タグの作成のコスト1秒あたり、どれだけ検索時間の短縮がされたかを計算することができる。タグ選択システムによるシーンタグの作成に費やされた時間は、全総計時間である656秒から各コンテンツに対する個人の最大時間の総計である314秒とした。その結果を表4.5に示す。

表4.5から、この実験における、シーン検索のコストパフォーマンスは、タグ選択システムによって作成したタグが最も高いという結果となった。よって、この観点で見た場合、タグ選択システムを用いたシーンタグの作成手法が最も優れていたと考えられる。

表 4.5: タグのコストパフォーマンス

タグ作成手法	コストパフォーマンス
オフラインでタグ付け	0.0348
オンラインでタグ選択	0.0369～0.0771

この実験の問題点として、被験者数と出題数が挙げられる。この実験では、被験者数は9人であり、出題数は9問ずつであるため、各シーンタグの作成手法に対して、27個のデータが獲得されたことになる。この値は、各手法を比較するために必要最低限の量ではあるが、十分な量であるとはとても言えない。被験者数も出題数も多ければ多いほど、検索精度や検索コストの比較のために、より正確なデータを獲得できると考えられる。この実験では、比較のための要素となり得る検索精度やビデオの視聴回数などのデータがどの手法でもほぼ等しい値が出てしまったため、これらの観点での比較をすることができなかった。今後の実験では、より多くの被験者によって、より多くのシーンを検索する必要があると考えられる。また、システムを一般公開することで、特別な被験者実験をすることもなく、より自然なデータを獲得することも考えている。

また、検索システムの改善点がいくつか挙げられる。まず、内部的な処理の問題がある。前節でも述べたように、検索結果のランク付けのアルゴリズムは深い考察に基づいて設計されたものではない。今後、より多くのビデオコンテンツを検索対象とする場合には、この点は必ず改善しなければいけない点である。また、現在は、検索クエリとタグの全文一致で検索されるが、部分一致でも検索できる必要がある。次に、インタフェースの改善点として、必ずタグをクエリとして用いるようなインタフェースにする必要があり、そのためにはインクリメンタル検索の改良が考えられる。さらに、より多くのシーン情報をビデオコンテンツそのものにアクセスすることなく閲覧できるような仕組みも課題の一つとして挙げられる。

第5章 関連研究

5.1 オンラインビデオアノテーションに関する研究

5.1.1 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

オンラインビデオアノテーションを獲得するためのシステムとして、映像シーン連動型掲示板コミュニケーションシステム SceneNavi がある [5]。SceneNavi は、Web 上に存在している動画を閲覧しているコミュニティの間でコミュニケーションを行うシステムである。映像の視聴に同期した掲示板型のコミュニケーションや、非同期での掲示板型のコミュニケーションを行うことができる。特徴は、映像を、映像アーカイビングシステム SceneCabinet [7] を用いることで、まとまった時間区間であるシーンへ分割しておき、各シーンに対してコミュニケーション空間を作り出すという点である。SceneCabinet は、映像からカット、テロップ、カメラの動き、音楽、人の声といった豊富なインデックスを、映像処理技術により高い精度で検出することができる。そのため、シーン単位でコミュニケーションが行われる。コミュニケーションの過程において、動画上の任意の時間に閲覧者のコメントが関連付けられる。また、実在する俳優のファンによるコミュニケーションサイトを利用して実験を行うことで、SceneNavi で行われるコミュニケーションに関する分析も行われている [6]。コミュニティの活性化や、アノテーションを利用した検索などの応用について研究が進められている。シーンをあらかじめ分割しておき、そこに情報を付与するという点が、本研究の手法との相違点である。シーンを自動解析によって分割する際に、意味的な情報が考慮されているかどうかは問題点であるが、今後の研究を進める上で参考となる研究である。

