NAIST-IS-MT1351018

修士論文

テキストを用いてユーザ意図を反映する映像要約

大谷 まゆ

2015年3月12日

奈良先端科学技術大学院大学
情報科学研究科 情報科学専攻

本論文は奈良先端科学技術大学院大学情報科学研究科に
修士 (工学) 授与の要件として提出した修士論文である。

大谷 まゆ

審査委員：

横矢 直和 教授　（主指導教員）
加藤 博一 教授　（副指導教員）
佐藤 智和 准教授　（副指導教員）
中島 悠太 助教　（副指導教員）

テキストを用いてユーザ意図を反映する映像要約*

大谷 まゆ

内容梗概

近年，スマートフォンや安価なビデオカメラの普及により，一般ユーザが大量に映像を撮影できるようになり，これらの映像をインターネットを通して公開することが一般的になった．その形態として映像とそれに付随するテキストを用いて，ユーザが日々の出来事や考えを表現するビデオブログがある．しかしながら，映像が大量に撮影された場合，撮影された雑多な映像の中からユーザの意図に合った映像を抽出し，編集することは手間を要する作業である．映像編集にかかるこれらの労力を削減可能な技術の一つに映像要約がある．これは長時間に及ぶ大量の映像から短いダイジェスト映像や代表フレームを提示することで内容を容易に把握可能とすることを目的としている．従来，様々な映像要約手法が研究されており，それらの多くは事前に設計された指標に従って映像をサンプリングすることで要約映像を作成する．しかし，それら多くの手法はユーザの意図を反映するように要約映像をコントロールすることができず，ビデオブログのための映像制作には適していない．

そこで本研究では，ユーザの意図に着目した要約映像の生成を目的とする．提案手法ではユーザが表現したい内容を記述したビデオブログのテキストを用い，そのテキストの内容に合った箇所を選択することで，ユーザの意図を反映した要約映像を生成する．提案手法はテキストと映像をそれぞれオブジェクトの集合として考える．テキストについては名詞を，映像については画面に写った人やものをオブジェクトして用いる．このとき，映像のオブジェクト表現を得るために，映像中のオブジェクトをアノテーションとして付与することが必要となる．そこ

*奈良先端科学技術大学院大学 情報科学研究科 情報科学専攻 修士論文, NAIST-IS-MT1351018, 2015 年 3 月 12 日.

で本研究ではまず，映像のアノテーション付与のためのインターフェースを提案する．提案手法ではオブジェクト表現を用いてテキストと類似する映像群を選択し，要約映像を生成する．具体的には，テキストと映像のオブジェクト表現の類似度を定義し，この類似度に基づく目的関数の最適化として，映像要約を定式化する．加えて，テキストと映像の類似度だけでなく，クラスタリングに基づく映像の優先度を目的関数に導入することで，よりテキストの内容に合った映像の選択を試みる．

本研究では提案手法を評価するためにユーザスタディを実施した．ユーザスタディでは要約映像をビデオブログに使用することを想定し，種々のベースライン手法と比較した．このユーザスタディでは主に次の 2 点について調査した．

1. ビデオブログ用動画の作成支援として提案手法が有効であるか
2. テキストに沿った内容の要約映像が実現できているか

この結果から，提案手法がビデオブログのための映像編集に有効であることを示すとともに，新たに得られた要約映像の要件に関する知見について述べる．

キーワード

映像要約，ビデオブログ，ユーザスタディ

User Intention-Oriented Video Summarization Based on Textual Description *

Mayu Otani

Abstract

The popularization of camera devices, such as smartphones or minicams, enables people to record videos in everyday life, and many users share the videos through the Internet. Video blogs have recently gained attention as expressive media, with which users can express themselves using videos and its supporting text. However, editing videos that well present user's intention is time consuming if he/she has a vast amount of videos. Video summarization can facilitate such an editing process. Video summarization is a technique to provide a compact representation, such as a short video or a set of keyframes for long videos. Typical methods sample frames or short video segments, called shots, based on preliminarily defined criteria. Unfortunately, these methods are not suitable for video blogs because their resulting video summaries do not take the user's intention into account.

This work proposes a novel video summarization method for video blogs considering the user's intention. The proposed method leverages the text, in which the user describes his/her intention, or stories, to generate video summaries consistent with the user's intention. In this method, the input text and videos are represented as sets of objects. In order to obtain object-based representation for text, we extracts nouns in the text as objects, and for videos, we use objects

*Master's Thesis, Department of Information Science, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-MT1351018, March 12, 2015.

annotated by the user to videos with our proposed interface. We then formulate video summarization as an optimization problem of an objective function, which involves the content similarity between the video summary and the text, as well as the content coverage over the text, defined by using the object-based representation of videos and text.

User study has been conducted to evaluate our video summarization method in terms of the following two points:

1. Whether generated video summaries are suitable for video blog posts
2. How well generated summaries represent the input text

The results have demonstrated the proposed method is superior to baseline methods in suitability to blog posts. We conclude this paper with remarks of requirements for video summaries.

Keywords:

Video summarizaiton, video blog, user study

目 次

図 目 次

表 目 次

1. はじめに

スマートフォンや安価なハンディカメラの普及により，映像を撮影し，インターネットを通じて公開することが一般的となった．そのような映像利用の方法の一つとして，映像とそれに付随するテキストを用いてユーザが日々の出来事や考えを表現するビデオブログが注目を集めている [1]．ビデオブログは多くの場合，図 1 のようにテキストとそれに対応する映像で構成されており，ユーザは撮影した映像をブログの内容に合わせて編集し公開する．

現在のビデオブログは，これに用いられる映像の内容から，ユーザがカメラの前でプレゼンテーションをする様子を記録した映像を用いるプレゼンテーション形式と，イベントやアクティビティを記録したホームビデオ形式の 2 種に大別される．両者はその撮影，編集工程が大きく異なる．前者のプレゼンテーション形式は事前に用意した進行に従って撮影するため，撮影された映像に不要なシーンが少ない．映像編集では，撮影された映像から不要箇所を除去するだけで，ビデオブログのための映像を作成することができる．このように，プレゼンテーション形式のビデオブログは撮影時に台本などの準備が必要であるものの、ビデオブログに載せる最終的な映像に近いものが撮影されるため，映像編集は比較的容易である．一方，後者のホームビデオは決まった進行がなく，不要なシーンを含む大量の映像を撮影しがちである．このため，映像編集では記事の内容に合った映像を生成するために，ユーザが雑多な映像からイベントの重要箇所などを抽出す

図 1　テキストが付随したビデオブログの例．

る必要がある．そのため，プレゼンテーション形式に比べて，ホームビデオ形式でのビデオブログ記事の制作は，映像編集に膨大な手間がかかるという問題がある．現在，ビデオブログはプレゼンテーション形式が主流であり [1]，ホームビデオ形式のビデオブログ制作は撮影が容易であるにもかかわらず，その編集の手間から盛んではない．

そこで本研究では，大量の雑多な映像からビデオブログのための映像へと自動編集する手法を提案し，一般ユーザが撮影したホームビデオからのビデオブログの制作にかかる労力を削減することを目的とする．映像編集にかかる手間を削減する方法として，映像要約手法の利用が考えられる [1]．映像要約は，視聴に時間がかかる映像から短いダイジェスト映像や代表フレームを抽出し，短時間での内容把握を可能とする技術である．

従来，映像要約の分野では，映像全体の内容の被覆度や重要なイベントやオブジェクトの有無などの指標に従い，映像をサンプリングする手法が盛んに研究されてきた [2, 3, 4]．しかし，これら従来手法は事前に設計した指標に基づき要約映像を生成しており，各ユーザおよびその時々の用途に応じた内容を含む要約映像を生成できない．一方で，ビデオブログとして用いられる映像は，そのブログでユーザが意図した内容を含む必要があることから，従来の映像要約手法はビデオブログのための映像生成に適さない．

そこで本論文ではユーザが記述したブログに付随するテキストにユーザの意図が表現されていると考え，そのテキストに沿った映像を選択する映像要約手法を提案する．そうすることで，提案手法は，従来の映像要約では考慮されていないユーザの意図を要約映像に反映し，ビデオブログに適した映像を生成する．この手法ではブログのためのテキストを用いることで，映像生成にかかる負担を削減しながら要約映像の制御を可能とする．

提案手法は図 2 に示すように，テキストと映像をオブジェクトの集合で表し，それらの類似度を用いた要約映像を生成する．オブジェクトはテキスト中の単語から自動的に抽出したものと，ユーザが映像中のオブジェクトをアノテーションとして付与したものを用いる．本研究では，映像へのアノテーション付与のためのインターフェースを開発する．また，テキストと映像の類似度に基づく目的関

