手順構造を考慮した作業映像からの手順書生成

西村太一
京都大学大学院情報学研究科
nishimura.taichi.43x@st.kyoto-u.ac.jp

橋本敦史
オムロンサイニックエックス株式会社
atsushi.hashimoto@sinicx.com

牛久祥孝
オムロンサイニックエックス株式会社
yoshitaka.ushiku@sinicx.com

亀甲博貴
京都大学 学術情報メディアセンター
kameko@i.kyoto-u.ac.jp

森信介
京都大学 学術情報メディアセンター
forest@i.kyoto-u.ac.jp

1 はじめに

手順書は，料理や科学実験などの一連の手続きを自然言語で記述した文書である．手順書を作業映像から自動的に生成できるようになれば，作業者が作業内容を検証したり，同じ結果を再現したりする上で有用である．本研究では，作業映像から手順書を生成することを目的とする．

正しく手順書を生成するためには，モデルは単に映像の情景を記述するのではなく，人が読んで作業を実施できるような一貫性を持った文を生成することが要求される．そのためには，材料や動作の依存関係を捉えること，つまり，(1) 動作や材料の順番が正しく，(2) ある手順の動作後の状態をもとに次の手順文を生成することが重要である．

我々は，材料や動作の依存関係を木構造をはじめとするグラフ構造で表現することで [1, 2, 3]，前述した要件を満たすモデル化が可能になると考えている．こうした構造を**手順構造**と呼び，図 1 にその一例を示す．この例では，手順 1 でトマトが切られ，手順 2 においてかぼちゃが切られ炒められる．こうして得られた切ったトマト，炒められたかぼちゃが手順 1，2 のノードにそれぞれ対応し，次に手順 3 で使われることを示している．以前の我々の研究では，動画ではなく写真列を対象に，同様の手順構造の半教師ありで学習しつつ，手順書を生成するモデルを提案した [2]．本研究では，このアイデアを動画列へ拡張し，かつ教師なしで手順構造をモデル化することで，一貫性のある手順書を生成する手法を提案する．

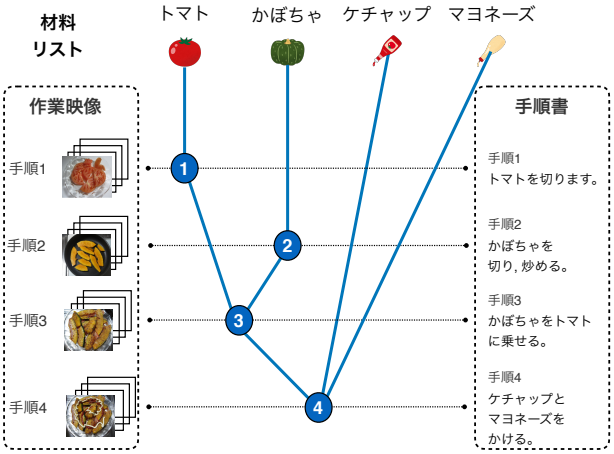

図 1 物体と動作の依存関係を表現した木構造.

実験では，映像から文を生成する従来手法と比較して文の自動評価尺度において性能が向上したこと，また，定性的評価において提案手法がある程度構造を考慮しつつ文を生成できていることを示した．

2 関連研究

画像や映像といった視覚情報から手順書を生成する研究が近年活発に行われている．手順書の中でも，料理ドメインは Web 上で大量にデータを集めやすく，動作や材料が多様であるため，特に注目を集めている．こうした研究の先駆けとして，Salvador ら [4] は完成写真からタイトル，材料リスト，レシピをまとめて生成する手法を提案した．その後，Chandu ら [5]，Nishimura ら [6] は，1 枚の完成画像ではなく，複数枚の作業途中の写真列からレシピを