5.1.2 番組実況チャットに基づく放送番組のビュー生成

日本最大の電子掲示板である 2ちゃんねる [27] では、テレビ番組の放送中にリアルタイムで番組について書き込みを行うという行為が絶えることなく行われている。その中から得られる情報から、番組コンテンツのインデキシングをしたり、盛り上がりを検出したりすることで、様々な視点で視聴可能なビューを生成する研究が宮森らによって行われている [10]。具体的には、2ちゃんねるで頻繁に書き込まれるアスキーアートや、盛り上がりや落胆を表すような単語をパターンマッ

チングによって検出する。開発されたシステムによって、シーンに対する視聴者の反応の大きさや、盛り上がり、落胆の度合いなどを利用した様々な視点による番組視聴を行うことができる。また、書き込みを行った人のID情報を利用することで、特定の視聴者の視点による番組の視聴を行うこともできる。映像コンテンツと電子掲示板は全く異なる閲覧ツール上に存在するが、電子掲示板に映像に対する記述が行われており、また、それが応用に利用可能であるため、この情報もオンラインビデオアノテーションであると言える。すなわち、この研究で行われている様々な視点によるビューの生成は、オンラインビデオアノテーションを利用して実現した応用の1つであると言える。この研究では述べられていないが、電子掲示板に書き込まれる情報を分析することで、シーン検索など、その他の応用にも利用可能である可能性が十分にある。また、大量のアノテーションが集まりやすいという点も大きな特徴であると言えるため、注目すべき研究であると考えられる。

5.1.3 ネットワーク簡易動画編集アプリケーション

オンラインでWebブラウザを用いて動画編集を行うことができるシステムが中村らによって開発されている[8]。このシステムは、ビデオ共有サイトにおいて、投稿したビデオコンテンツの編集をしたいというニーズがあるという考えに基づいて開発された。このシステムは、コンテンツファイルそのものを直接編集するのではなく、同期マルチメディア記述言語SMIL (Synchronized Multimedia Integration Language) [21] で書かれた記述ファイルを生成することで、動画の編集を行う仕組みである。ユーザはWebブラウザ上で編集を行い、編集終了後にはサーバ側でSMILによる記述ファイルが自動生成される。そして、コンテンツファイルとSMIL対応プレーヤーによって、編集されたビデオコンテンツを閲覧することができる。オンラインで編集作業を行っているが、編集作業は1人で行うシステムであるため、作業そのものはオフラインビデオアノテーションに近いと言える。ただし、この研究では述べられてはいないが、コンテンツそのものとは別のファイルを作成し、コンテンツの変換を可能にしているため、その編集履歴はオンラインビデオアノテーションとして利用可能である可能性がある。コメントの投稿やブログへの引用とは異なるオンラインビデオアノテーションの獲得手法であると考えられるため、新たなアノテーション獲得手法を考案する上での参考にしたい。

5.1.4 分散環境下における複数人による動画編集

オンラインでWebブラウザを用いて、動画に対してメタデータの作成、編集、共有ができるシステムの応用として、分散環境下で複数人によって動画編集作業を行う研究が牛木らによって行われている[9]。開発されたシステムでは、ビデオ

コンテンツを視聴しながら、サーバを介してメタデータを共有することが可能であり、「動画上に貼り付けられたメタデータを他者が見る、それをきっかけとして、他者がメタデータを貼り付ける」というコミュニケーションを行うことができる。そのシステムを動画編集に特化したものとして応用しようというアプローチである。複数人が動画編集を行うということは、議論を行いながら操作を行う協調作業であると考え、分散環境下でそれを行う際に生じる議論の発散を防ぐための手法が提案されている。

5.1.5 コミュニケーションを活発化させるインターネット街頭TV

TV ストリーミングとオンラインチャットシステムを融合したコミュニケーションシステムである *TV.com*² (ティービー・ドットコム・スクエア) が森田らによって開発されている [11]。このシステムは、街頭テレビを仮想空間内に持ち込むという概念によって開発され、Web ブラウザ上でストリーミングビデオを視聴しながらコミュニケーションを行う。TV ストリーミングによってユーザに共通な話題を与えることで、コミュニケーションが活発化されるということを評価実験によって実証している。コミュニケーションそのものに重点を置いている研究であるため、アノテーションの獲得や利用に関しては述べられていない。しかし、映像コンテンツを中心としたコミュニケーションは、オンラインビデオアノテーションの研究に欠くことはできないので、関連研究として取り上げた。

