1G1-6

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

Video Scene Retrieval Using Online Video Annotation

増田 智樹[*1] 山本 大介[*1] 大平 茂輝[*2] 長尾 確[*3]
MASUDA, Tomoki YAMAMOTO, Daisuke OHIRA, Shigeki NAGAO, Katashi

[*1]名古屋大学 大学院情報科学研究科
Graduate School of Information Science, Nagoya University

[*2]名古屋大学 エコトピア科学研究所
EcoTopia Science Institute, Nagoya University

[*3]名古屋大学 情報メディア教育センター
Center for Information Media Studies, Nagoya University

In this paper, we propose an efficient method for scene tag extraction from online video annotation (e.g., comments to video scenes). For evaluating the method by applying extracted information to video scene retrieval, we have developed a video scene retrieval system based on scene tags (i.e., tags associated with video scenes). We also have developed a tag selection system that online users can select appropriate scene tags from automatically created data using online video annotation. Furthermore, we performed subject experiments on selecting tags and video scene retrieval. Through these experiments, we found that scene tags extracted by using a tag selection system have higher cost performance than other scene tags created with a conventional client-side video annotation tool.

1. はじめに

近年のインターネット技術の発達やブロードバンド回線の普及によって、Web 上に膨大な数のビデオコンテンツが存在するようになった。また、YouTube[*1]などのビデオ共有サービスの出現によって、誰でも手軽に Web 上にビデオを公開することができるようになったため、専門家が作成するビデオだけでなく一般人が作成したビデオが非常に大量に存在するようになった。それに伴い、それらのビデオコンテンツに対して検索や要約などといった応用に対する要求が非常に高まっており、今後もさらに高まっていくと考えられる。

それらの応用を実現するためには、ビデオの内容情報を記述するビデオアノテーションが必要不可欠であり、多くの研究が行われている [1]。筆者らは近年 Web 上で行われているビデオコンテンツを中心とした自然なコミュニケーション活動からビデオの内容情報をオンラインビデオアノテーションとして獲得するシステム Synvie を開発した [2]。また、一般公開実験によってアノテーションデータの収集を行っている[*2]。

オンラインビデオアノテーションは、アノテーションの作成コストが低いことや、多様な人間による意味情報が含まれる可能性があるという利点があるが、利用価値の低い情報も含まれるため情報の選別が必要であると考えられる。そこで、本研究では、オンラインビデオアノテーションを基に価値の高いアノテーションの選別法、利用法を提案する。具体的には、オンラインビデオアノテーションの選別を行い、それをビデオシーン検索に利用する仕組みの提案を行う。

まず、オンラインビデオアノテーションを基にビデオシーンに対するタグ（以後シーンタグと呼ぶ）の作成を行った。さらに、シーン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、シーン検索の被験者実験を行った。

本稿を通して、オンラインビデオアノテーションとその選別法、利用法の有用性を検証した。

連絡先: 増田 智樹，名古屋大学大学院情報科学研究科 長尾研究室，愛知県名古屋市千種区不老町，TEL,FAX:052-789-5878，masuda@nagao.nuie.nagoya-u.ac.jp

∗1 YouTube: http://www.youtube.com/

∗2 Synvie Public Beta Service: http://video.nagao.nuie.nagoya-u.ac.jp/

2. 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。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 3 種類の手法でシーンタグを作成した。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 秒であり、ビデオの種類は教育、物語、エンターテインメントなど様々なものがある。

次章のシーン検索実験によって、作成されるタグの有用性の比較評価を行うために、3 種類の手法によってシーンタグの作成を行った。

2.1 専用ツールを用いたタグ付け

ビデオを視聴しながら、任意の開始時間、終了時間を指定してタグを付与することができるツールを利用し、１人のアノテータによってタグ付けを行った。アノテータは、ビデオの制作者ではなく、各ビデオに関する詳細な知識を待たないため、映像や音声から客観的に得られる情報であるシーンのイベント情報や、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をできる限り詳細かつ網羅的にシーンタグとして付与した。この手法は従来から行われてきた人手によるオフラインビデオアノテーションの一種である。

ここで、シーンタグの付与に費やした時間を、シーンタグ作成のためのコストとした。その時間は、１コンテンツあたり平均で約 1480 秒、最長で 3692 秒、最短で 582 秒であった。

