手順構造を考慮した手順書からの作業画像検索

迫田航次郎
京都大学大学院情報学研究科
sakoda.kojiro.48z@st.kyoto-u.ac.jp

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

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

ある手順書に基づいて作業を行う場合，手順書に記されている文章だけでなく，その手順における物体の状態を写した画像を共に参照することで作業者は手順を再現しやすくなる．本研究では手順書を入力として，各手順に対応する画像を検索することを目的とする (図 1).

手順書から作業画像を検索するためには，モデルは単に対応する手順の内容だけでなく，過去の手順で使用した材料や動作との因果関係も考慮する必要がある．我々は，この因果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ができる [1, 2, 3]. 本研究ではこうした構造を**手順構造**と呼び，図 1 にその一例を示す．この例では，手順書の (1) で材料リストにあるアスパラガスとじゃがいもを茹で，(2) でそれらを皿に並べている．(4) では (2) の結果と (3) で刻んだチーズを合わせてオーブンで加熱している．手順構造はこれらの手順間の因果関係を表現しており，各材料が手順構造の葉に，各手順が節点に対応している．本研究では，手順書と，材料リスト，そして手順構造を入力として与え，手順構造を効率よく学習しつつ作業画像を検索する手法を提案する．

実験では，手順構造を考慮しない従来の画像検索モデルと比較して画像検索性能が向上したことが分かった．また，実際に検索した結果をもとに，手順構造を考慮することの有用性を示した．

2 関連研究

手順書とその実施映像を対象とした研究は近年活発に行われている．中でも，料理ドメインは Web 上でデータを集めやすく，材料や動作の種類が豊富

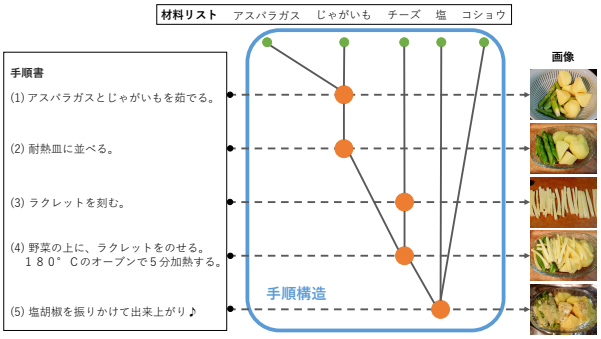

図 1 本研究の課題概要．手順書，材料リスト，手順構造を入力として，手順書の各手順に対応する作業画像を検索する．

であるため注目を集めている．本研究で利用する Cookpad Image Dataset [4] は約 170 万のレシピ，画像列，材料リストを含む世界最大規模のマルチモーダルデータセットである．

こうした大規模なデータセットを活用して，レシピとその実施映像との間で共通の表現を学習する研究も行われている．Salvador ら [5] は，テキスト側の入力として手順と材料を，画像の入力としてレシピの完成画像をそれぞれエンコードし，共有潜在空間を学習することで，レシピから完成画像を検索する手法を提案した．

しかし，材料と動作の因果関係を考慮しつつ検索を行うような研究は未だ十分に取り組まれてはいない．レシピの内容をコンピュータが理解するために，こうした関係を手順構造で表現する取り組みも行われている [1, 2]．さらに，近年では，手順構造を画像付きで理解する試みも行われており，そのためのマルチモーダルデータセットとして，vSIMMR [3] というデータセットも提案されている．本研究では，vSIMMR データセットを用いて，材料と動作の因果関係を考慮しつつ，手順書の各手順に対応する

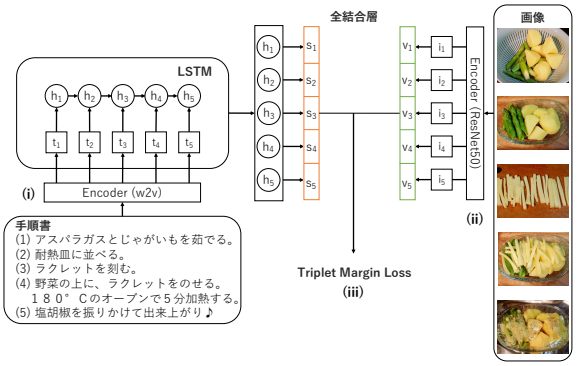

図 2　従来の画像検索モデル (ベースラインモデル).

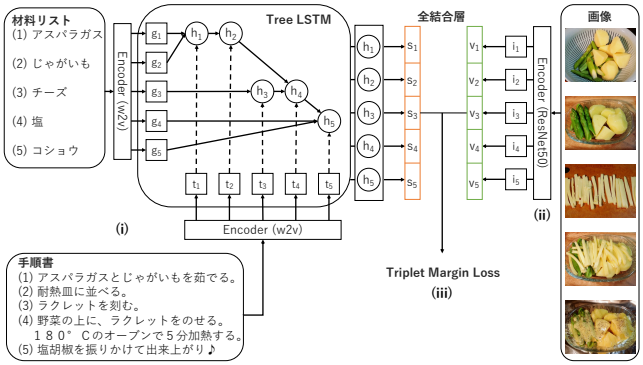

図 3　提案手法モデル.