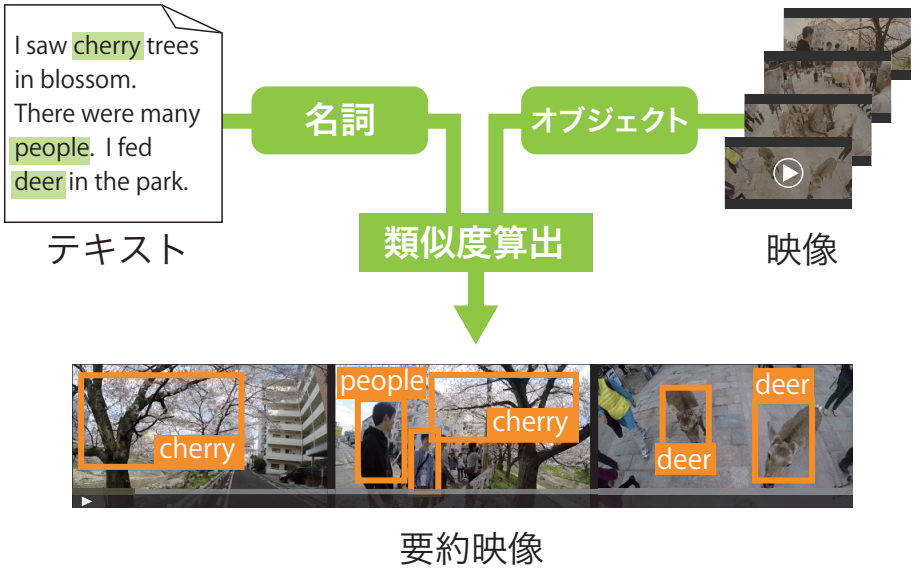

図 2　テキストと映像のオブジェクト集合に基づく類似度を用いた映像要約.

数を設計し，映像要約を目的関数の最適化問題として定式化する．この目的関数を動的計画法を用いて近似的に最適化することで，要約映像に含まれる映像を決定する．実験では，要約映像を評価するために，20 人の被験者に対しユーザスタディを実施し，提案手法がビデオブログのための映像編集に対し有効であることを示す.

2. 関連研究と本研究の位置づけ

本章では，関連研究として映像要約とオブジェクトのアノテーション付けに関する従来研究をまとめた上で，本研究の位置づけを明らかにする．

2.1 映像要約に関する従来研究

本研究ではビデオブログ制作のための映像の自動編集を実現するために映像要約手法を用いる．映像要約とは，映像の視聴にかかる手間を削減するために，ユーザがより短い時間で内容を把握できるコンパクトな表現を生成する技術である．具体的には，短いダイジェスト映像や重要箇所のフレームを抜き出した画像を生成する．

映像要約手法は一般的に，ショット分割などの前処理，重要箇所抽出，要約画像・映像生成の 3 段階の処理で構成される．映像の重要箇所の推定には，映像やそれに付随するメタデータなどの様々な情報が用いられる．本節では重要箇所推定に用いられる情報の観点から従来手法をまとめる．

2.1.1 入力映像のみに基づく映像要約

入力映像のみから得られる情報を用いた手法はこれまで数多く提案されている基本的なアプローチである．そのなかでも視覚的な低レベル特徴量を用いた映像要約手法は近年に至るまで活発に研究されている [5, 6, 7, 8]．これには，クラスタリングに基づいて映像中からその映像をよく表す箇所を抽出する手法 [5] や，特徴量の際立った変化から重要箇所を抽出する手法が含まれる [7, 8]．例えば Laganière ら [8] は時間方向，空間方向の特徴量変化を用いて顕著なシーンを抽出している．

多くの映像は視覚情報に加え，人の発話や音楽などの音声情報を伴うことから，音声に着目した手法が提案されている [9, 10, 11]．音声を用いることの利点としては，発話内容を解析することで，視覚情報からの推定が困難な映像の意味の扱いが可能となることや，大きな音や突然の音など，人の注意を引きつける音声を

考慮した重要箇所の推定が可能となる点などがあげられる．Ma ら [9] は，音声情報を考慮した映像の顕著度に基づく映像要約を提案した．一方 Taskiran ら [10] は，音声認識を用いて発話された単語を取得し，その重要度に基づいて要約映像に取り入れる箇所を選択した．

またこのような特徴量ベースの手法とは異なるアプローチとして，映像中のストーリーやイベントの進行などの構造を利用した手法が提案されている [12, 13]．近年では Lu ら [13] が，あるイベントが別のイベントを引き起こす様子を映像に出現するオブジェクトから推定し，イベント列からなる要約映像を生成した．

対象映像についての事前知識が得られる場合などは，発生するイベントやオブジェクトを検出することで，映像の重要箇所を検出することができる [14, 15, 16]．例えば，スポーツ映像であれば得点のシーンが重要と考えられることから，これらのシーンを抽出することで，試合内容を要約することができる．Babaguchi ら [14] はアメリカンフットボールの映像を対象とし，得点のシーンや優位なチームの入れ替わりなどに着目して，試合における各イベントの重要度を定義し，ショット選択に用いた．

2.1.2 映像に付随する情報やインターネットの情報を用いた映像要約

映像のみから推定することが困難な映像の意味などは，映像の内容を表現した他のメディアから得られる場合がある．それらのメディアを解析することで重要箇所を推定する手法が盛んに研究されている [14, 17, 18, 19]．例えば上記の Babaguchi ら [14] の手法は，スポーツ映像の重要イベントの検出のために公開された試合記録を用いる．また，Sang ら [18] は映画に付随する台本を用いて，キャラクターの出番やセリフに基づくシーンの重要度を定義した．一方，インターネット上の画像や動画の利用も検討されている．例えば Khosla ら [19] の手法はインターネットの画像から学習した結果を用いて，入力映像から代表的な構図で撮影された箇所を抽出している．

2.1.3 ユーザに関する情報を用いた映像要約

よりユーザの要望に合った要約映像を生成することを目的として，ユーザに関する情報を用いる手法が提案されている [14, 20, 21]．Syeda-Mahmood ら [20] の手法は隠れマルコフモデルを用いてユーザの視聴行動をモデル化し，視聴者の状態予測に基づいた映像評価から要約映像に取り入れる箇所を選択して映像のプレビューを作成する．一方，Aizawa ら [21] はユーザの脳波を用いた手法を提案した．この手法はライフログ映像を対象に，ユーザが集中状態にあるときの映像から要約映像を生成する．また，上記の Babaguchi ら [14] の手法では，各ユーザの嗜好に基づく要約映像を生成するために，ユーザの好みのチームや選手などの情報が取り入れられた．

2.2 アノテーション付与に関する従来研究

提案手法では，映像をオブジェクトの集合として扱う．そこで，本研究ではオブジェクトの有無とその位置をアノテーションとして映像に付与する．従来，画像や映像へのアノテーション付けには，膨大な画像や映像の処理にかかる手間と，アノテーションの誤りや表記の揺れの問題があった．これらの問題を解消するために，インターフェースや機械学習の分野で様々な手法が研究されている．本節ではそれらの手法を，人手による手法，機械学習を用いた自動化手法，人による入力を補助的に利用した手法の 3 つに分類して紹介する．

2.2.1 人手によるアノテーション付与

人手によるアノテーション付与はインターネット上では広く行われており，Flickr や Youtube など多くのメディア共有サービスでは，ユーザによるオブジェクトやシーンのラベル付け機能を提供している．また，Amazon Mechanical Turk [22] などのクラウドソーシングサービスによる大規模なアノテーション付与も利用されている．しかしながら，人手によるアノテーション付与は対象の数に比例して膨大な労力がかかり，また，習熟していないユーザによるアノテーションは多くの誤りや対象およびラベルのばらつきを含む [23]．この問題を解決するためのゲー

ムデザインやインターフェースが開発されており，Von Ahn ら [24] は誤りとばらつきが少ないアノテーション付与を実現するために，2 人のプレーヤーがお互いに付与するラベルを予想するゲームを考案し，これを用いてインターネット上で画像に対するラベルを収集した．また，Russell ら [25] はブラウザ上で動作するアノテーションツールを開発し，多数のユーザで協力的にアノテーションを付与および修正することを可能とした．

2.2.2 人手を介さない自動アノテーション付与

画像や映像のアノテーションは，オブジェクトの多クラス識別の問題として考えられる．機械学習の分野では，PASCAL Visual Object Classes (VOC) Challenge や TREC Video Retrieval Evaluation (TRECVID) を筆頭に，データセットの公開，多クラス識別や内容解析手法の性能評価が例年実施され，様々な手法が提案されてきた [26, 27, 28, 29]．これらのデータセットにはオブジェクトラベルやバウンディングボックスなどのメタデータが付与されており，それらを訓練データとして識別器の学習に用いることができる．しかし，事前に用意されたデータセットを用いる場合，学習可能なオブジェクトは提供されているオブジェクトカテゴリに限られるため，実際の画像および映像のオブジェクト検出に適用することは困難である．

一方，任意のオブジェクトの学習を目的に，インターネット画像からのモデル学習手法が提案されている [30, 31]．Chen ら [30] は画像検索の結果からオブジェクトの識別器および，学習した要素同士の類似や包含関係など，視覚的特徴に関する知識を獲得する手法を開発した．また，Fergus ら [31] は画像検索の結果の学習に pLSA を拡張したモデルを用いて，位置や回転といった見えの変化に頑健なオブジェクト検出を試みた．

さらに，これら識別器の学習に加え，オブジェクトの候補領域の抽出 [32] や画像群に共通する要素の位置推定 [33] など，識別の高速化や画像へのメタデータ獲得を目的とした周辺技術も盛んに研究されている．しかしながら，人手を介さず自動的にオブジェクトの識別や位置の特定することは，実用において十分な精度を達成することは困難である [23].