5.2 オフラインビデオアノテーションに関する研究

5.2.1 Informedia プロジェクト

カーネギーメロン大学の Informedia プロジェクト [12] では機械処理によって全自動でビデオアノテーションの作成を行うシステムを開発している。画像認識、音声認識、自然言語処理技術を統合することでビデオの自動索引付けを行うシステムである。クローズドキャプションを利用したり、文字・音声の自動認識を行うことで、ビデオに対してキーワードによる索引を自動で生成し、キーワードによるビデオ検索を可能にしている。また、TF-IDF 法 [18] を組み合わせることで、ビデオの重要なシーンを検出し、ビデオ簡約 (スキミング) も実現している [13]。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、非常に大規模な研究である。

この研究は、全自動でビデオに対するアノテーションを作成できるため、人手を必要としないという非常に大きな利点がある。しかし、シーン検索を高精度で実現しているとは言い難く、またビデオ簡約に関しても意味的な情報が考慮され

て簡約できているかについて問題が残っている。また、機械処理の精度を大幅に向上させるということは非常に困難なことである。そのため、この研究で行われているような機械処理によるアノテーションと、オンラインビデオアノテーションを組み合わせるという手法が最も優れていると考えられる。

5.2.2 Web 上の情報を利用したアノテーションの作成

オフラインビデオアノテーションを作成するために、Web 上の情報を利用するという手法も存在する。Dowman らの研究 [14] では、ニュース映像やラジオ放送に対して、音声認識によって得られた情報を基に、映像のセグメンテーションを行い、また得られたキーワードを用いて BBC ニュースサイトから、映像に関するニュース記事を検索することで、そのドキュメントを映像の各セグメントに関連付ける。また、ドキュメント内のキーワードの種類あるいは上位概念の識別を行うことで、意味的な検索を可能にしている。例えば、キーワードの種類を person で指定して Sydney を検索すると、地名の Sydney は検索にヒットしない。この研究の問題点は、音声認識の際の認識エラー率の高さである。しかし、Web 上のテキストから映像の情報を抽出してくるという手法は、オンラインビデオアノテーションにも共通している部分があり、非常に興味深い。

5.2.3 ビデオアノテーションツール

第2章で代表的なビデオアノテーションツールとして、Video Annotation Editor [4] について述べた。

同様なツールとして、リコーが開発した MovieTool [15] がある。これは MPEG-7 [20] 記述によるビデオアノテーションツールである。MPEG-7 を記述するために開発されたものであり、ビデオコンテンツに対するアノテーションを容易に作成することが可能である。特に、ストーリー・シーン・カットといった階層構造表現も含めた MPEG-7 データを記述することが可能であるという特徴がある。MPEG-7 スキーマを動的に取り込む仕組みがあり、MPEG-7 スキーマで定義されているすべてのメタデータを記述できる。また、MPEG-7 記述の編集において利用可能なタグ候補を表示する機能や、MPEG-7 の文法を定義した MPEG-7 スキーマと MPEG-7 記述との整合性のチェックを行う機能がある。

また、IBM が開発した MPEG-7 対応アノテーションツールである VideoAnnEx Annotation Tool [16] がある。これは、動画コンテンツに対して、Static Scene, KeyObjects, Event の意味属性とともに情報を付与することができるツールである。Windows 標準のツリービューを利用することにより意味属性をカテゴリ分けできる利点がある一方、ツリーが大きくなるにつれどこにどの意味属性があるの

か分かりにくく、また、ツリーの一部分を閉じてしまうと、そのツリー以下に意味属性のチェックが入っていても分かりにくいという弱点がある。

このようなビデオアノテーションツールをオンラインビデオアノテーションで利用することはないが、アノテーションの記述方法や、属性情報などは、オンラインビデオアノテーションでも非常に重要な問題であるため、関連研究として取り上げた。

