DEIM Forum 2019 G5-5

写真付き手順書生成のための実施映像からのフレーム選択

西村　太一† 　橋本　敦史†† 　山肩　洋子††† 　森　信介††††

† 九州大学芸術工学部芸術情報設計学科　〒 811–0032 福岡県福岡市南区塩原 4 丁目 9 番 1 号
†† オムロン サイニックエックス株式会社　〒 113–0033 東京都文京区本郷 5 丁目 24 番 5 号
††† 京都大学 情報学研究科　〒 606-8501 京都府京都市左京区吉田本町
†††† 京都大学 学術情報メディアセンター　〒 606–8501 京都府京都市左京区吉田本町
E-mail: †nishimura.taichi.104@s.kyushu-u.ac.jp, ††atsushi.hashimoto@sinicx.com,
†††{yamakata,forest}@i.kyoto-u.ac.jp

あらまし　作業の手順を伝える上で，テキストは分量や時間などを正確に伝えることができる重要な表現形式である．一方で，数値化しにくい作業の目標状態などについては，画像などによる直感的な情報伝達の方が優れている場合もある．本研究では，調理を題材として作業者の内容理解・手順実施補助のための写真付き手順書を得ることを目標とする．入力として手順書と未編集の作業実施映像を与え，作業内容を伝達するのに適したフレームを映像から手順ごとに選択する．我々は，手順書と実施映像のそれぞれに現れる手順実施の上で重要な物体が共起するときが手順書の各文と動画フレームの組み合わせとして相応しいと仮定し，お互いをその物体に基づいて特徴ベクトルに変換し，手順と動画フレームの間での相応しさとして余弦類似度を計算した．さらに，物体に視覚的な変化がある時が手順実施の上で重要なシーンであると考え，動画フレームごとにシーンの重要度を加えたものをスコアとし，Viterbi アルゴリズムを用いて写真付き手順書を生成する手法を提案する．また，2 名の被験者に各レシピ文に付与していて相応しいと感じられる動画フレームを選んでもらうことで作成した正解データと比較し，提案手法の性能を評価した．

キーワード　食, フレーム選択, マルチモーダル

1 はじめに

作業の手順を伝える上で，テキストは分量や時間などを正確に伝えることができる重要な表現形式である．一方で，数値化しにくい作業の状態などについては，画像などによる直感的な情報伝達の方が優れている場合もある．本研究の目的は，調理を題材として作業者の内容理解・手順実施補助のための写真付き手順書を得ることである．入力として手順書と未編集の作業実施映像を与え，作業内容を伝達するのに適した動画フレームを映像から各手順ごとに選択する．読者は生成された写真付き手順書を見ることによって，より正確に手順内容を理解できるようになることが期待される．

本研究では，手順書と実施映像のそれぞれに現れる手順実施の上で重要な物体が共起するときが手順書の各手順に映像中の動画フレームが付与されていて相応しい組み合わせになると仮定した．この仮定に基づき，手順書からは固有表現認識器を用いて，実施映像からは物体認識技術を用いて手順実施のための重要な物体を抽出した．そして，それをもとに双方を特徴ベクトルに変換し，余弦類似度を計算することで，手順と映像の動画フレームの間での相応しさを計算した．この相応しさの値を，本研究では手順と映像の動画フレームとの間の適合度と呼ぶこととする．この適合度をもとに，Viterbi アルゴリズムを用いて手順書と調理動画の間で最も組み合わせとして相応しいと考えられる組み合わせを求める．

実験では，KUSK Dataset [1] の調理動画を実施映像として利用する．調理動画に対応するようにレシピがそれぞれ存在し，これを手順書として用いた．

本稿は以下のように構成する．第 2 章で関連研究について述べる．第 3 章で提案手法について述べ，第 4 章で提案手法について定量的に評価を行う手法を検討した後に評価を行い，結果を考察する．第 5 章でまとめと今後の課題について述べる．

2 関連研究

本章では，関連研究として，まず，自然言語と画像や動画を統合する研究全体について述べる．次に，自然言語を手順書とし，動画をその実施映像とすることで，それらのアライメントを獲得する研究について述べる．

2.1 言語と画像・映像間との複数のモダリティの統合

自然言語と画像や動画などを統合する研究は深層学習の発展により近年活発に行われている．たとえば，画像や動画を入力としてその内容をキャプション生成する研究 [2] [3] や，画像や動画を言語と同じ潜在空間に埋め込む研究 [4] などがある．それぞれの手法を調理動画・画像とレシピに応用した研究も行われている．牛久ら [5] は調理動画を入力として，レシピを出力するモデルを提案した．深層学習のための調理動画とレシピの対が大量に存在しないため，物体認識とキャプション生成を別々に行い，レシピの生成を行う手法を試みている．また，Salvador