2.2.3 人手による入力を補助的に利用したアノテーション付与

人手によるアノテーション付与は高い精度を得ることが可能だが，労力が膨大になりがちである．また，自動化手法は，人手による作業を必要としない一方で，十分な精度が得られない．これらの特徴を補い合う手法として，人手による入力を補助的に利用した手法 [34, 35] が提案されている．Jain ら [34] は確率的 k-NN 法を提案し，そのカーネルの学習に，識別結果の確度が低いデータから順にユーザがラベル付けをして訓練データに加える能動学習を用いることで，全データのラベル付けを必要としない多クラス識別の精度向上を試みた．また，Sigurbjörnsson ら [35] は Flickr におけるラベルの出現頻度や一枚の画像に付与されるラベル数等を分析し，ユーザが入力した少数のラベルをもとに，候補となるラベルを推薦することで容易なラベルの拡張を試みた．

2.3 本研究の位置付け

人手による映像編集において，撮影された映像の視聴は時間がかかり，また要約映像の長さを考慮した重要箇所選別は，映像編集に習熟していないユーザにとって手間のかかる複雑な作業である．本研究は，このような映像編集をテキストの執筆と単純なアノテーション作業に置き換えることで，一般的なユーザが容易に映像を制作可能な手法を提案する．

本研究は，ユーザが記述したテキストを用いて要約映像を生成する．多くの従来手法は映像の重要性を事前に設計した指標に基づき決定する．一方，提案手法において，映像の重要性はその内容とテキストに表現されるユーザの意図に依存して決定されるものであり，このような研究はこれまでなされていない．また，従来のユーザに関する情報を用いた手法は，ユーザの嗜好や無意識の状態を用いていたが，本研究ではユーザが映像に関して記述したテキストを用いる．これにより，ユーザが明示的に要約映像を制御可能な映像要約手法を実現する．

また，本研究では映像へのオブジェクトのアノテーションを人手により付与する．対象とする映像は多様なオブジェクトを含むため，機械学習を用いた方法で実用上十分な検出精度を達成することは現状では困難である．そこで本研究では，

人手による映像のアノテーション付与にかかる労力を削減するためのインターフェースを提案する．提案インターフェースでは，アノテーション付け対象としてキーフレームを抽出する．次に，映像にオブジェクトとバウンディングボックスの候補をユーザに提示することで，アノテーション付与にかかる手間の削減を試みる．

本論文はまず，テキストを用いてユーザ意図を反映する映像要約手法について 3 章で説明する．そこで，映像中のオブジェクトをアノテーションとして付与するためのインターフェースを提案し，映像のオブジェクトによる表現を与える．また，ショットの内容に基づいたクラスタリングを用い，テキストで言及されているシーンの映像を優先的に選択する方法について述べる．次に，提案手法の評価実験について 4 章で述べる．まずアノテーション付与のための提案インターフェースに関するアンケートの結果を示す．次に，提案手法により生成された映像を評価するユーザスタディについて説明し，考察を述べる．最後に 5 章で本研究をまとめ，今後の展望を述べる．

3. テキストを用いてユーザ意図を反映する映像要約

3.1 提案手法の概要

本研究では，ビデオブログの投稿時に映像に付随するテキストがユーザの意図を表すものと考え，テキストに基づいて要約映像を生成することで，ユーザの意図に沿った要約映像を生成する．以下に，ビデオブログのための要約映像の要件を定める．

(1) テキストで記述された内容に関する映像で構成される．
(2) 要件 (1) を満たす範囲で多様な映像を含む．
(3) 冗長でない．

要件 (1) はビデオブログでの公開を目的とする本研究特有のものであり，ほか 2 つは映像要約手法で広く取り入れられている要件である．

提案手法の概要を図 3 に示す．入力の映像群を V，テキストを T とする．要約映像の長さ L はユーザにより与えられる．映像とテキスト中に出現するオブジェ

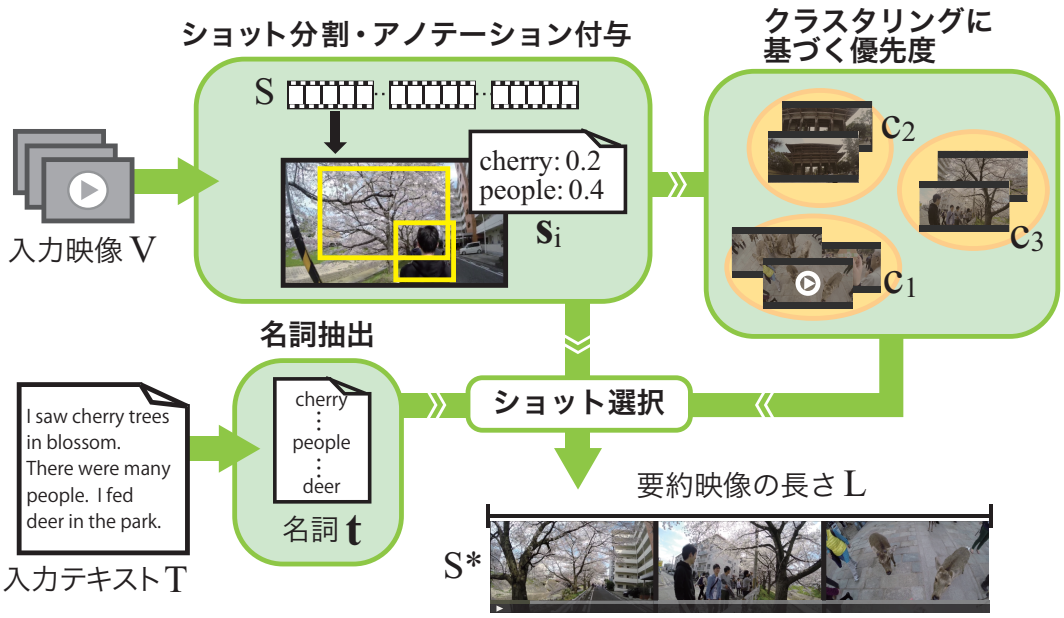

図 3 撮影対象についてのテキストに基づく映像要約の概要．

クトの集合を $O = \{o_1, \ldots, o_N\}$ とし，映像とテキストをこのオブジェクトの集合により表す．提案手法では，まず映像群を短いショットに分割し，ショットの内容を表す特徴として，含まれるオブジェクトのラベルを付与する．一方，テキストにおいては，名詞はオブジェクトに対応すると考え，テキストをオブジェクトの集合として表す．提案手法では基本的に要約映像に含まれるオブジェクト集合がテキストを表すオブジェクト集合と一致するようにショットを選択し，それらを時系列に従って並べることで要約映像を生成する．ここで，ショットは映像の一部のみを表すものであることから，単一のショットに含まれるオブジェクトだけでは，その内容の記述が不十分であると考えられる．そこで本研究では，撮影した時刻や含まれるオブジェクトに大きな隔たりがないショット群を同一シーンで撮影されたものとし，映像に対して付与された撮影時刻とオブジェクト集合に基づいてショットをクラスタリングすることにより，ショットが属するシーンを得る．提案手法はこのシーンを考慮して，ショット群からショットを選択する．具体的には，ショットのクラスタリング結果を考慮した映像とテキストの類似度に基づく目的関数を設計し，この関数を最適化することによりショットを選択する．

3.2 テキスト中の名詞抽出

テキスト T を名詞の集合として表すために，まずテキストから名詞を抽出し，単数形および複数形の表記の差を吸収するために得られた名詞をレンマ化 [36] する．それらの中からテキストの特徴としては過度に一般的な語を辞書を用いて除去する [36]．テキスト T は抽出された名詞を用いて $\mathbf{t} = (t_1, t_2, \ldots, t_N)$ で表す．$t_n = 1$ はオブジェクト o_n に対応する名詞が含まれていることを示し，それ以外は 0 とする．

3.3 ショット分割

本研究では未編集の映像群を対象としており，これらの映像は長時間撮影された映像を含むことがある．そこで要約映像での利用を考慮し，入力映像群 V を短いショット群に分割する．提案手法ではこれらをオブジェクトによって表した