第6章 まとめと今後の課題

6.1 まとめ

本研究では、映像コンテンツの柔軟な検索を目指して、まず、オフラインビデオアノテーションとオンラインビデオアノテーションのそれぞれの特徴について考察を行った上で、オンラインビデオアノテーションに着目して研究を行った。

具体的には、オンラインビデオアノテーションシステム Synvie の一般公開実験から獲得されるアノテーション情報をオンラインビデオアノテーションとして利用した。

オンラインビデオアノテーションをビデオの任意のシーンに対するタグ、すなわちシーンタグに加工して利用することを提案し、オンラインビデオアノテーションとオフラインビデオアノテーションの比較を行うために、それぞれの手法でシーンタグの作成を行った。

また、オンラインビデオアノテーションから、より質の高いシーンタグを作成するためのシステムとして、オンラインタグ選択システムを開発した。そして、タグ選択システムを利用した被験者実験を行うことで、シーンタグの作成を行った。

それぞれの手法で作成したタグを比較するために、タグを利用した新しい発想のビデオシーン検索システムを開発した。このシステムは、付与されているシーンタグの特徴を考慮し、有効利用するためのインタフェースとなっている。具体的には、シーンサムネイル表示、シーンタグ一覧表示とともにシーンタグの付与されている箇所がハイライトされたタイムラインシークバーを利用することで、ビデオコンテンツに直接アクセスすることなしに、ビデオのシーン情報を Web ブラウザ上で閲覧できる。

作成したそれぞれのシーンタグをシーン検索システムに利用してシーン検索の被験者実験を行うことで、シーンタグの比較を行った。実験から、タグのコストパフォーマンスの点で、オンラインタグ選択システムによって作成されたシーンタグが最も優れているという結果を得ることができた。

本研究を通して、オンラインビデオアノテーションを利用することで、比較的低コストでシーン検索に有用なシーンタグを作成することが可能であることを実証し、オンラインビデオアノテーションの有用性を示すことができた。

6.2 今後の課題

6.2.1 タグの付与されていないシーンの検索法の実現

今後の大きな課題として、タグの付与されていないシーン、すなわち、オンラインで獲得されるアノテーションが直接関連付けられていないシーンの検索が考えられる。本研究で開発した検索システムでは、タイムラインシークバーのスライダーによってタグの付与されていないシーンを閲覧することができるため、実験でもそのようなシーンを検索することが可能であった。しかし、そのためにはそのシーンの周辺に検索の手がかりとなるタグが付与されている必要があり、さらに、検索結果から直接的にそのシーンを検索することができない。そのため、タグの付与されていないシーンにも情報を付与する仕組みが必要である。特に、オンラインビデオアノテーションはその量がコンテンツによって偏りが生じる可能性が高いため、アノテーションが少量であるコンテンツに対する検索も可能にするために、アノテーションの不足を補う手法を実現することは、非常に重要な問題であると考えられる。

6.2.2 シーンタグ作成法の改良

オンラインビデオアノテーションから、より多くの情報を抽出することで、より有用なシーンタグを作成できると考えられる。それを実現するために、シーンタグを作成するための仕組みには様々な改良点が考えられる。具体的には、形態素解析の際に細かく分割しすぎた語の取り扱いや、タグの属性の定義などが考えられる。また、タグを付与するタイムコードを広げるために、カット検出などの機械処理による情報を組み合わせるということも考えられる。これらによって、本研究では情報を付与することができなかったシーンにもタグを付与することができる可能性がある。本研究では、被験者実験でのみオンラインタグ選択システムを使用した。一般公開によってユーザが利用するためには、システムを自然に利用させるような仕組みやインタフェースが必要である。そのためには、タグを選択しながらコミュニケーションを行う仕組みや、ゲーム感覚でタグ選択を行う仕組みなどが考えられる。

6.2.3 アノテーションをより大量に獲得するためのシステムの開発

オンラインビデオアノテーションにおいて無視できない問題として、アノテーションの獲得の問題がある。オンラインビデオアノテーションは、アノテーションの量や質がシステムのユーザに依存するため、その量の増加や質の向上のためには、より多くのユーザにシステムを利用してもらう必要がある。本研究でアノ