ら [6] はレシピ・材料と調理後の完成写真を同じ潜在空間に埋め込むモデルを提案し，調理画像とレシピの検索タスクにおいて既存のモデルを大きく上回る精度を実現した．

2.2 手順書と実施映像の間でのアライメント獲得

手順書と実施映像を入力として，手順を説明する上で相応しい画像を実施映像から抽出して組み合わせる先行研究として，Malmaud ら [7] の研究がある．この研究では，まず，YouTube 上の調理動画で，説明文にレシピ全文が記述してあるものを大量に収集した．次に，各動画から音声認識を用いて各動画の音声を文章で書き起こし，各レシピ文と動画中で認識した文章の間で隠れマルコフモデルを用いてモデル化しアライメントを獲得した．最後に，動画中の視覚的情報を用いることで，より洗練されたアライメントをレシピ文と動画フレームの間で獲得する手法を提案した．この研究では，レシピ文と調理動画内での作業内容の発話と調理動画が必要であるのに対し，我々の研究では入力として与えるのは調理動画とレシピ文のみで，音声を必要としない点が異なる．さらに，実施映像が編集済みであり，本研究に用いるものは未編集である点も異なる．

Naim ら [8] は未編集の実施映像として動作ごとにあらかじめ区切られた映像と手順書を入力として，手順書に現れる物体に着目し，隠れマルコフモデルと統計的機械翻訳のモデル [9] を組み合わせて用いて各手順と実施映像の一区切りとの間，また各手順の物体と実施映像内の物体との間でアライメントを獲得する手法を提案している．この研究では，映像をあらかじめ動作単位で区切り，その区間と手順書の手順の組み合わせを得ている．一方，本研究ではそのような区切りを入れずに実施映像と手順書を入力としている点が異なる．

Bojanowski ら [10] の研究では，未編集の実施映像と手順書に加えて，各手順の実施映像中の範囲をアノテーションしたものも入力として与え，手順書の各手順と実施映像の各フレームをそれぞれ特徴ベクトルへ変換した後，各手順がどの実施映像の範囲を指すかを教師ありで学習させる手法を提案している．先行研究では直接各手順が実施映像のどの範囲に当たるかをアノテーションして学習しているが，本研究の提案手法ではこのようなアノテーションを必要としない点が異なる．

3 提 案 手 法

本研究の目的は入力として手順書とその自然な実施映像を与え，手順書の内容理解・作業実施補助のための写真付き手順書を得ることである．この目的を果たす上で問題となるのは以下の 3 点である．

(a) 映像の動画フレームと手順文がどれくらい付与されていて相応しいかを測定するための適合度の定義

(b) 未編集・発話なしの実施映像において各手順にとって重要なシーンの推定方法

(c) 実施映像・手順書全体を通して最も相応しい動画フレームと手順文の対応づけを選択するための手法

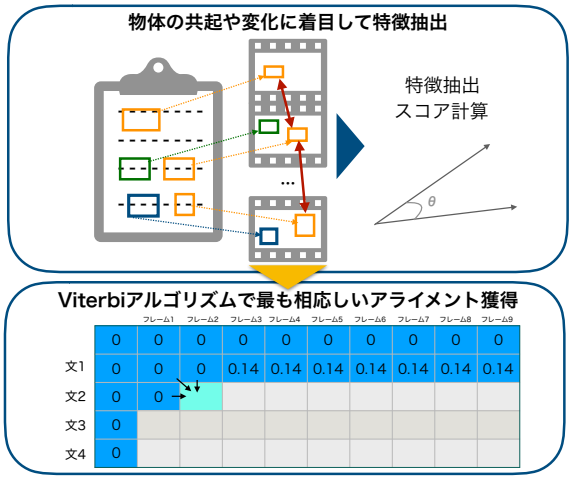

図 1 手 法 概 要

この問題点を解決するための手法を図 1 に概要を示す．(a) を解決するために，本研究では手順書の各手順と写真の適合度を以下のように計算した．まず，手順文中に現れる手順を実施するうえで必要となる重要な物体が映像中に現れる時の動画フレームがその手順文にとって相応しい動画フレームであると仮定し，手順文と動画フレームを表層の情報に着目して特徴ベクトルに変換した．この特徴ベクトルを用いて余弦類似度を計算し，これを適合度とした．次に，(b) を解決するために，物体の状態の変化が大きい時が調理を実施する上で重要なシーンと考えた．そして，物体の変化を訓練済みニューラルネットの出力のユークリッド距離とすることで，調理各シーンごとに調理を実施する上でのシーンの重要度を計算した．最後に，(c) を解決するために，Viterbi アルゴリズムを利用する．(a) で求めた手順と動画フレーム間の適合度と，(b) で求めたシーンの重要度をスコアとし，手順書と実施映像全体を通してスコアの合計が最大となるようなアライメントを獲得することで，各手順文に対応する写真を選択した．