め，オブジェクトが出現，あるいは消失する箇所でショットを分割する．提案手法では隣接するフレームのオブジェクトマッチングに基づくショット分割 [37] を用いる．通常，同一ショット内では，隣接するフレーム間で同一のオブジェクトが含まれるため，それらが記述子によりマッチングされるが，ショットの変わり目ではその前後で写るオブジェクトが変化するため，オブジェクトのマッチングが失敗する．この問題に対して，Huang らの手法 [37] はショット変化とオブジェクトマッチング結果の関係に基づき，オブジェクトがマッチングされない箇所を，隣接するフレーム間の対応する記述子の数を用いて検出し，ショットを分割する．具体的には，映像中の全隣接フレーム間での対応する記述子の数を算出し，その極小値をとる箇所を求める．この極小点をショットの変わり目の候補とし，そこから対応する記述子の数が閾値より多いものなど，ショットの変わり目として不適合なものを除くことでショットの変化を検出する．

3.4 アノテーション付与

続いて，ショットに対してオブジェクトの有無とそのバウンディングボックスをアノテーションとして付与する．ショット分割における仮定から，ショット内でオブジェクトは出現・消失しないものとし，その中央フレームをキーフレームとしてアノテーションを付与する．アノテーション付与のための手法として，文献 [38] の手法など一般物体認識の適用や，3.2 節で得られた名詞をクエリとしたウェブ画像検索結果を用いた検出器の学習 [31] などが考えられる．しかし，本研究が対象とする映像に含まれるフレームは同一フレーム内に多数のオブジェクトが含まれているなど，一般的な物体検出手法が対象とする画像とは異なる性質を持つ．また，既存手法には，出現するオブジェクトの検出器を学習するためのデータセットがノイズを含まず，バウンディングボックスやセグメンテーションなどのメタデータが与えられることを前提とするものもあり [38]，そのようなデータセットを映像要約の度に生成することは現実的ではない．これらのことから，本研究では画像に対してオブジェクトに対応するラベルとそのバウンディングボックスのアノテーションを人手で付与する．

本研究では人手によるアノテーション付けを補助するために，図 4 のようなイ

図 4　アノテーション付与のためのインターフェース．

ンターフェースを開発した．ユーザは表示されたキーフレーム上のオブジェクトを矩形で囲い，そのオブジェクトのラベルを右のリストから選択する．もし，目的のラベルがリストに存在しない場合，ユーザはラベル名を新たに入力することでリストに追加する．

提案インターフェースはこの作業を補助するために，事前に抽出したオブジェクトの候補領域を図 4 のようにユーザに提示する．この候補領域の抽出には，オブジェクトの (i) 閉じた境界を持つ，(ii) 周囲と異なる見た目を持つ，(iii) 顕著性が高い，という 3 つの性質に基づき算出されるオブジェクトらしさの尺度を用いる [32]．またラベルの候補としてユーザが記述したテキストから抽出した名詞を，リストに加える．このように候補領域と候補ラベルを提示することで，ユーザが一から矩形を描画し，ラベルを入力する手間を削減する．

3.5 フレーム内の位置を考慮したオブジェクトの重み付け

提案手法はラベル付けされた各オブジェクトに対して，その重要度を表す重みを付与する．オブジェクトの重要度としては，文献 [39] の手法など，従来提案されてきたさまざまな手法が適用可能であるが，本研究では，中央に大きく写るオブジェクトを重要とみなす単純な尺度を用いる．ここで V に含まれるショットの集合を $S = \{\mathbf{s}_i | i = 1, \ldots, I\}$ とし，$\mathbf{s}_i = (s_{i,1}, \ldots, s_{i,N})$ はショット i に含まれるオブジェクト o_n の重み $s_{i,n}$ により構成される．このバウンディングボックスに含まれる領域 $\Omega_{i,n}$ を用いて，重み $s_{i,n}$ は

$$s_{i,n} = \int_{\Omega_{i,n}} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \mathrm{d}\mathbf{x} \tag{1}$$

により与える．このとき $\mathcal{N}$ はフレーム中央 $\boldsymbol{\mu}$ を平均，$\boldsymbol{\Sigma}$ を分散とするガウス関数であり，$\mathbf{x}$ は画像上の位置を表す.

3.6 ショット選択

提案手法では目的関数 $f(S, \mathbf{t})$ を設計し，それを近似的に最適化することで，要約映像に含まれるショット集合 $S^* \subset S$ を求める．目的関数は，映像とテキストの類似度，およびテキストで記述されたシーンに関する内容の被覆度を用いて以下のように定義する．

$$f(S, \mathbf{t}) = \mathrm{Sim}(S, \mathbf{t}) + \beta \mathrm{Cvrg}(S, \mathbf{t}) \tag{2}$$

ここで β は各項のバランスを調整するためのパラメータである．第 1 項は要約映像とテキストの類似度に，第 2 項はテキストで記述されたシーンに関する内容の被覆度に関する項であり，それぞれ 3.1 節で述べた要約映像の要件 (1)，要件 (2) に対応する．また，第 2 項は被覆度を高めるために様々なオブジェクトを取り入れるため，間接的に要件 (3) にも対応する．前述のように，ショット単体に含まれるオブジェクトのみを用いて，テキストとショットの類似度を算出することは困難であると考えられるため，ショットのクラスタリングに基づく優先度を定義

する．この優先度を考慮した目的関数を近似的に最大化するショットの組み合わせを動的計画法 [40] により得る．

3.6.1 クラスタリングに基づくショットの優先度

ショットとテキストの内容の類似度は，ショット単体に含まれるオブジェクトだけでは判断できない．例えば，ユーザが「川沿いの桜」についてのテキストを入力として与えたとき，入力映像に川沿いで撮影された桜と公園で撮影された桜のショットが含まれる場合，各ショットのオブジェクトラベルだけを考慮すると，この 2 つを区別することができない．この問題に対処するため，ショットをシーンに対応するクラスタに分け，各クラスタに含まれるオブジェクト集合をそのシーンの大局的特徴とし，その特徴がテキストに類似したものを優先する．これにより，ショット単体のオブジェクトのみを考慮した場合よりも，テキストとの内容が近いショットを要約映像に取り入れることが可能となる．

ショットのクラスタリングには Affinity Propagation [41] を用いる．ここでは出現オブジェクトと撮影時刻が近いものは同一シーンを撮影していると考え，ショット間の類似度を $i \neq j$ の場合について以下のように与える．

$$A(\mathbf{s}_i, \mathbf{s}_j) = \exp\left[-\frac{\lambda \min(|\tau_i - \tau_j|, \theta)}{M}\right] + \gamma J(\mathbf{s}_i, \mathbf{s}_j) \tag{3}$$

ここで τ_i はショット i のキーフレーム位置，M は S に含まれる全フレーム数であり，$J(\cdot, \cdot)$ は次式で定義される重み付き Jaccard 係数である．

$$J(\mathbf{s}_i, \mathbf{s}_j) = \frac{\sum_n \min(s_{i,n}, s_{j,n})}{\sum_n \max(s_{i,n}, s_{j,n})} \tag{4}$$

また式 (3) の λ，γ は各項の重みに関するパラメータであり，θ は考慮する撮影時刻の差の最大値である．ショットの自分自身への類似度 $A(\mathbf{s}_i, \mathbf{s}_i)$ はクラスタ数を決定するパラメータであり，$A(\mathbf{s}_i, \mathbf{s}_j)$ の中央値，または最小値が用いられる．クラスタに含まれるショットが少ないと大局的な特徴を得られないため，ここではクラスタ数を抑えるために，文献 [41] の手法に従って $A(\mathbf{s}_i, \mathbf{s}_j)$ の最小値を採用する．ここで，クラスタ k を $\mathbf{c}_k = (c_{k,1}, c_{k,2}, \ldots, c_{k,N})$ で表す．$\mathbf{c}_k$ に含まれるショッ

トがオブジェクト o_n のラベルを付与されるとき $c_{k,n} = 1$，それ以外を 0 とする．このクラスタ $\mathbf{c}_k$ とテキスト $\mathbf{t}$ を用いてショット i の優先度 p_i を $p_i = J(\mathcal{C}_i, \mathbf{t})$ とする．このとき $\mathcal{C}_i$ はショット i が属するクラスタを表す．優先度 p_i はショット i のクラスタのオブジェクト集合 $\mathcal{C}_i$ とテキストのオブジェクト集合 $\mathbf{t}$ が類似するとき高い値をとる．