テーションを利用した Synvie の一般公開実験でも、実験開始当初は、ユーザ数が伸び悩み、アノテーションの量が非常に少ないという問題があった。公開期間の経過とともに少しずつアノテーションの量も増加したが、現在でも十分に満足できる量であるとは言い難く、また、コンテンツによる偏りも大きい。そのため、アノテーションが一定以上獲得された 27 個のビデオコンテンツのみを本研究で利用した。アノテーションの量が増加することで、検索精度や検索コストが向上し、また、要約などのその他の応用も可能になる可能性があると考えられる。

多くの人間を集めるためには、システムを自然に利用するための動機付けを与えなければならない。そのためには、システム自体の使いやすさや、おもしろさ、話題性などが必要である。また、Synvie のように、ビデオを共有してアノテーションを獲得するシステムである場合は、ビデオコンテンツ自体のおもしろさも非常に重要な要素であるが、著作権や肖像権の問題が絡んでくるため運用には注意が必要である。そのため、ニコニコ動画 [33] のような、ビデオコンテンツ自体は共有せずに、ビデオを中心としたコミュニケーションを行うシステムによってアノテーションを獲得するという手法も有効であると考えられる。

また、アノテーションを大量に獲得するだけでなく、その中に含まれる有用な情報の割合が高くなるような手法を考える必要がある。

6.2.4 ビデオコンテンツの新しい応用の考案と実現

本研究では、オンラインビデオアノテーションをシーン検索に利用することで、その有用性を実証したが、ビデオコンテンツには様々な応用が存在する。検索以外には、要約や変換などがすでに提案されているが、まだ提案されていない新しい形の応用があると考えられる。特にオンラインビデオアノテーションは、比較的新しい研究分野であるため、その特性を十分に活かした応用や、新たな応用のためのアノテーション獲得手法を考案していく必要があるだろう。

参考文献

- [1] Katashi Nagao. “Digital Content Annotation and Transcoding,” Artech House Publishers, 2003.
- [2] 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, 特集「便利で身近な音楽情報処理」, Vol.48, No.1, pp.258–273, 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, pp.13–18, 2006.
- [4] Katashi Nagao, Shigeki Ohira, Mitsuhiro Yoneoka. “Annotation-Based Multimedia Summerization and Translation,” In Proceedings of the Nineteenth International Conference on Computational Linguistics (COLING-2002), 2002.
- [5] 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, 2002-GN-43, No.31, pp.37–42, 2002.
- [6] 畠田聡, 宮川和, 森本正志. “ファンコミュニティサイトにおける映像シーン連動型掲示板コミュニケーションの分析”, 電気情報通信学会技術報告, HCS2005-50, pp.69–74, 2005.
- [7] 谷口行信, 南憲一, 佐藤隆, 桑野秀豪, 児島治彦, 外村佳伸. “SceneCabinet: 映像解析技術を統合した映像インデクシングシステム”, 電子情報通信学会論文誌, Vol.J84-D-II, No.6, pp.1112–1121, 2001.
- [8] 中村誠, 大堀順也, 池橋貫司, 依田育生. “ネットワーク簡易動画編集アプリケーションの開発”, 2001年電子情報通信学会総合大会, 2001.
- [9] 牛木一成, 西田真啓, 中筋文, 湊賢治. “動画像メタデータを用いたインタラクション機構によるダイナミックコミュニティ形成と動画編集コラボレーション”, 2003年電子情報通信学会総合大会, 2003.