2.2 オンラインビデオアノテーションからの自動抽出

Synvie は、Web 上でビデオの任意のシーンに対してコメントの投稿、ブログへの引用を行うことができるビデオ共有システムである。本研究では 2006 年 7 月 1 日から 11 月 30 日までに一般公開実験によって獲得されたアノテーションデータを利用した。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 で獲得される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。Cabocha[3] を利用してテキスト情報の形態素解析を行うな

どの処理によってアノテーションデータから自動的にシーンタグの生成を行った。シーンタグ生成までの処理手順は次のとおりである。

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いしまうカナ 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. 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。

これらの処理はすべて機械処理によって自動で行うことができ、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。

この手法では、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た。

2.3 タグ選択システムを用いた抽出

Synvie で獲得さ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。ブログやコメント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。実際に自動生成された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い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タグなどが含まれていた。そのため、前節でオンラインビデオアノテーションから自動生成し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。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と考えられる。

この選別を機械処理によって全自動で行うことが最も理想的であるため、いくつかの手法によってそれを試みた。まず、TF-IDF(Term Frequency-Inverse Document Frequency)[4] の原理によって重み付けを試みたが、この手法は大量のドキュメントを必要とするため、ドキュメント数が十分でない現状では有効な選別を行うことができなかった。次に、Google Web API*3を利用して、コンテンツ投稿時に付与されたタグとの共起関係によって重み付けを行ったが、シーンに関わらず一般的なキーワードの重みが高くなってしまい、この手法でも有効な選別を行うことができなかった。やはり、シーンタグの選別を機械処理で行うことは非常に困難な問題であり、有効な選別を行うためには少なからず人間の力が必要であると考えられる。そのため、シーンタグの選別を人手によって行うシステムを開発した。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大量にかけてしまえば、オンラインビデオアノテーションを利用した効果が現象してしまうため、なるべく個人の負担を小さく、効率的にタグの選別を可能である仕組みにする必要がある。具体的には、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複数人によって選択するシステムである。ビデオを視聴中にシーンタグの付与されているタイムコードになるとビデオが一時停止し、そのシーンを閲覧することでタグの選別を行う。

*3 Google Code: http://code.google.com/

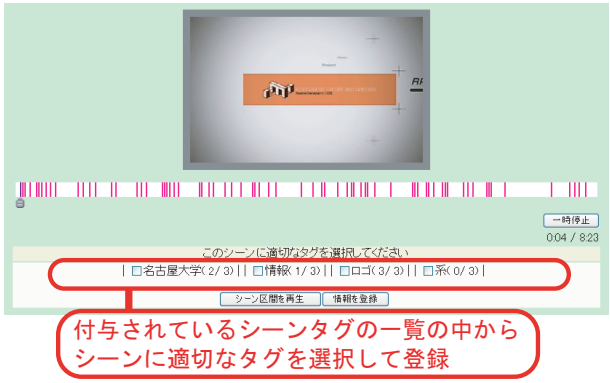

図 1: タグ選択システムの画面例

表 1: 各手法におけるタグの作成コスト

タグ作成手法	作成コスト (秒)
オフラインでタグ付け	1480
自動抽出	0
オンラインでタグ選択	314

このシステムを用いて被験者実験を行うことで、タグの選別を行った。各ビデオコンテンツに対する被験者は 2 人または 3 人である。

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、タグの選択に各個人が費やした時間とした。この時間はタグの選択中の時間を自動的に取得しておくことで算出した。その時間は、1 コンテンツあたり平均で約 314 秒で、2.1 でタグ作成に費やした平均時間の約 5 分の 1 であった。また、最長で 1121 秒、最短で 33 秒であった。

この実験によって、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、最多で 277 個、最少で 7 個のシーンタグが作成された。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 36.2 %が人間の目によってそのシーンを検索するためのタグとして適切だと判断された。

各作成手法におけるタグの作成コストを比較したのが表 1 である。

3. 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利用した Web ページをそれぞれ作成し、被験者実験を行った。

3.1 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ビデオ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