3.1 前 処 理

まず，動画フレームと手順文を特徴ベクトルに変換するために，それぞれの表層に現れる特徴に基づいて双方を特徴ベクトルに変換した．この手法を図 2 に示す．手順文からは，後述する手順の固有表現を抽出し，特徴ベクトルに変換する．また，映像側から表層に現れる特徴を得るために，映像を動画フレームの連続として捉え，各動画フレームごとに物体認識を行った．物体認識の結果をもとに，動画フレームに現れる物体を特徴ベクトルとして変換した．

3.1.1 手順書の特徴ベクトルへの変換方法

本研究では，手順書・実施映像の例として調理を題材に扱うこととする．手順書から調理を実施するうえで重要な物体を抽出するために，まず，我々は調理分野特有の固有表現としてレシピ固有表現（r-NE）を定める [13]．これは通常の固有表現と同様に単語列とタグの組で表現される．特にタグについては調理に関係したものに限定している（表 1）．r-NE タグは固有表現認識器である PWNER [14] を用いて認識した物体，レシピ

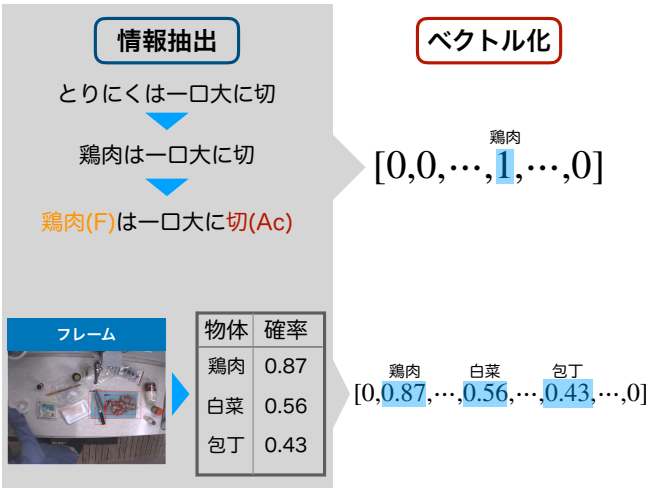

図 2　前処理概要

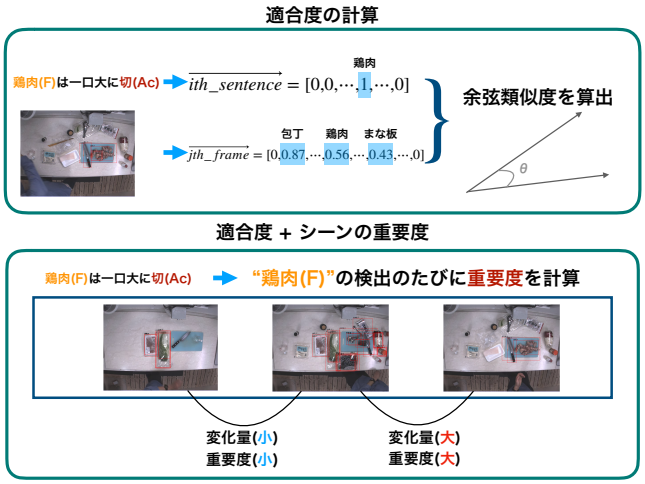

図 3　スコア計算手法

文にそれぞれ自動的に付与することも可能であり，r-NE タグのうち，F と T に限定すれば 95%の精度で認識することができる．しかし，本研究ではレシピの各単語に人手でアノテーションされているものを利用することを前提とする．

表 1　r-NE タグ一覧

r-NE タグ	意味
F	食材
T	道具
D	経過時間
Q	分量
Ac	調理者の動作
Af	食材の動作
Sf	食材の状態
St	道具の状態

次に，手順書をレシピ固有表現の中の動作 (Ac) ごとに文章を分割する．これは，1 行のレシピ文には複数の動作が存在していることもあり，どの動作にどの動画フレームを結びつけるべきか不明確なこともあるためである．分割して得られた分割単位のそれぞれを指示句と呼ぶこととし，この各指示句と動画フレームとの間でアライメントを獲得することとする．さらに，レシピ内の料理用語の表現と，認識カテゴリの表現を合わせるために，料理オントロジー [15] を用いて正規化を行なう．

最後に，各指示句をそれぞれ特徴ベクトルに変換する．ただし，このプロセスで利用する r-NE は F と T のみである．なぜなら，物体認識によって認識できるのは F と T の物体のみであるので，動画からは F と T 以外の情報を抽出することができないためである．よって，レシピ側は各レシピ文中の r-NE の物体を指すインデックスの指す値が 1，それ以外が 0 となるように変換する．