3.6.2 テキストと要約映像の類似度関数

共通のオブジェクトを多数持つショット集合 S とテキスト T は似た内容を表現すると考え，式 (2) の右辺第二項の $\mathrm{Sim}(S, \mathbf{t})$ を以下のように定義する．

$$\mathrm{Sim}(S, \mathbf{t}) = J(\boldsymbol{\phi}(S), \mathbf{t}) \tag{5}$$

このとき，$\boldsymbol{\phi}(S)$ は S に含まれるショット $\mathbf{s}_i$ をショットの優先度 p_i によって重み付けしたものの和

$$\boldsymbol{\phi}(S) = \sum_{\mathbf{s}_i \in S} p_i \mathbf{s}_i \tag{6}$$

とする．この項はテキスト T に表れるオブジェクトを含むショットを要約映像に取り入れる効果がある．特にクラスタに基づく優先度が高く，かつテキストと共通するオブジェクトを含むとき，そのショットは優先的に選択される．一方で，同一オブジェクトが多く出現する場合，類似度が減少するため，そのようなオブジェクトを含むショットは抑制される．

3.6.3 内容の被覆度

式 (2) の第 2 項は，テキストで記述されたシーンに関するオブジェクトを広く要約映像に取り入れることで，3.1 節の要件 (2) を満たすことを目的としている．また，クラスタリングに基づく優先度が高いショットに含まれているオブジェクトは，テキストのオブジェクトに含まれていなくても，テキストで言及されているシーンに現れていると考えられるので，それらが要約映像に含まれることを許容

アルゴリズム 1 ショット選択アルゴリズム

入力: 映像群 $V = \{\mathbf{s}_1, \ldots, \mathbf{s}_M\}$, テキスト $\mathbf{t}$, 要約映像の長さ L

出力: $S^* \subseteq V$

```
S_{i,0} = {} ∀i = 1, ..., M
for i : 1 to M do
  for j : 1 to L do
    S' = S_{i-1,j}
    S'' = S_{i-1,j-l(s_i)} ∪ s_i
    if f(S', t) > f(S'', t) then
      S_{i,k} = S'
    else
      S_{i,k} = S''
    end if
  end for
end for
S* = arg max_{S_{n,k}} f(S_{n,k})
```

することで要件 (3) の冗長性を抑制する．内容の被覆度は以下のように定義する．

$$\mathrm{Cvrg}(S, \mathbf{t}) = J(\boldsymbol{\phi}(S), \boldsymbol{\psi}(\mathbf{t})) \tag{7}$$

$\boldsymbol{\psi}(\mathbf{t})$ はテキスト T と関連するクラスタに含まれるオブジェクトの集合を表しており，$\boldsymbol{\psi}(\mathbf{t})_n = 1$ はオブジェクト o_n がテキストとの類似度が閾値 ρ 以上のクラスタ集合 $\{\mathbf{c}_k | J(\mathbf{c}_k, \mathbf{t}) \geq \rho\}$ のショットに含まれることを意味する．

3.6.4 目的関数の最適化

提案手法は，文献 [40] を参考に，動的計画法を用いて目的関数 $f(S, \mathbf{t})$ を近似的に最大化するショットを選択する．具体的には，アルゴリズム 1 により f を近似的に最大とする $S^* \subseteq V$ を得る．ここで，$l(\mathbf{s}_i)$ はショット i の長さとする．

4. 実験

本研究では提案したアノテーション用インターフェースと生成された要約映像を評価するためにユーザスタディを実施した.

4.1 データセット

映像データセットとして約 80 分にわたる 42 個の映像を用いた．これらの映像は 1 日の間に撮影され，車内，川辺，公園などのシーンを含む．また入力テキストとして図 5 に示す 3 つのテキストを用いた．テキストはそれぞれ異なる場面に関するものであり，テキスト 1 は川辺での花見，テキスト 2 は鹿の餌付け，テキスト 3 は南大門観光の様子を描写している.

On a warm day in March, we went to Nara Park. Before getting to Nara Park, we went to Saho river. There were cherry trees along the river. The river is well known for cherry blossom, and many people visit during the season of blossom. I took many videos of other students. One of the students, Nakashima used a special camera for his study. He took some videos, carrying the camera along the river. It was a beautiful place and I want to visit there next spring again.

We went to Nara Park. A lot of deer were around the Nandaimon. There were also a few cracker shops, and many tourists enjoyed feeding deer. I bought some crackers and deer immediately gathered around me.

Nandaimon is a famous gate in the Nara Park. I saw a statue of Nandaimon. There were many people.

図 5 入力テキスト．上からテキスト 1，テキスト 2，テキスト 3.

4.2 アノテーションインターフェースの評価実験

提案したアノテーション付与のためのインターフェースを評価するため，ユーザスタディを実施した．本実験において，ユーザはオブジェクトの候補領域を提

表 1 実験条件

	データセット 1	データセット 2
グループ 1	候補領域有り	候補領域無し
グループ 2	候補領域無し	候補領域有り

示する場合と，しない場合の両方を利用して映像にアノテーションを付与し，それぞれの操作に関するアンケートに回答した．

4.2.1 実験環境

実験にはアノテーション付与対象として 10 枚のキーフレームからなるデータセットを 2 つ（データセット 1, データセット 2）用いた．ユーザは表 1 に示すように，2 つのグループ（グループ 1，グループ 2）に分けられ，グループ 1 はデータセット 1 に候補領域が提示されるインターフェース (A) を用い，データセット 2 には候補領域が提示されないインターフェース (B) を利用してアノテーションを付与した．グループ 2 はデータセットを入れ替えて同様の作業をした．このとき，2 つのインターフェースの作業順は各ユーザごとに変更した．本実験では 9 人のユーザに対してユーザスタディを実施し，グループ 1 を 5 人，グループ 2 を 4 人とした．アンケートでは，ユーザがそれぞれのインターフェースについて,（質問 1）自身が撮影した映像のアノテーション付けをする際に使用したいか，1 点（そう思わない）から 5 点（とてもそう思う）で採点した．また,（質問 2）提示した候補領域について煩わしかったか，あるいは便利であったか，1 点（煩わしかった）から 5 点（便利だった）で採点した．

4.2.2 アノテーションインターフェースの評価実験の結果と考察

アンケートの結果を図 6 と図 7 に示す．図 6 の質問 1 に関するグラフは，それぞれのインターフェースが獲得した得点の平均を表す．同図から，候補領域を提示する提案インターフェース (A) が，提示しない場合 (B) よりわずかであるが高い評価を受けたことがわかる．

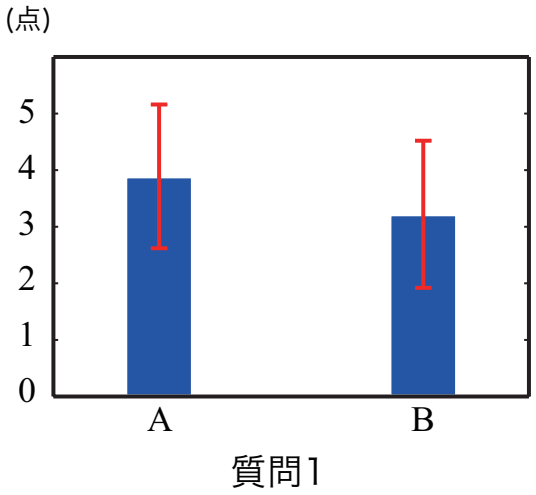

図 6　質問 1 に対する回答の平均値と標準偏差．質問 1 の A は候補領域を提示するインターフェース，B は候補領域を提示しないインターフェースを表す．

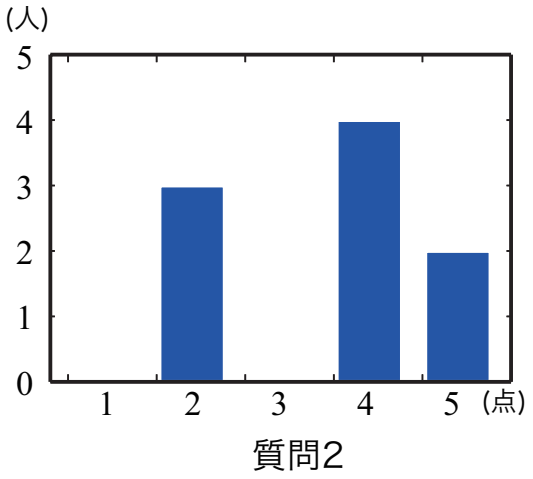

図 7　質問 2 の結果．横軸が候補領域についての点数，縦軸がその点数をつけた人数を表す．

また，図 7 の質問 2 に関するグラフは，それぞれの点数をつけた人数である．この図からわかるように，候補領域に関しては，便利であるという回答が多く得られた．一方で，2 点（たまに煩わしい）とする回答も複数得られた．候補領域の

提示が操作を煩わせる場面として，複数の候補領域が重なる場合，領域の選択や変形等の操作が意図した通りにならないことが報告された．このことはインターフェース上で候補領域を表す矩形の前後関係が明示されていないことが原因と推察され，矩形の表示方法の改善が今後の課題となる．また，リストのラベルが多くなると，目的のラベルを探すために時間がかかることが指摘された．この問題の解決には，他のキーフレームに対してユーザが使用したラベルから，使用可能性の高いラベルを推定し，リストの上位に表示するなどの方法が考えられる．