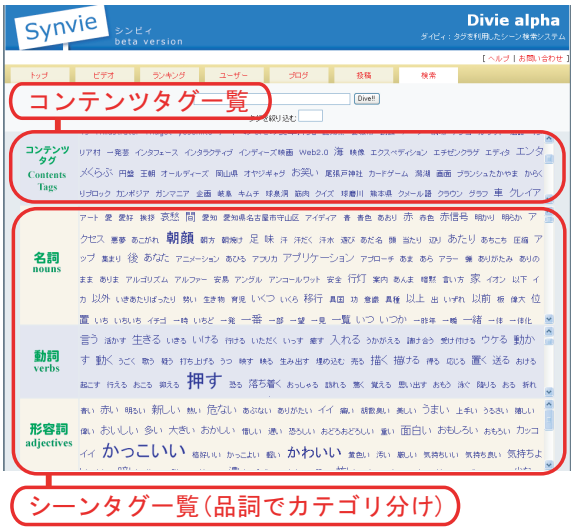

図 2: シーン検索システムトップページ

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。

以上のような特徴を考慮し、検索クエリに対して、各シーンをランキング付けして表示するのではなく、まずビデオをランキング表示し、ビデオコンテンツファイルに直接アクセスすることなくビデオの内部情報を Web ブラウザ上で閲覧することでシーンを検索する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おける検索の流れは次のとおりである。

1. 登録されている全てのビデオに対するシーンタグの一覧から任意の数のタグを選択し、検索を行う。
2. ビデオがランキング表示される。各ビデオには検索クエリとして利用されたタグが付与されている箇所を反映したタイムラインシークバーと、付与されているシーンタグの一覧が表示される。
3. シーンタグの一覧の中からタグを選択することで、タイムラインシークバーに反映させることができる。
4. タイムラインシークバーを動かすことで任意のタイムコードに対するサムネイル画像を閲覧する。
5. ビデオを任意のタイムコードから再生する。

図 2 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、コンテンツ投稿時に付与されたタグ、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、それぞれのカテゴリの中では、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は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ため、キーボードを使ったテキスト入力を行うことなくタグを利用することができる。また、タグをインクリメンタル検索することが可能である。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のことである。具体的には、タグを絞り込むためのテキストフィールドにひらがな

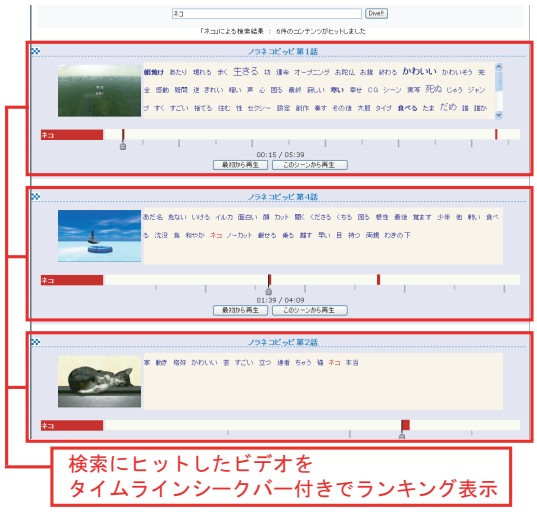

図 3: 検索結果ページの画面例

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。内部的な処理として、あらかじめ読み込んでおいたタグを JavaScript で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せず、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。

検索結果として、ビデオコンテンツが、内部情報をブラウザ上で閲覧するためのタイムラインシークバー付きでランキング付けされて表示される（図 3）。また、シークバーのタイムコードに対応したサムネイル画像、シーンタグの一覧が表示される。シークバーを動かすとそのタイムコードに対応してサムネイル画像が切り替わるため、タグの付与の有無に関わらず、任意のタイムコードに対するサムネイル画像を閲覧することができる。また、このタイムラインシークバー上には、クエリとして使用したタグが時間軸上のどのタイムコードに付与されているかがハイライト表示されるため、検索したいシーンのキーワードがタグとして付与されている箇所のサムネイル画像を容易に閲覧することができる。さらに、シーンタグの一覧からタグをシークバー上に 1 クリックで追加することができ、それを繰り返すことで、検索したいシーンを絞り込んでいき、任意のタイムコードからの再生することで、ビデオシーン検索を実現する。ビデオシーン検索を行った画面例を図 4 に示す。