3.1.2　実施映像の特徴ベクトルへの変換方法

動画フレームから特徴ベクトルを得るために，まず，Faster R-CNN [11] を用いて，各動画フレーム中の物体認識を行う．これによって各動画フレーム中に存在すると推定される物体名，物体の認識の確度，物体の動画フレーム中の位置を長方形で獲得する．物体認識をするタイミングは，作業内容が変更されたと考えられる調理者が物体を手に取ったり置いたりした動画フレームとし，これをキーフレームとして残し，それ以外の動画フレームは除外する [12]．次に，物体認識の結果をもとに，特徴ベクトルへ変換する．各動画フレーム内に認識された r-NE の物体を指すインデックスの指す値に認識した r-NE の確率値を代入し，その動画フレーム内で認識しなかった物体のインデックスには 0 を入れるように変換する．

3.2　スコアの計算方法

次に，上記で得られた特徴ベクトルをもとに指示句と動画フレームの間の相応しさを計算する手法を検討する．図 3 に概要を示す．

3.2.1　余弦類似度

指示句中に現れる物体が動画中に現れる時の動画フレームがその指示句にとって相応しい動画フレームであると仮定し，前述した方法で得られた特徴ベクトルをもとに，余弦類似度を用いて指示句と動画フレームの適合度を計算する．

3.2.2　シーンの重要度の考慮をしたスコアの計算

しかし，余弦類似度のみでは，物体確度が高いものに大きく影響を受けることが考えられる．たとえば，「ねぎをフライパンで炒め」という指示句に対して，求めたい動画フレームはフライパン上で炒められているねぎであるが，フライパンで炒められているねぎを物体認識すると認識率が下がるため，その結果ねぎが原型で残っているフレームを選ぶという問題点がある．そのため，各レシピ文中の F と T を動画中で見た目が変化した時が調理をする上で重要なシーンと考え，重要なシーンに高いスコアを割り当てることで，シーンの重要度を加味したスコアを手法も併せて提案する．シーンの重要度を計算するには，各レシピ文中に含まれる F と T を各動画フレームで再検出の度に訓練済みニューラルネットワークを用いて特徴ベクトルを抽出し，そのユークリッド距離を物体の形状の変化として求める．

3.3　アライメントの獲得方法について

上述したスコアをもとに，Viterbi アルゴリズムを用いてレ

シピ文と動画フレーム間でスコアの合計が最大になるようなパスを探索する．前向きプロセスでは漸化的にテーブルの値を以下のように更新し，後ろ向きプロセスでスコアの和が最大となるパスを再帰的に辿ることで，アライメントを獲得することができる．

$$table[i][j] = \begin{cases} table[i-1][j-1] + score & (score > 0) \\ \max\{table[i-1][j], table[i][j-1]\} & (otherwise) \end{cases}$$

4 実験と考察

KUSK Dataset のレシピ文と作業実施映像を与え，前章で提案した手法を実行し，写真付き手順書を生成した．本実験で実行した手法は以下の 3 つである．

(a) cos

(b) cos+viterbi

(c) cos+viterbi+scene

(a) の cos は，全指示句と動画フレームをそれぞれ特徴ベクトルに変換し，各指示句ごとに最も余弦類似度が高い動画フレームを選択する手法である．(b) は，前章の 3.2.1 で計算した適合度をスコアとして，3.3 を実行したものである．つまり，指示句と動画フレームとの間で余弦類似度を計算し，これをスコアとして Viterbi アルゴリズムを実行することによって，各指示句と動画フレームとの間でアライメントを獲得する方法である．(c) は，前章の 3.2.1 で計算した適合度に，3.2.2 で計算したシーンの重要度を加算してスコアとし，3.3 を実行したものである．つまり，(b) で計算したスコアに，シーンの重要度を足し合わせて改めてスコアとし，Viterbi アルゴリズムを実行する手法である．

次に，提案した手法の性能を評価するために，指示句とそれに付与している写真の組み合わせがどの程度相応しいかを測定する方法を検討する．提案手法によって，指示句と指示句に付与した写真の組み合わせと，人手で指示句に付与した写真の組み合わせが同じである数が多ければ多いほど，提案手法の性能がよいと考えた．そのために，まず，2 名のアノテータに，指示句に対して，その指示句を読んで付与されていて相応しいと思われる写真をキーフレームの中から選択してもらった．アノテータ間でアノテーションの基準がぶれないように，以下の基準を設けた．

- 各指示句に対して，付与されていて相応しいと感じられる写真を選ぶこと．
- 複数選択可能である．また，該当する写真がなければ，該当なしも選んでもよい．
- 画面の中央で動作が行われている写真を選ぶこと．画像の端で指示句の動作が行われているものは選んではならない．