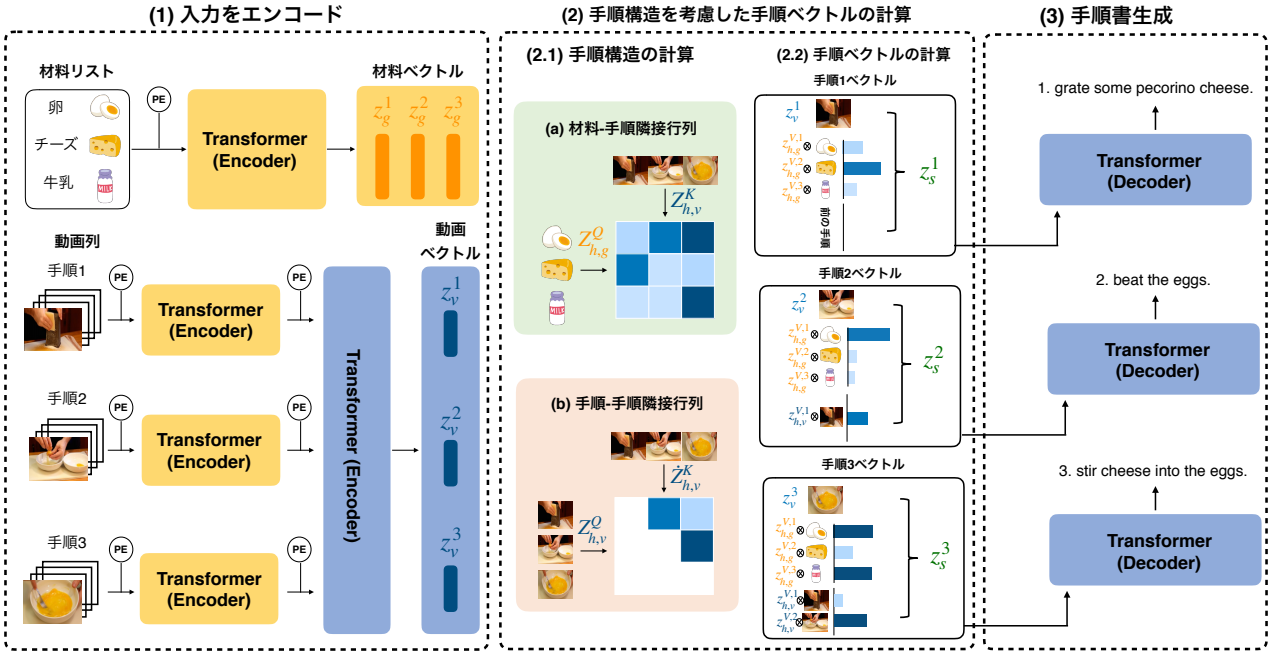

図 2　提案手法の概要図.

生成する課題と手法を提案した．作業映像から手順書を生成する研究も行われている．Shi ら [7] は作業映像と，映像中の音声の書き起こしを入力にして手順書を生成する手法を提案している．

また，手順書の理解に向けて，材料や動作の依存関係をグラフで構造的に表現する研究も行われてきた [2, 1, 3]．近年ではこれを拡張し，手順書に付与された画像情報も活用してマルチモーダルに木構造で表現する取り組みも行われている [2, 8]．本研究ではこの木構造の手順構造を教師なしで，入力の作業映像から獲得しつつ，手順書を生成する手法を提案する．

3　提案手法

提案手法は大きく分けて 3 つのプロセスからなる．(1) まず，最初に動画列と材料リストをエンコードし，動画，材料のベクトル表現を獲得する．(2) 次に，手順構造を獲得するために，得られた動画，材料ベクトルから材料-手順隣接行列，手順-手順隣接行列を計算する．計算した隣接行列を用いて，手順ベクトルを計算する．(3) 最後に，計算した手順ベクトルを入力に手順書の各手順文を生成する．

3.1　入力のエンコード

動画列のエンコード．動画列をエンコードするために，動画列の各動画の先頭にエンコードを意味する [ENC] ラベルを結合して Transformer [9] を用いてエンコードする．そして，[ENC] ラベルに対応するベクトルを動画ベクトルとして得る．次に，動画列の時系列情報を考慮するために，各動画ベクトルを別の Transformer へ入力し，動画列の時系列情報を考慮した動画ベクトル $\boldsymbol{Z_v} = (z_v^1, z_v^2, \ldots, z_v^N)$ を得る．ここで，N は動画列に含まれる動画数である．

材料リストのエンコード．各材料を単語の分散表現に変換した後，材料リスト内の関係を考慮するため，Transformer を用いて材料ベクトル $\boldsymbol{Z_g} = (z_g^1, z_g^2, \ldots, z_g^M)$ を得る．ここで，M は材料リストに含まれる材料数である．

3.2　手順構造を考慮した手順ベクトルの計算

(2.1) 手順構造の計算．得られた動画ベクトル，材料ベクトルを用いて手順構造を計算する．節 1 で述べた通り，手順構造は木構造として表現することができ，これは材料がどの手順へ接続するか，ある手順がどの手順へ接続するのかを材料-手順，手順-手順の隣接行列を計算することで得られる．本研究では，この 2 つの行列計算をマルチヘッドアテンションとして定式化する．まず，材料ベクトル，動画ベクトルから各ヘッドごとの材料ベクトル $\boldsymbol{Z_{h,g}} = (z_{h,g}^1, z_{h,g}^2, \ldots, z_{h,g}^M) \in \mathbb{R}^{|\boldsymbol{h}| \times M \times d_h}$，動画ベクトル $\boldsymbol{Z_{h,v}} = (z_{h,v}^1, z_{h,v}^2, \ldots, z_{h,v}^N) \in \mathbb{R}^{|\boldsymbol{h}| \times N \times d_h}$ を得て，材料-手順 $P_{i,j}^{ingr} \in \mathbb{R}^{|\boldsymbol{h}| \times M \times N}$，手順-手順行列