4.3 ホームビデオを用いた要約映像生成実験

上記のデータセットから提案手法により要約映像を生成した．このとき，要約映像の最大長 $L = 20$ 秒 とし，各パラメータの値は，複数の入力テキストから生成された結果から，経験的に $\mathbf{\Sigma} = \mathrm{diag}(8w, 8h)$，$\beta = 0.25$，$\lambda = 5$，$\theta = 36000$，$\gamma = 0.25$，$\rho = 0.1$ とした．ただし，w，h はフレームの幅と高さである．

図 5 のテキストを用いて提案手法により生成した要約映像を図 8 に示す．テキスト 1 を提案手法に入力として与えた結果，“cherry blossom”や “river”，“camera” などがアノテーションとして付与されたショットが選択された．これらのショットはテキストで描写された場面を撮影したものである．またテキスト 2 を用いた場合にも，“deer”などのオブジェクトがアノテーションとして付与されたショットが選択され，テキストで描写された鹿の餌付けについての要約映像が得られた．本実験で用いたデータセットの映像は，アノテーションがとして “deer”が付与されたショットを複数の場面に含むが，テキストの内容と合致しない場面のショットは要約映像に含まれていない．これはクラスタリングに基づく優先度を考慮したことで，そのようなショットの選択が抑制された結果であると考えられる．ただし，L を長く設定すれば，優先度の低いショットも選択される場合がある．この問題の対処法としては，高い優先度を持つショット全体の長さも考慮し，それよりは短く L を定めるなどの値の調整が考えられる．テキスト 3 を用いて生成された要約映像も同様に，テキストで言及された場面からショットで構成されていることが分かる．これらの結果から，提案手法はテキストで言及された場面からショットを選択可能であることが確認された．

テキスト 1

テキスト 2

テキスト 3

図 8 要約映像に含まれる各ショットのキーフレーム．上からテキスト 1，テキスト 2，テキスト 3 を用いて生成された映像.

表 2 評価に用いた映像の入力テキストと生成手法.

映像	入力テキスト	手法
(a)	なし	定間隔のサンプリング
(b)	なし	クラスタリングに基づく手法
(c)	テキスト 1	提案手法
(d)	テキスト 1	関連内容の被覆度を除去
(e)	テキスト 1	関連内容の被覆度，優先度を除去
(f)	テキスト 2	提案手法
(g)	テキスト 2	関連内容の被覆度を除去
(h)	テキスト 2	関連内容の被覆度，優先度を除去
(i)	テキスト 3	提案手法
(j)	テキスト 3	関連内容の被覆度を除去
(k)	テキスト 3	関連内容の被覆度，優先度を除去

4.4 生成映像の主観評価実験

映像要約手法の評価は，要約映像の明確な真値が存在しないため，困難な問題である．本研究の場合，ユーザによる編集であっても，生成される映像はユーザの技量にも左右されるため，その映像がユーザにとって最良のものとは限らず，評価のための真値映像として用いることは妥当ではない．そこで本研究では，ユーザスタディを実施し，真値映像の代わりに種々のベースライン手法と比較することで提案手法を評価した．ユーザスタディでは，生成された要約映像のビデオブログでの利用を想定し，1. ビデオブログ用映像の作成支援として提案手法が有効であるか，2. テキストに沿った内容の要約映像が実現できているか，の 2 点について，アンケートにより調査した．

4.4.1 実験環境

比較実験に用いた映像とその入力テキストおよび生成手法を表 2 に示す．これらの映像はユーザ意図を考慮しない要約映像 (a)，(b) とテキストを用いてユーザ意図を考慮した要約映像 (c)〜(k) に分けられる．(a) は入力映像を一定間隔でサンプリングした最も単純な要約である．(b) は 3.6.1 項で述べた手法で得られたクラスタの代表ショットから，なるべく多くのオブジェクトを含むようにショット

を選択したものである．映像 (c)，(f)，(i) は提案手法により生成された映像であり，それぞれテキスト 1，テキスト 2，テキスト 3 のユーザ意図を考慮したものである．これら提案手法による映像とユーザ意図を考慮せずに生成された要約映像を比較することで，ビデオブログの制作における提案手法の有効性を確認する．残りの映像も入力テキストに基づき内容が決定されるため，ユーザ意図を考慮している映像に分類されるが，これらは目的関数を一部変更し生成された映像であり，提案手法による出力映像 (c)，(f)，(i) への各項の影響を調査することを目的とする．映像 (d)，(g)，(j) は内容の被覆度 $\mathrm{Cvrg}(S, \mathbf{t})$ を除いたものであり，映像 (e)，(h)，(k) は被覆度に加え，クラスタリングに基づく優先度を定数とすることで無効化したものである．

本実験では，20 人のユーザに対してアンケートを実施した．その際，(表 2) の映像は不規則に並び替えられ，図 9 のようにユーザに提示された．ユーザはアンケート回答中これらの映像を繰り返し視聴できるものとした．

4.4.2 評価項目

本実験では 2 種のアンケートを実施した．一つ目のアンケートでは，要約映像をビデオブログ記事に掲載することを想定し，表 2 の各映像を比較した．その際，被験者は図 1 のようなブログ記事を見て，そこに埋め込む映像としてのふさわしさを表 2 の 11 個の要約映像について 1 点（ふさわしくない）から 5 点（とてもふさわしい）で採点した．その際，異なる入力テキストに対する要約映像の評価を調査するため，被験者を 3 つのグループに分け，それぞれに図 5 のテキスト 1，テキスト 2，テキスト 3 を持つブログ記事を提示し，アンケートを実施した．

二つ目のアンケートでは提案手法の要約映像を，(i) 各テキストの内容をどれだけ良く表しているか，(ii) 内容は冗長でないか，(iii) 全体の映像の内容をどれだけ含んでいるか，の 3 つの観点から調査した．一つ目の質問は提案手法がテキストの内容に合った映像を生成可能か確かめることを目的としており，残りの二つは，従来用いられてきた要約映像の評価指標における，提案手法の性質を調査するためのものである．

図 9　アンケート回答時にユーザに提示された要約映像視聴画面.

4.4.3　評価実験の結果と考察

図 10 に 1 つめのアンケートの結果を示す．同図から，テキスト 2，テキスト 3 を持つブログ記事に関しては提案手法が高い評価を得たことがわかる．一方でテキスト 1 のブログ記事についてはクラスタリングに基づく手法 (b) が最も高く評価された．(b) に含まれるショットのキーフレームを図 11 に示す．映像 (b) は，

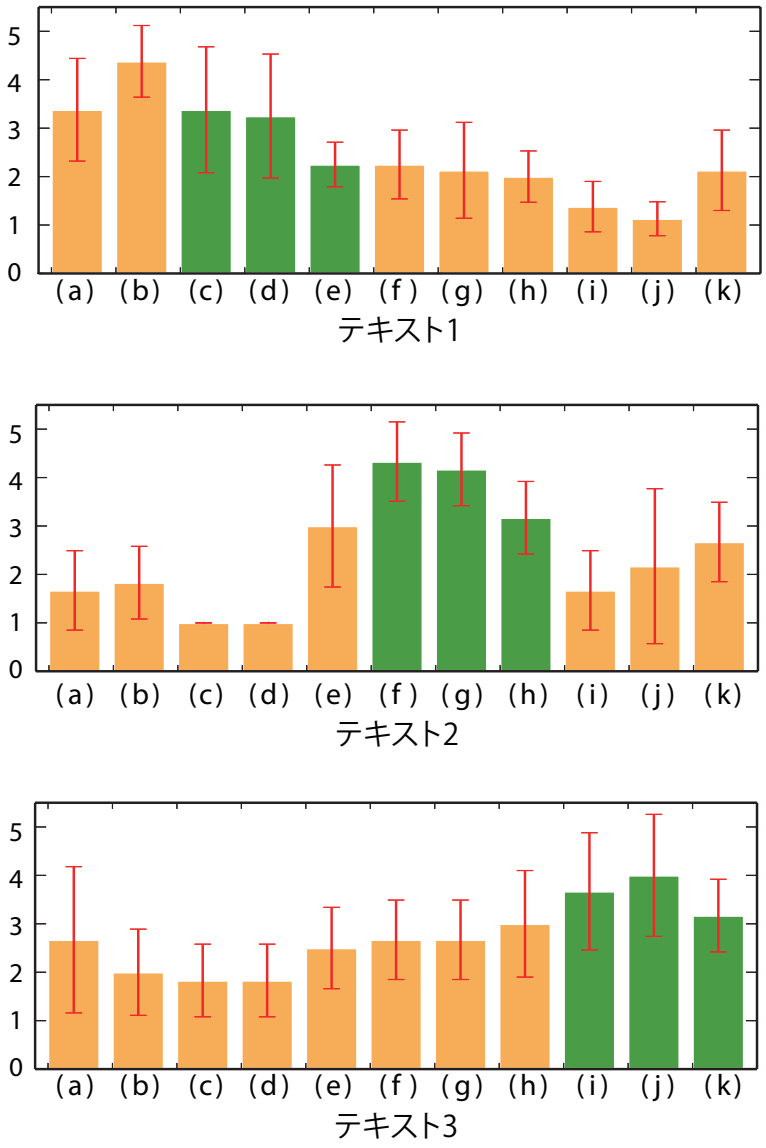

図 10　使用テキスト別のビデオブログ用映像としての評価の平均値と標準偏差. ブログ記事と同様のテキストを入力に用いた要約映像を緑で示す.