次に，提案手法で付与した写真がアノテータの選んだ写真の中にある数を数え，指示句の数で割ることで正解率を計算した．この指標では，アノテータが一枚も写真を選択しない場合は，提案手法も写真を選択しないことで正解とする．さらに，価値のある動画フレームをどの程度当てられているかを調べるために，再現率も計算した．この指標は，アノテータが写真を少なくとも 1 枚以上付与した指示句の中で，提案手法が付与した写真がアノテータの選んだ写真の中にある割合に相当する．計算式は以下の通りである．

$$\text{正解率} = \frac{\text{写真付与に正解した指示句数}}{\text{全指示句数}}$$

$$\text{再現率} = \frac{\text{写真付与に正解した指示句数}}{\text{1 枚以上写真が付与されている指示句数}}$$

4.1 評価手法の妥当性の検討

次に，上述した評価手法が適切であるかを検討する．評価手法が適切であるためには，人手でアノテーションしたもののぶれが少ない必要がある．それを確かめるために，以下の予備実験を行った．まず，2 名のアノテーション結果のうち，一方を自動付与した写真付き手順書とみなし，もう一方を正解の写真付き手順書とみなした．そして，2 つの写真付き手順書を比較して正解率と再現率を計算した．それを交互に繰り返し，平均の正解率と再現率を計算した．その結果，平均正解率は 75.1%，再現率は 74.8%となった．以上の結果より，2 人の付与した写真はある程度共通していると考えられ，これと比較することで，提案手法の性能を評価できると考えられる．よって，評価手法の妥当性を確かめることができた．また，自動生成した写真付き手順書の評価結果と予備実験の結果を比べることで，手法の性能を評価できると考えられる．

4.2 統 計 情 報

2 名のアノテータのアノテーション結果を集計した．その結果を表 4.2 に示す．レシピの種類は 7 種類で，撮影した動画は 12 種類ある．ハイフンの前の数字がレシピの種類を表し，ハイフンの後の数字は同じレシピの中で調理動画を識別するために付与した値である．全ての指示句の数は 198 文あり，2 名のアノテータによってそれぞれにアノテーションを行ったため，アノテーション結果の総指示句数は 396 文となる．このうち，写真が少なくとも 1 枚付与された指示句は 255 文であり，一枚も写真が付与されなかった指示句は 141 文である．

ID : 2-1, 3-1, 3-2 については，半数以上の指示句に 1 枚も写真が付与されなかった．その他のレシピについては，半数以上の指示句に少なくとも 1 枚は写真が付与された．

	1-1	1-2	2-1	3-1	3-2	4-1	4-2	5-1	5-2	6-1	6-2	6-3
指示句数	18	18	19	11	11	26	26	15	15	13	13	13
総指示句数	36	36	38	22	22	52	52	30	30	26	26	26
写真が付与された指示句数	22	29	15	10	11	28	37	19	27	22	17	18
「該当なし」数	14	7	23	12	11	24	15	11	3	4	9	8

表 2　アノテーションの統計情報

また，図 4 に 1 つの指示句に対して同時に付与された写真の枚数の分布を示す．最も多いのは「該当なし」の 141 文であ

る．最も写真が付与された指示句には 42 枚の写真が付与されていた．

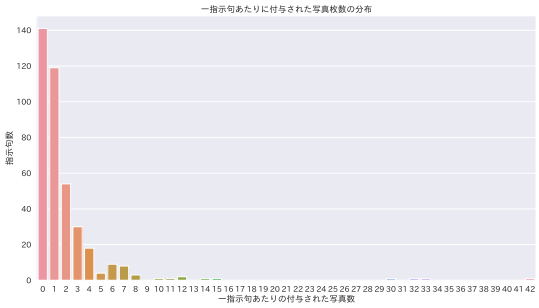

図 4　1 つの指示句に付与されたアノテーション枚数の分布

4.3　提案手法の実験設定

Faster R-CNN の学習データとして，KUSK Object Dataset [12] を用いた．これは各レシピに対応する調理の作業動画を 3 つのカメラで撮影したものについて，食材や調理道具についてアノテーションを行なったものである．また，訓練済みのニューラルネットワークとして，ImageNet で学習した PyTorch の ResNet50 [16] を用いた．

4.4　結　　果

本研究では評価する手法として，3 つの手法を実行した．これらの手法と予備実験の結果を併せて図 5，図 6 にその結果を示す．正解率の平均値は余弦類似度のみの場合は 23%，余弦類似度に Viterbi アルゴリズムを組み合わせ場合が 27%，さらにシーンの重要度を計算して組み合わせた場合は 30%となった．また，再現率の場合は，それぞれ 4%，9%，13%となった．なお，4.1 節で測定した人間同士の正解率・再現率はそれぞれ 75.1%・74.8%であり，図中の human という項目に該当する．