$P^{inst}_{i,j} \in \mathbb{R}^{|\boldsymbol{h}|\times N\times N}$ を以下のように計算する．

$$P^{ingr}_{i,j} = \mathrm{Softmax}(\frac{\boldsymbol{Z}^Q_{h,g}(\boldsymbol{Z}^K_{h,v})^T}{\sqrt{d_h}}) \quad (1)$$

$$P^{inst}_{i,j} = \mathrm{Softmax}(\frac{\boldsymbol{Z}^Q_{h,v}(\dot{\boldsymbol{Z}}^K_{h,v})^T}{\sqrt{d_h}}) \quad (2)$$

ここで，$\boldsymbol{Z}^Q_{h,g}, \boldsymbol{Z}^K_{h,v}, \boldsymbol{Z}^Q_{h,v}, \dot{\boldsymbol{Z}}^K_{h,v}$ はそれぞれ別の 1 層の線型結合層で変換されたベクトルである．計算した材料-手順隣接行列 $P^{ingr}_{i,j}$ では，i 番目の材料が j 番目の手順に接続する確率を，手順-手順隣接行列 $P^{inst}_{i,j}$ は，i 番目の手順が j 番目の手順に接続する確率を表す．ただし，$P^{inst}_{i,j}$ については，ある手順は前の手順に遡って繋がらないという仮定のもと，$i \geq j$ 番目の要素には非常に小さい値 $\epsilon(= 1.0 \times 10^{-5})$ が挿入されている．

(2.2) 手順ベクトルの計算．次に，得られた隣接行列を用いて，各手順に対応するベクトルを計算する．$P^{ingr}_{i,j}$ の j 列目は j 番目の手順で使われる材料の重みが，$P^{inst}_{i,j}$ の j 列目は j 番目の手順と依存関係にある手順の重みが計算されているとみなすことができる．よって，j 番目の手順における各材料，手順の隣接行列の重みと，材料，動画ベクトルを掛け合わせることによって，j 番目の手順での材料と動画の重要度を加味したベクトルを得ることができる．具体的には，材料，動画ベクトルを 1 層の線型結合層で変換したベクトル $\boldsymbol{Z}^V_{h,g}, \boldsymbol{Z}^V_{h,v}$ と，$P^{ingr}_{i,j}, P^{inst}_{i,j}$ の j 列目を用いて，j 手順目と関係のある材料ベクトル，動画ベクトルを得る．

$$\boldsymbol{Z}^j_{h,g} = [P^{ingr}_{1,j}\boldsymbol{z}^{V,1}_{h,g}, P^{ingr}_{2,j}\boldsymbol{z}^{V,2}_{h,g}, \ldots, P^{ingr}_{M,j}\boldsymbol{z}^{V,M}_{h,g}] \quad (3)$$

$$\boldsymbol{Z}^j_{h,v} = [P^{inst}_{1,j}\boldsymbol{z}^{V,1}_{h,v}, P^{inst}_{2,j}\boldsymbol{z}^{V,2}_{h,v}, \ldots, P^{inst}_{N,j}\boldsymbol{z}^{V,N}_{h,v}] \quad (4)$$

ここで，$[\cdot]$ はベクトルの結合を表す．こうして得られるベクトルの各ヘッドのベクトルを結合し，j 番目の動画ベクトル $\boldsymbol{z}^j_v$ を結合することによって，手順書を生成するのに必要な j 番目の手順ベクトル $\boldsymbol{z}^j_s$ を計算する．

3.3 手順書の生成

最後に，得られた手順ベクトルから対応する手順文を生成し，それらを結合することで手順書を得る．手順書を $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N)$ とし，各 $\boldsymbol{y}_j$ は手順文を表すとする．この時，損失関数 L は全ての動画列と手順書のペアからなるデータセット D に対して，以下の負の対数尤度が最小となるように計算を行う．

$$L(\boldsymbol{\theta}) = -\sum_D \sum_j \log p(\boldsymbol{y}_j | z^j_s; \boldsymbol{\theta}) \quad (5)$$