図 11　クラスタリングに基づく手法で生成された映像 (b) の各ショットのキーフレーム．

図 11 の (6)，(7) のようにテキスト 1 で言及された場面以外のショットを含む一方で，(3)，(4)，(5) のように，テキスト 1 で記述された川辺と桜のショットを多数含んでいたことが，高評価の要因の一つであると考えられる．さらに (b) には「川辺に向かう移動中」のような，テキスト 1 のシーンの導入となるようなショット (1)，(2) が含まれていた．このようなショットはストーリーを補強する効果があり [13]，ユーザに好印象を与える重要な要因となったと考えられる．また提案手法を一部変更して生成した映像の点数との比較から，関連内容の被覆度の項は評価にほとんど影響しないこと，一方でクラスタリングに基づく優先度がビデオブログに適した映像を生成するために効果的であることが確認された．これらの結果から，ビデオブログ用の映像として提案手法が有効である一方，導入的なショッ

トを取り入れることによる改善可能性が示唆された．

2 つ目のアンケートの結果を図 12，13，14 に示す．図 12 に示すように，質問 (i) の回答に関しては，提案手法が各々の入力テキストに関して高い評価を得ており，適切に入力テキストに対応する映像が生成できていることが確認できた．一方で，図 13，14 からわかるように，冗長性と全体の内容の被覆度に関しては定間隔のサンプリング (a) やクラスタリングに基づく手法 (b) の方が高得点を得た．全体の内容の被覆に関しては，提案手法はユーザが記述したテキストに関するショットを重点的に用いて映像を作成することを目的としているため，全体の内容を広く取り入れることは考えていない．冗長性に関しては，提案手法もテキストの内容に関するショット内で抑制するように試みているが，やはり，要約映像に取り入れるシーンを限定するため，他の手法に比べて類似ショットが含まれやすくなっている．しかし，提案手法はビデオブログのための映像としては十分な評価を得ており，(ii) および (iii) の結果は本研究の目的を損なうものではないと考えられる．

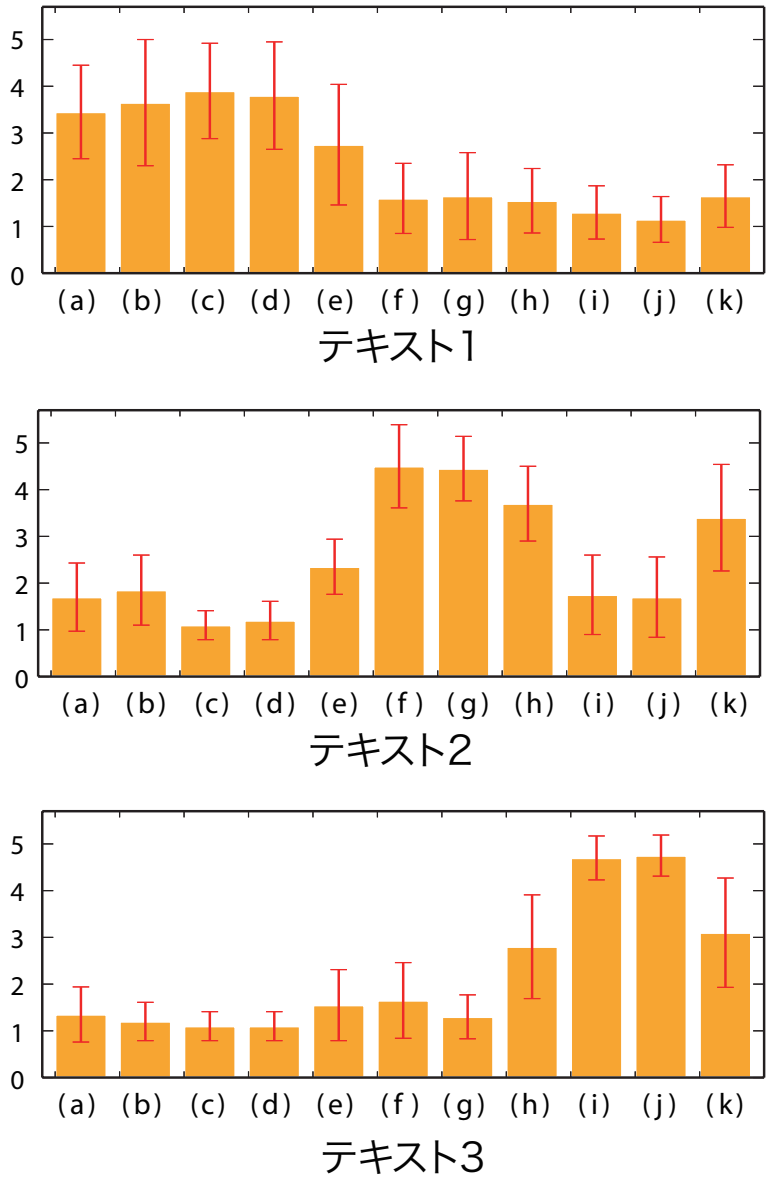

図 12　質問 (i) 映像が表現できているテキストの内容に対するスコアの平均値と標準偏差.

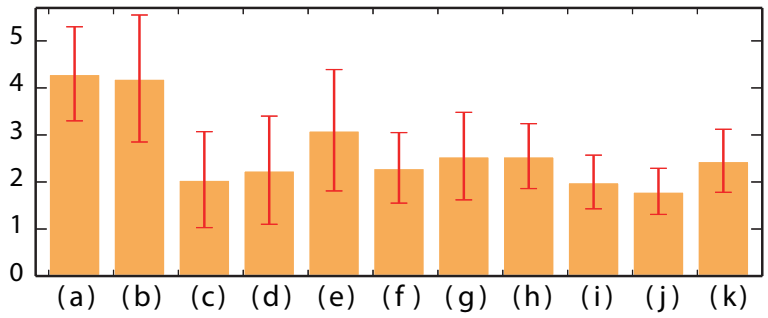

図 13　質問 (ii) 映像の冗長性の低さに対するスコアの平均値と標準偏差.

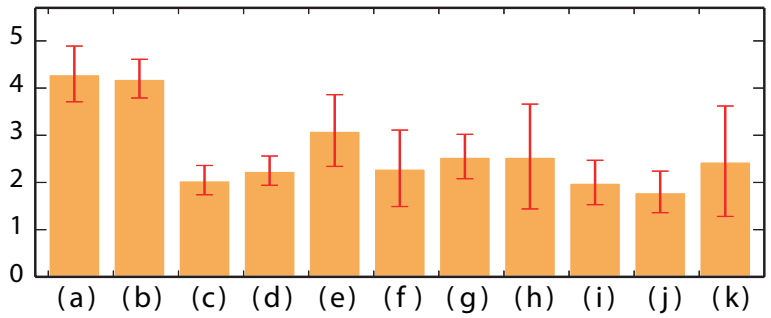

図 14　質問 (iii) 全体の内容の被覆に関するスコアの平均値と標準偏差.