アノテーターは 1 つの指示句に対して，写真を付与しないことも，一枚以上選択することもできる．アノテーターの付与した写真の枚数に応じて提案手法の正解率がどう変化するのかを明らかにするため，1 つの指示句にアノテーションされた枚数に対する正解率を測定した．図 7 にその結果を示す．

1 つの指示句に対して，写真が付与される種類の枚数が 15 枚以上になる場合は少ないため，グラフは滑らかにならない．8 枚以上では同じようになるため，図 7 のグラフは 15 枚を最大としている．

4.5　考　　察

以上の結果より，人間の結果と比較して，提案手法の性能は人手の結果よりも下回る結果となった．本節では，人手によって付与された写真と，提案手法によって付与された写真を見比べることで，提案手法の長所と短所を定性的に考察を行う．

4.5.1　アノテーション枚数と正解率の変化

アノテーターの付与した写真の枚数に応じて提案手法の正解率がどう変化するのかを明らかにするため，1 つの指示句にアノテーションされた枚数に対する正解率を測定した (図 7)．

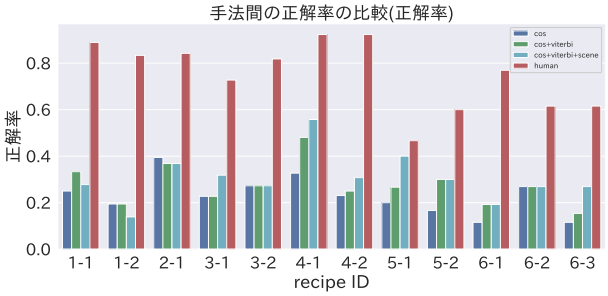

図 5　正　解　率

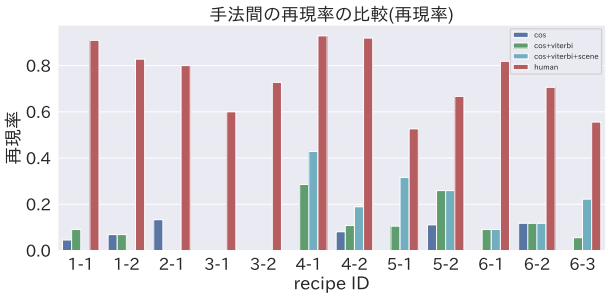

図 6　再　現　率

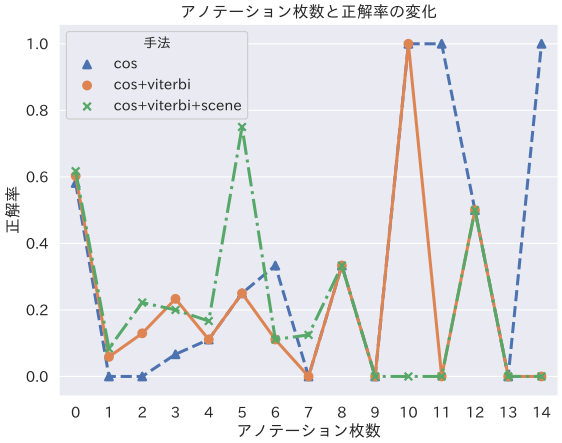

図 7　アノテーション枚数につき正解率

我々は，1 つの指示句に対して，アノテーション枚数が増えるに従って正解率が高くなると仮定した．しかし，提案手法の正解率の変化を見比べる限り，余弦類似度に Viterbi を組み合わせた手法と，さらにシーンの重要度をスコアに組み合わせた手法との間での差は大きくないと考えられる．

4.5.2　提案手法が写真付与に成功するフレームの考察

前節の結果を参照すると，指示句と動画フレームの相応しさの尺度を余弦類似度で定義し，Viterbi アルゴリズムと組み合わせることによって，正解率，再現率共に性能の向上がみられた．よって，Viterbi アルゴリズを用いることで，レシピ全体を通しての相応しさが高くなるような組み合わせを得ることができたと言える．