ここで，$\boldsymbol{\theta}$ はモデル全体のパラメータを表す．また，デコーダには Transformer を用いる．

表 1 YouCook2 材料データセットの統計情報．

	訓練	検証
レシピ数	1,235	429
レシピあたりの手順数	7.73	7.67
レシピあたりの材料数	10.37	10.27

4 実験

4.1 実験設定

データセット．作業映像と手順書のデータセットとして YouCook2 [10] を利用した．YouCook2 は作業映像に対し，手順書の各手順に対応する区間がアノテーションされたデータセットである．本研究では，この区間を抽出することで，各手順に対応する動画列を作成している．YouCook2 には材料のアノテーションが行われていないため，材料のアノテーションを 3 人のアノテータを介して行なった．アノテーションの結果得られたデータを YouCook2 と併せて，本研究では YouCook2 材料データセットと呼ぶこととする．データセットの統計情報を表 1 に示す．なお，YouCook2 ではテストセットは公開されていないため，材料のアノテーションは行っていない．以後の実験結果は全て検証セットで評価されたものである．

比較モデル．本研究と比較するモデルとして，同じく Transformer ベースで動画列から文章を生成するモデルである Transformer-XL [11]，MART [12] を用意した．元々のモデルは材料を入力としていないため，提案手法との公平な比較のために，提案手法と同様に材料リストをエンコードした後，入力の動画ベクトルと結合してモデルを学習させた．また，提案手法の中でどの要素が効果があったのかを調べるために，提案手法から要素を除いたモデルも学習して比較した．

4.2 定量的評価

文の自動評価尺度である BLEU, METEOR, CIDEr-D を用いて生成した手順書を評価した．表 2 に評価した結果を示す．全ての評価尺度において，既

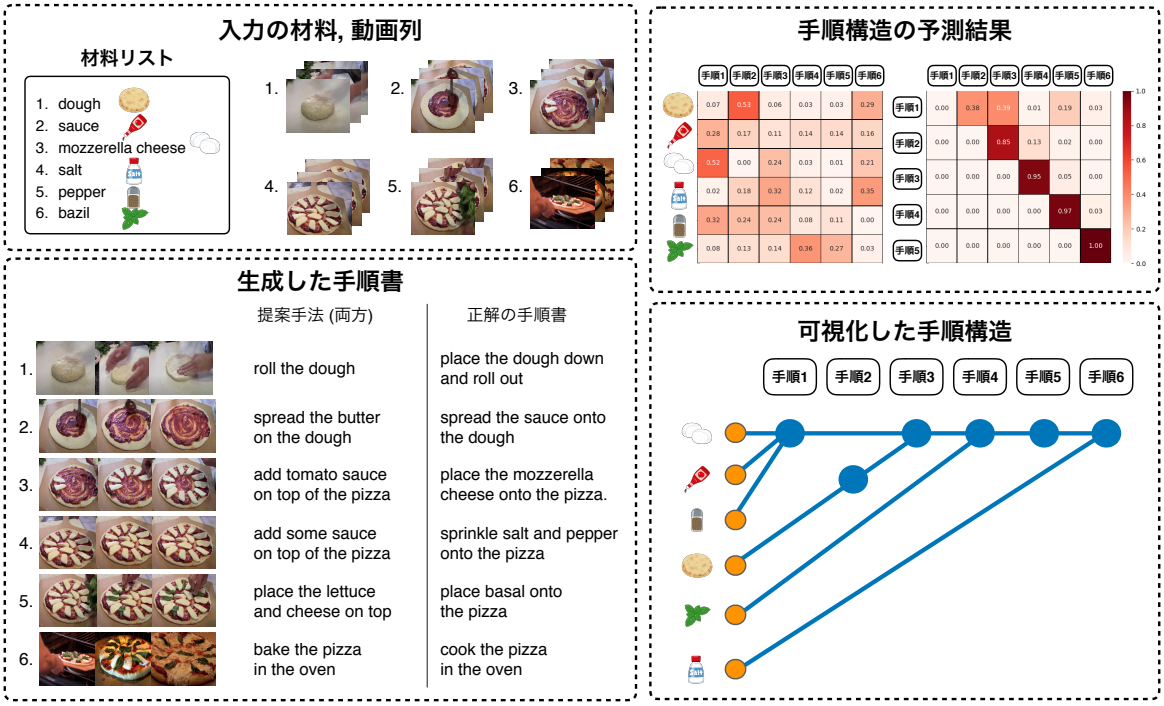

図 3　手順書の生成結果．手順構造の予測結果はヘッドの平均値で計算されたものである．また，可視化した手順構造は隣接行列の各行で最大の値をサンプリングすることで得られたものである．