- [10] 宮森恒, 中村聡史, 田中克己. “番組放送チャットに基づく視聴者視点を利用した放送番組のビュー生成”, DEWS2005, 4B-i9, 2005.
- [11] 森田篤史, 金谷裕幸, 西本一志, 国藤進. “*TV.com*²: コミュニケーションを活性化させるインターネット街頭 TV”, DICOMO2002, 情報通信学会シンポジウムシリーズ, Vol.2002, No.9, pp.429–432, 2002.
- [12] D. Wactlar, T. Kanade, A. Smith, and M. Stevens. “Intelligent access to digital video: Informedia project,” IEEE Computer, Vol.29, No.5, pp.140–151, 1996.
- [13] Michael A. Smith, Takeo Kanade. “Video Skimming and Characterization through the Combination of Image and Language Understanding,” IEEE International Workshop, ICCV98, 1998.
- [14] Mike Dowman, Valentin Tablan, Hamish Cunningham, Borislav Popov. “Web-Assisted Annotation, Semantic Indexing and Search of Television and Radio News,” World Wide Web Conference 2005, 2005.
- [15] Ricoh. Movie Tool, <http://www.ricoh.co.jp/src/multimedia/MovieTool/> , 2002.
- [16] Ching-Yung Lin, Belle L. Tseng, and John R. Smith. VideoAnnEx Annotation Tool, <http://www.research.ibm.com/VideoAnnEx/> , 2002.
- [17] Taku Kudo, Yuji Matsumoto. “Fast Methods for Kernel-Based Text Analysis,” ACL-03. 2003.
- [18] G. Salton, A. Wong, and C.S. Yang. “A Vector Space Model for Automatic Indexing,” Communications of the ACM, Vol.18, No.11, pp.613–620, 1975.
- [19] W3C. Extensible Markup Language (XML), <http://www.w3.org/XML/> , 1996.
- [20] MPEG. MPEG-7, <http://ipsi.fraunhofer.de/delite/Projects/MPEG7> , 2002.
- [21] W3C. Synchronized Multimedia Integration Language (SMIL 2.0), <http://www.w3.org/AudioVideo/> , 2002.
- [22] Google Code, <http://code.google.com/> .
- [23] Google Suggest, <http://labs.google.com/suggest/> .
- [24] Synvie Public Beta Service, <http://video.nagao.nuie.nagoya-u.ac.jp/> .

- [25] YouTube, <http://www.youtube.com/> .
- [26] Wikipedia, <http://ja.wikipedia.org/wiki/> .
- [27] 2ちゃんねる, <http://www.2ch.net/> .
- [28] mixi, <http://mixi.jp/> .
- [29] Flickr, <http://www.flickr.com/> .
- [30] del.icio.us, <http://del.icio.us/> .
- [31] はてなブックマーク, <http://b.hatena.ne.jp/> .
- [32] Ask ビデオ, <http://video.ask.jp/> .
- [33] ニコニコ動画 (β) , <http://www.nikovideo.jp/> .

謝辞

本研究を進めるにあたり、指導教員である長尾確教授には、研究に対する姿勢や心構えといった基礎的な考え方から、研究に関する貴重な御意見、論文執筆に関する御指導など、大変お世話になりました。心より御礼申し上げます。

大平茂輝助手には、研究に関することから技術的なことまで幅広く御指導、御意見を頂き、大変お世話になりました。心より御礼申し上げます。

友部博教さん、梶克彦さんには、ゼミ等で貴重な御意見を頂き、また、研究室の楽しい雰囲気づくりをしていただきました。山本大介さんには、本研究を進めるにあたり絶対不可欠であるオンラインビデオアノテーションシステム Synvie に関することや、研究に対する基礎的な進め方やアドバイス、技術的な御指導など様々な面で大変お世話になりました。土田貴裕さんには、Java や JavaScript など、実装に関して非常に未熟である私の質問にいつでも答えていただき、また、様々なアドバイスを頂き非常にお世話になりました。ここに御礼申し上げます。

成田一生さん、林亮介さん、伊藤周さん、石戸谷顕太郎さん、金田哲広君には、研究や実装に関することや、研究室生活における様々な面でお世話になりました。ありがとうございました。

長尾研究室秘書である金子幸子さんには、研究室生活や学生生活の様々な面でお世話になりました。ありがとうございました。

最後に、影ながら見守っていただき、生活を支えていただいた両親にも最大限の感謝の気持ちをここに表します。ありがとうございました。