検索に利用されたタグや、再生されたタイムコードなどの情報をフィードバックすることで検索をより良くしていく仕組みなど、新たな機能の実装や設計などを現在も進行中である。

3.2 シーン検索実験

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。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

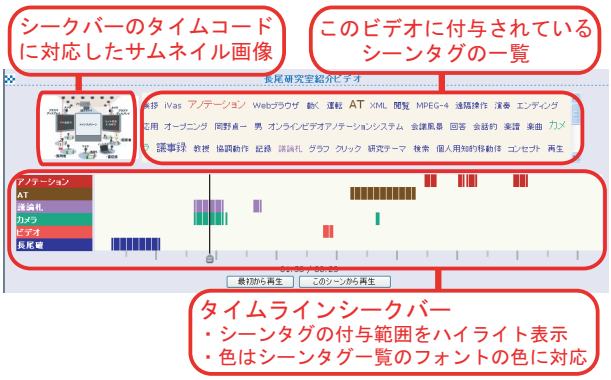

図 4: ブラウザ上でのビデオ内部情報の閲覧

表 2: シーン検索実験結果

タグ作成手法	平均検索時間 (秒)	クエリ送信数
オフラインでタグ付け	118.1	132
自動抽出	169.6	202
オンラインでタグ選択	145.4	156

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。実験用のトップページを用意することで、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。

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

今回の実験では、全被験者がすべての出題に対して正しいシーンを発見できたため、その観点ではシーンタグ作成手法の比較を行うことはできなかった。そのため、シーン検索に費やされた時間によってシーンタグの作成手法の比較を行った。その実験結果が表 2 である。平均検索時間の増加に対応してクエリ送信数も増加しているため、タグの違い以外の何らかの影響によって検索時間に違いが出たわけではないと考えることができる。表から、オフラインで専用ツールを用いたタグ付け、オンラインでタグ選択システムによって抽出、オンラインビデオアノテーションから自動抽出の順で検索に時間がかかったということがわかる。つまり、シーンタグの作成に人手をかけた 2 つの手法では、検索時間が短縮された。シーンタグの作成コストの点では、オンラインビデオアノテーションから自動生成が最も優位性があるが、検索コストが特に高く、また、今後膨大な数になっていくビデオコンテンツに対して検索時間を短縮するということは不可欠な問題であり、そのためには少なからず人手をかける必要があると考えられる。

表 3: 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	3.48
オンラインでタグ選択	7.71

そのため、オフラインで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 (費用対効果)、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。タグのコストパフォーマンス C は次の式によって計算した。

$$C = \frac{(\text{自動抽出による平均検索時間}) - (\text{平均検索時間})}{(\text{シーンタグの作成に費やされた時間})} \times 100 \quad (1)$$

この式から、タグの作成のコスト 100 秒あたり、どれだけ検索時間が短縮されたかを計算することができる。その結果を表 3 に示す。

表 3 から、コストパフォーマンスの観点で比較を行った場合、今回の実験では、タグ選択システムを用いたシーンタグの作成手法に優位性が見られた。

4. まとめと今後の課題

4.1 まとめ

本研究では、オンラインビデオアノテーションを利用し、また、アノテーションの選別を行うことで、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できた。これによってオンラインビデオアノテーションの有用性とその選別の有効性を実証することができた。

また、シーンタグを用いた新しい発送のビデオシーン検索システムを開発し、オンラインビデオアノテーションの利用法の提案を行った。

4.2 今後の課題

今後の課題として、より大量のデータに基づいての検証が挙げられる。大量のデータを収集するためには、被験者による実験だけでなく、一般公開実験によってより自然にデータを収集するということが考えられる。また、ビデオシーン検索システムのインタフェースの改善や取得するデータの検討、機能の拡張、検索アルゴリズムの改善などや、シーンタグの自動生成の手法の改善などが課題として挙げられる。

参考文献

[1] Katashi Nagao. “Digital Content Annotation and Transcoding”, Artech House Publishers, 2003.

[2] 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, p13-18, 2006.

[3] Taku Kudo, Yuji Matsumoto. “Fast Methods for Kernel-Based Text Analysis”, ACL-03, 2003.

[4] 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.