表 2　文の自動評価尺度による評価．最も性能が高いものは太字に，次に性能が高いものには下線を引いている．

	材料	BLEU1	BLEU4	METEOR	CIDEr-D
Transformer-XL		31.3	5.8	13.1	20.8
MART		36.9	7.0	14.8	31.7
Tansformer-XL + 材料	✓	34.1	6.6	14.9	28.4
MART + 材料	✓	37.9	8.8	15.9	36.9
提案手法 (動画のみ)		39.0	7.2	15.4	31.1
提案手法 (材料-手順のみ)	✓	**40.4**	**8.9**	**16.6**	42.4
提案手法 (手順-手順のみ)		38.9	7.7	15.5	33.1
提案手法 (両方)	✓	40.2	8.3	16.3	**44.3**

存モデルと比較して提案手法が最も高い性能を示した．中でも，**材料-手順のみ**が BLEU1，BLEU4，METEOR において最も高い性能を示しており，材料-手順の隣接行列を計算することが有効であることが分かる．一方，**手順-手順のみ**のモデルは，**動画のみ**のモデルに比べ性能は向上している．しかし，全てを組み込んだ**両方**のモデルでは CIDEr-D を除いて**材料-手順のみ**のモデルより悪化する結果となっており，材料-手順と手順-手順の依存関係を同時に考慮する方法については未だ検討が必要である．

4.3　定性的評価

図 3 に提案手法が生成した手順書，正解の手順書，及び手順構造の予測結果を示す．提案手法はある程度正しく生成できた手順もある一方で (手順 1，手順 6)，誤った材料や動作の記述があるものも見られる (手順 2，手順 3)．手順構造の予測結果を可視化すると，手順-手順の依存関係はある程度正しく捉えられているが，材料-手順については正しく予測できなかった．その結果，異なる手順で異なる材料を使用して手順書を生成していると考えられる．このことから，手順構造を正しく計算しつつ，手順書の生成に効果的に反映させるようにモデルを改善することを検討している．

5　おわりに

本研究では，作業映像から手順書を生成する課題に取り組んだ．正しく手順書を生成するためには，材料と動作の依存関係をモデルが理解する必要がある．本研究では，こうした材料と動作の依存関係を木構造として明示的に表現しつつ，文生成モデルに組み込むことで正しく手順書を生成する手法を提案した．実験では，文の自動評価尺度による評価と定性的評価を行い，既存の文生成手法と比較して正しく手順書を生成できていることを確認した．

今後は検証セットに手順構造のアノテーションを行い，人のアノテーションとモデルの予測結果がどの程度一致するのか調査する．また，手順書の一貫性をより詳しく評価するために，人手評価および別の評価尺度を検討する．

参考文献

[1]Shinsuke Mori, Hirokuni Maeta, Yoko Yamakata, and Tetsuro Sasada. Flow graph corpus from recipe texts. In *LREC*, 2014.

[2]Taichi Nishimura, Atsushi Hashimoto, Yoshitaka Ushiku, Hirotaka Kameko, Yoko Yamakata, and Shinsuke Mori. Structure-aware procedural text generation from an image sequence. *IEEE Access*, 2020.

[3]Chloé Kiddon, Ganesa Thandavam Ponnuraj, Luke Zettlemoyer, and Yejin Choi. Mise en place: Unsupervised interpretation of instructional recipes. In *EMNLP*, 2015.

[4]Amaia Salvador, Michal Drozdzal, Xavier Giro i Nieto, and Adriana Romero. Inverse cooking: Recipe generation from food images. In *CVPR*, 2019.

[5]Khyathi Chandu, Eric Nyberg, and Alan W Black. Storyboarding of recipes: Grounded contextual generation. In *ACL*, 2019.

[6]Taichi Nishimura, Atsushi Hashimoto, and Shinsuke Mori. Procedural text generation from a photo sequence. In *INLG*, 2019.

[7]Botian Shi, Lei Ji, Yaobo Liang, Nan Duan, Peng Chen, Zhendong Niu, and Ming Zhou. Dense procedure captioning in narrated instructional videos. In *ACL*, 2019.

[8]Pan Liang-Ming, Chen Jingjing, Wu Jianlong, Liu Shaoteng, Ngo Chong-Wah, Kan Min-Yen, Jiang Yugang, and Chua Tat-Seng. Multi-modal cooking workflow construction for food recipes. In *ACMMM*, 2020.

[9]Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[10]Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018.

[11]Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In *ACL*, 2019.

[12]Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara Berg, and Mohit Bansal. Mart: Memory-augmented recurrent transformer for coherent video paragraph captioning. In *ACL*, 2020.