さらに，シーンの重要度を計算し，スコアとして余弦類似度に加え入れることで，正解率，再現率の性能が向上した．その

図 8　シーンの重要度を加味することの影響例

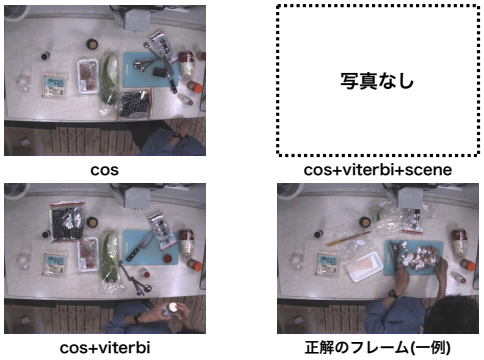

図 9　Faster-RCNN が検出に失敗するケース

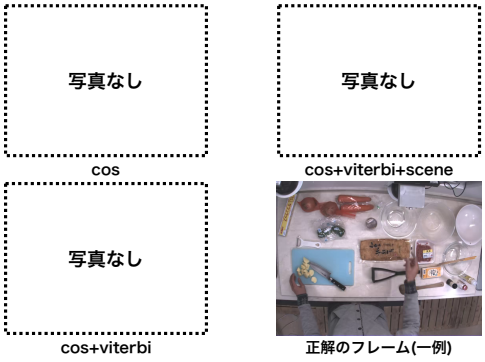

図 10　テキスト上に物体の名称が省略されるケース

要因として，今まで検出はできていたが，組み合わせとして選ばれることのなかった Faster-RCNN の検出確度が小さい動画フレームもシーンの重要度を加味すると組み合わせの候補になったためであると考えられる．図 8 にその例を示す．左から順に余弦類似度のみ，余弦類似度と Viterbi アルゴリズムを組み合わせたもの，そしてさらにシーンの重要度を計算して加味したもの，人手によってアノテーションされた写真の順である．

4.5.3　提案手法が写真付与に失敗するフレームの考察

図 5 と図 6 を比較すると，すべての手法で正解率を再現率が下回るという結果となった．よって，提案手法と人手で手順に写真をつける方法の間には，そのやり方に大きく異なる点があると考えられる．そのため，まず提案手法の根本的な問題点を指摘し，人手で付与したアノテーション結果と比較しながら考察を行う．余弦類似度の値が 0 になってしまうため，提案手法が原理上ミスしてしまうのは以下のようなケースである．

- Faster-RCNN の検出ミスによって，映像から特徴ベクトルの抽出に失敗するケース
- 物体の名称の省略によって，テキストから特徴ベクトルの抽出に失敗するケース

前者は，以下の図 9 のようなケースである．括弧内は後ろの文章から補完して終止形にしている．今回チューニングしたモデルは，小麦粉の正例として袋に詰まっている状態で学習しているため，鶏肉の上にまぶされた小麦粉を検出に失敗している．その結果，余弦類似度が 0 となるため，人手のアノテーションのような正解のフレームを付与することができない．

これを解決するためには，アノテーションの枚数を増やす必要がある．しかし，本研究では，一般的な物体検出のデータセットを用いて訓練済みのモデルに，KUSK Object Dataset を用いて一つ抜き法を用いて転移学習を行ったものを用いており，学習するためのデータセットの数を増やすことは非常に高コストである．

後者は，図 10 のようなケースである．括弧内は前後の文章から補完している．赤字のじゃが芋は前の指示句に含まれるものである．このケースは，レシピのフローグラフ [17] を用いることで解決する可能性がある．指示句には少なくとも一つは動作が含まれているため，その動作を実行する上で必要な物体をフローグラフによって探し出し，補完することができたならば，その指示句を実行する上で必要な物体を含めて特徴ベクトル化することで，性能を向上し得る可能性がある．

4.6　調理以外の一般の手順書と実施映像への本手法の適用

本研究では，調理を題材に，手順書と実施映像から写真付き手順書を生成したが，調理のみに適用可能な手法を用いていないため，調理以外の一般のドメインにおいても利用することができる．その場合に，必要となるのは以下の 2 点である．

- テキストから手順実施の際に重要な物体のを抽出するための固有表現認識器
- 動画から手順実施の際に重要な物体を抽出するための物体検出器

提案手法で計算した余弦類似度やシーンの重要度は，上の抽出器を用いて手順実施において重要な物体を抽出することで計算できる．前者の固有表現認識器は，適用したいドメインに応

じて学習させることで作成することができる．レシピコーパスの場合，3000 文ほどの学習データを用いることで，調理に重要な物体である F と T の認識精度は 95%である [13]．適用させたいドメインに応じて固有表現認識器の性能は変化するが，重要な物体は高頻度で出現するものであるため，十分に抽出することは可能であると考えられる．後者の物体検出器も，適用したいドメインの動画から，重要な物体に長方形でアノテーションさせて学習データを作成することで学習することができる．ただし，テキストにアノテーションすることに比べて，物体検出のためのアノテーションは高コストであるという問題がある．しかし，一般の物体検出で学習させた物体検出器から転移学習を行うことによって，少量のアノテーションデータを用いて十分な性能で利用することができる．本研究では，KUSK Object Dataset [12] を用いて一つ抜き法で学習させることで，物体検出器を作成している．正解率は 74%であり，十分に重要な物体を認識することができる．以上の考察より，学習データを作成して抽出器を学習させることで，各ドメインに適用することができると考えられる．