5. まとめ

本研究ではビデオブログのためのテキストを用いてユーザが映像の内容を制御可能な映像要約手法を提案した．そのために，まず映像へのアノテーション付与のためのインターフェースを開発した．提案インターフェースでは，オブジェクトのバウンディングボックスの候補をユーザに提示することで，アノテーション付与にかかる労力の削減を試みた．提案手法は，このようにして得られたテキストと映像のオブジェクト集合を用いて，ビデオブログのための映像が記事のテキストの内容と類似するという考えのもと，目的関数を設計し，その最適化問題を解くことで要約映像を生成した．

本論文では実際のホームビデオを対象に複数の入力テキストを用いて要約映像を生成し，テキストの内容に合った映像が生成可能であることを実験的に示した．また，ユーザスタディを実施し，提案手法がビデオブログのための映像制作に有効であることが示された．この実験結果から，ビデオブログ用映像の評価においては，映像の冗長性の低さや全体の内容をどれだけ網羅できているかといった，従来の要約映像の指標は重要ではなく，映像の内容がビデオブログのテキストと整合していることが重要であることが分かった．また，ショット単体だけでなく，類似したショットをクラスタリングし，その内容がテキストの内容に近いものを優先することで，より適切な場面からショットが選択されることが確認できた．現在，映像要約手法の評価を目的としたテキストと映像の組みからなるデータセットの構築も進んでおり [42]，今後このような一般に公開されたデータセットを用いた評価実験により，提案手法の適用可能性および手法の制約を確認することが必要となる．

一方で，映像のアノテーション付与には課題が残されている．本研究では，映像中のオブジェクトの検出の自動化は実用上困難であると判断し，人手によるアノテーション付与を採用し，そのためのインターフェースを開発したが，要約対象の映像が長くなるほどアノテーション付与にかかる負担は大きくなる．この問題を解決するためには，オブジェクトの候補領域算出の高精度化や，映像中の同一物体の検出によるラベル伝播など，アノテーションインターフェースのさらなる機能の拡充が必要となる．また，テキスト中で同一カテゴリのオブジェクトが

異なる語で表現されることもあり，そのような表記の揺れを考慮したラベル名決定方法も解決すべき課題である．

謝辞

本研究を進めるにあたり，懇切なるご指導，ご鞭撻を賜りました視覚情報メディア研究室 横矢直和 教授に心より感謝致します．また，本研究の遂行にあたり，有益なご助言，ご鞭撻を頂いたインタラクティブメディア設計学研究室 加藤博一 教授に厚く御礼申し上げます．さらに，本研究を進めるにあたり，終始細やかなご指導，ご助言を頂いた視覚情報メディア研究室 佐藤智和 准教授に厚く御礼申し上げます．また，本研究を行うにあたり，多大なるご助言，ご鞭撻を賜った視覚情報メディア研究室 中島悠太 助教に心より感謝致します．特に，中島 悠太 助教には本研究の着想およびテーマ設定から研究の遂行，論文執筆，発表練習その他公私にわたる様々なご指導をいただきました．本研究を進めるにあたり，的確なご助言，ご鞭撻を頂いた視覚メディア研究室 河合紀彦 助教に深く御礼申し上げます．また，研究室生活において様々な支援をして頂いた，視覚情報メディア研究室秘書 石谷由美 女史に厚く御礼申し上げます．また，実験に際して，お忙しい中，被験者として協力してくださいました方々に深く感謝致します．最後に，研究のみならず研究生活全般においてお世話になりました視覚情報メディア研究室の諸氏に深く感謝致します．

参考文献

[1] W. Gao, Y. Tian, T. Huang, and Q. Yang, "Vlogging: A survey of videoblogging technology on the web," *ACM Computing Surveys*, vol. 42, no. 4, pp. 15:1–15:57, 2010.

[2] B. T. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 3, no. 1, 2007.

[3] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank, "A survey on visual content-based video indexing and retrieval," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 41, no. 6, pp. 797–819, 2011.

[4] A. G. Money and H. Agius, "Video summarisation: A conceptual framework and survey of the state of the art," *Journal of Visual Communication and Image Representation*, vol. 19, no. 2, pp. 121 – 143, 2008.

[5] Y. Gong and X. Liu, "Video summarization using singular value decomposition," *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. 174–180, 2000.

[6] B. Zhao and E. P. Xing, "Quasi real-time summarization for consumer videos," *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2513 – 2520, 2014.

[7] D. DeMenthon, V. Kobla, and D. Doermann, "Video summarization by curve simplification," *Proceedings of ACM International Conference on Multimedia (MM)*, pp. 211–218, 1998.

[8] R. Laganière, R. Bacco, A. Hocevar, P. Lambert, G. Païs, and B. E. Ionescu, "Video summarization from spatio-temporal features," *Proceedings of ACM TRECVID Video Summarization Workshop*, pp. 144–148, 2008.

[9] Y.-F. Ma, L. Lu, H.-J. Zhang, and M. Li, "A user attention model for video summarization," *Proceedings of ACM International Conference on Multimedia (MM)*, pp. 533–542, 2002.

[10] C. Taskiran, Z. Pizlo, A. Amir, D. Ponceleon, and E. Delp, "Automated video program summarization using speech transcripts," *IEEE Transactions on Multimedia*, vol. 8, no. 4, pp. 775–791, 2006.

[11] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1553–1568, 2013.

[12] C.-W. Ngo, Y.-F. Ma, and H.-J. Zhang, "Video summarization and scene detection by graph modeling," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 2, pp. 296–305, 2005.

[13] Z. Lu and K. Grauman, "Story-driven summarization for egocentric video," *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2714–2721, 2013.

[14] N. Babaguchi, Y. Kawai, T. Ogura, and T. Kitahashi, "Personalized abstraction of broadcasted american football video by highlight selection," *IEEE Transactions on Multimedia*, vol. 6, no. 4, pp. 575–586, 2004.

[15] F. Wang and C.-W. Ngo, "Rushes video summarization by object and event understanding," *Proceedings of the International Workshop on TRECVID Video Summarization*, pp. 25–29, 2007.

[16] Y. J. Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization," *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1346–1353, 2012.

[17] R. Hong, J. Tang, H.-K. Tan, C.-W. Ngo, S. Yan, and T.-S. Chua, "Beyond search: Event-driven summarization for web videos," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 7, no. 4, pp. 35:1–35:18, 2011.

[18] J. Sang and C. Xu, "Character-based movie summarization," *Proceedings of ACM International Conference on Multimedia (MM)*, pp. 855–858, 2010.

[19] A. Khosla, R. Hamid, C.-J. Lin, and N. Sundaresan, "Large-scale video summarization using web-image priors," *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2698–2705, 2013.

[20] T. Syeda-Mahmood and D. Ponceleon, "Learning video browsing behavior and its application in the generation of video previews," *Proceedings of ACM International Conference on Multimedia (MM)*, pp. 119–128, 2001.

[21] K. Aizawa, K. Ishijima, and M. Shiina, "Summarizing wearable video," in *Proceedings of IEEE International Conference on Image Processing (ICIP)*, vol. 3, 2001, pp. 398–401.

[22] M. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's mechanical turk a new source of inexpensive, yet high-quality, data?" *Perspectives on Psychological Science*, vol. 6, no. 1, pp. 3–5, 2011.

[23] M. Wang, B. Ni, X.-S. Hua, and T.-S. Chua, "Assistive tagging: A survey of multimedia tagging with human-computer joint exploration," *ACM Computing Surveys*, vol. 44, no. 4, pp. 25:1–25:24, 2012.

[24] L. Von Ahn and L. Dabbish, "Labeling images with a computer game," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pp. 319–326, 2004.

[25] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: a database and web-based tool for image annotation," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 157–173, 2008.

[26] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[27] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," *Proceedings of ACM International Workshop on Multimedia Information Retrieval (MIR)*, pp. 321–330, 2006.

[28] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," arXiv:1409.0575, 2014.

[29] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3485–3492, 2010.

[30] X. Chen, A. Shrivastava, and A. Gupta, "NEIL: Extracting visual knowledge from web data," *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pp. 1409–1416, 2013.

[31] R. Fergus, L. Fei-Fei, and P. Perona, "Learning object categories from google ' s image search," *Proceedings of International Conference on Computer Vision (ICCV)*, vol. 2, pp. 1816–1823, 2005.

[32] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object?" *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 73–80, 2010.

[33] K. Tang, A. Joulin, L.-J. Li, and L. Fei-Fei, "Co-localization in real-world images," *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1464–1471, 2014.

[34] P. Jain and A. Kapoor, "Active learning for large multi-class problems," *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 762–769, 2009.

[35] B. Sigurbjörnsson and R. Van Zwol, "Flickr tag recommendation based on collective knowledge," *Proceedings of International Conference on World Wide Web (WWW)*, pp. 327–336, 2008.

[36] S. Bird, "NLTK: The natural language toolkit," *Proceedings of Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING-ACL)*, pp. 69–72, 2006.

[37] C.-R. Huang, H.-P. Lee, and C.-S. Chen, "Shot change detection via local keypoint matching," *IEEE Transactions on Multimedia*, vol. 10, no. 6, pp. 1097–1108, 2008.

[38] T. Malisiewicz, A. Gupta, and A. A. Efros, "Ensemble of exemplar-svms for object detection and beyond," *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2011.

[39] Y. Nakashima and N. Yokoya, "Inferring what the videographer wanted to capture," *Proceedings of IEEE International Conference on Image Processing (ICIP)*, pp. 191–195, 2013.

[40] R. McDonald, "A study of global inference algorithms in multi-document summarization," *Proceedings of the European Conference on IR Research (ECIR)*, pp. 557–564, 2007.

[41] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, pp. 972–976, 2007.

[42] S. Yeung, A. Fathi, and L. Fei-Fei, "Videoset: Video summary evaluation through text," arXiv:1406.5824, 2014.