5 まとめと今後の課題

本研究では，まず，手順書と未編集の実施映像を入力として与え，各手順に作業者の内容理解・実施補助のためにふさわしい動画フレームを映像から選択するという問題を考えた．次に，それを解決する手法として，各手順と実施映像の各動画フレームで間でそれぞれに現れる重要な物体が共起する時が組み合わせとして相応しいと仮定した．その物体を特徴ベクトルに変換し，cos 類似度を計算することで，手順と実施映像の各動画フレームとの間での相応しさとし，Viterbi アルゴリズムを用いることで手順全体と実施映像全体で最も相応しいと考えられる組み合わせを得た．さらに洗練された組み合わせを得るために，我々は物体の視覚的な変化に着目した．視覚的に物体が大きく変化する時が，作業を実施する上で重要な動画フレームになりうると考えたためである．これらの複数の提案手法と人手によるアノテーションの結果と比較して評価を行った．

提案手法は人手の性能よりも下回る結果となったものの，人手のアノテーション結果と見比べることにより，手法の問題点を考察を行った．その結果，一部正解することはできているものの，物体検出側や，テキスト側の検出ミスや省略によって，特徴ベクトルを得る段階で失敗していることが提案手法の性能の低さの結果であることが明らかとなった．

今後の課題として，手順内容に含まれているが，実施映像から抽出できないため利用していない動作情報を含めるため，深層学習を用いてお互いを共有の潜在空間に埋め込んだ上で比較する手法などを用いて，より密なベクトルを得た後に本手法を再度試すことを検討している [4]．

文　献

[1] Hashimoto, A., Sasada, T., Yamakata, Y., Mori, S., and Minoh, M., “KUSK Dataset: Toward a direct understanding of recipe text and human cooking activity” *Proc. of UbiComp*, ACM(2014)

[2] Vinyals, O., Toshev, A., Bengio, S., and Erhan, D., “Show and Tell: A Neural Image Caption Generator” *Proc. of CVPR*, IEEE(2016)

[3] Guo, Z., Gao, L., Song, J., Xu, X., Shao, J. and Shen, H. T., “Attention-based LSTM with Semantic Consistency for Videos Captioning” *Proc of ACMMM*, ACM(2016)

[4] Zhang, Y. and Lu, H., “Deep Cross-Modal Projection Learning for Image-Text Matching” *Proc of ECCV*, IEEE(2018)

[5] Ushiku, A., Hashimoto, H., Hashimoto, A. and Mori, S., “Procedural Text Generation from an Execution Video” *Proc of IJCNLP*, IEEE(2017)

[6] Salvador, A., Hynes, N., Aytar, Y., Marin, J., Ofli, F., Weber, I., and Torralba, A., “Learning Cross-modal Embeddings for Cooking Recipes and Food Images” *Proc of CVPR*, IEEE(2017)

[7] Jonathan, M., Jonathan, H., Vivek, R., Nick, J., Andrew, R., Kevin, M., “What’s Cookin’? Interpreting Cooking Videos using Text, Speech and Vision” *Proc of NAACL*, ACM(2015)

[8] Naim, I., Song, Y. C., Liu, Q., Kautz, H., Luo, J., and Gildea, D., “Unsupervised alignment of natural language instructions with video segments” *Proc. of AAAI*, ACM(2014)

[9] Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L., “The Mathematics of Statistical Machine Translation: Parameter Estimation” *Computational Linguistics*(1993)

[10] Bojanowski, P., Lajugie, R., Grave, E., Bach, F., Laptev, I., Ponce, J., and Schmid, C., “Weakly-Supervised Alignment of Video With Text” *Proc. of ICCV*, IEEE(2015)

[11] Ren, S., He, K., Girshick, R. and Sun, J., “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” *Advances in Neural Information Processing Systems* (2015)

[12] Hashimoto, A., Mori, S., Iiyama, M. and Minoh, M., “KUSK Object Dataset: Recording Access to Objects in Food Preparation” *Proc. of ICME*, IEEE(2016)

[13] 笹田鉄郎, 森信介, 山肩洋子, 前田浩邦, 河原達也, “レシピ用語の定義とその自動認識のためのタグ付与コーパスの構築” 自然言語処理 (2015)

[14] Sasada, T., Mori, S., Kawahara, T. and Yamakata, Y., “Named Entity Recognizer Trainable from Partially Annotated Data” *Proc. of PACLING*, ACM(2015)

[15] 土居 洋子, 辻田 美穂, 難波 英嗣, 竹澤 寿幸, 角谷 和俊, “料理レシピと特許データベースからの料理オントロジーの構築” 電子情報通信学会 MVE/CEA 研究会 (2014)

[16] He, K., Zhang, X., Ren, S., and Sun, J., “Deep Residual Learning for Image Recognition” *Prof of CVPR*, IEEE(2016)

[17] 前田 浩邦, 山肩 洋子, 森 信介, “手順文書からの意味構造抽出” 人工知能学会論文誌 (2017)