画像を検索する手法を提案する.

3　作業画像の検索モデル

図 2 に従来の画像検索モデル (ベースラインモデル) の，図 3 に提案手法モデルの概要を示す．また，以下の節でそれぞれのモデルについて説明する.

3.1　ベースラインモデル

ベースラインモデルは 3 つの処理に分けることができる．(i) で手順列を入力に，LSTM を用いて手順書の時系列情報を考慮した特徴ベクトルを得る．(ii) で画像列を入力に，全結合層を用いることで画像の特徴ベクトルを得る．最後に，(iii) で (i) と (ii) で得られた手順と画像の特徴ベクトルを 3 層の全結合層により共有潜在空間上に埋め込む.

(i) **手順列の入力：**手順列の入力部は 1 層の LSTM と 2 層の全結合層からなる．まず，各手順を構成する単語を word2vec を用いて分散表現に変換し，その平均を手順ベクトル列 $\boldsymbol{T} = (\boldsymbol{t_1}, \boldsymbol{t_2}, ..., \boldsymbol{t_k}, ..., \boldsymbol{t_K})$ として得る．次に，手順ベクトル列を LSTM に入力し，手順書の時系列情報を考慮した手順ベクトル列 $\boldsymbol{H} = (\boldsymbol{h_1}, \boldsymbol{h_2}, ..., \boldsymbol{h_k}, ..., \boldsymbol{h_K})$ を以下のように計算して得る.

$$\boldsymbol{i_k} = \sigma(\boldsymbol{W}^{(i)}\boldsymbol{t_k} + \boldsymbol{U}^{(i)}\boldsymbol{h_{k-1}} + \boldsymbol{b}^{(i)}) \quad (1)$$

$$\boldsymbol{f_k} = \sigma(\boldsymbol{W}^{(f)}\boldsymbol{t_k} + \boldsymbol{U}^{(f)}\boldsymbol{h_{k-1}} + \boldsymbol{b}^{(f)}) \quad (2)$$

$$\boldsymbol{o_k} = \sigma(\boldsymbol{W}^{(o)}\boldsymbol{t_k} + \boldsymbol{U}^{(o)}\boldsymbol{h_{k-1}} + \boldsymbol{b}^{(o)}) \quad (3)$$

$$\boldsymbol{u_k} = \tanh(\boldsymbol{W}^{(u)}\boldsymbol{t_k} + \boldsymbol{U}^{(u)}\boldsymbol{h_{k-1}} + \boldsymbol{b}^{(u)}) \quad (4)$$

$$\boldsymbol{c_k} = \boldsymbol{i_k} \odot \boldsymbol{u_k} + \boldsymbol{f_k} \odot \boldsymbol{c_{k-1}} \quad (5)$$

$$\boldsymbol{h_k} = \boldsymbol{o_k} \odot \tanh(\boldsymbol{c_k}) \quad (6)$$

次に，LSTM の出力を 2 層の全結合層に入力し，共有潜在空間上の手順ベクトル列 $\boldsymbol{S} = (\boldsymbol{s_1}, \boldsymbol{s_2}, ..., \boldsymbol{s_k}, ..., \boldsymbol{s_K})$ を得る.

(ii) **画像列の入力：**画像列の入力部は ResNet50 [6] と 3 層の全結合層からなる．まず，画像列の各画像を ImageNet [7] で学習済の ResNet50 を用いて分散表現 $\boldsymbol{I} = (\boldsymbol{i_1}, \boldsymbol{i_2}, ..., \boldsymbol{i_k}, ..., \boldsymbol{i_K})$ に変換し，全結合層に入力することで，共有潜在空間上の画像列ベクトル $\boldsymbol{V} = (\boldsymbol{v_1}, \boldsymbol{v_2}, ..., \boldsymbol{v_k}, ..., \boldsymbol{v_K})$ を得る.

(iii) **共有潜在空間の学習：**Triplet margin loss [8] を用いて，手順と画像の共有潜在空間を学習する．Triplet margin loss は基準となるベクトル $\boldsymbol{a}$，正例ベクトル $\boldsymbol{p}$，負例ベクトル $\boldsymbol{n}$ を用いて以下のように書ける.

$$L = \max(0, d(\boldsymbol{a}, \boldsymbol{p}) - d(\boldsymbol{a}, \boldsymbol{n}) + \epsilon) \quad (7)$$

なお，$d(\cdot)$ は 2 つのベクトルのユークリッド距離を表し，ϵ はハイパーパラメータである.

本研究では，手順を基準として得られる 3 つ組 $(\boldsymbol{s}^{\mathrm{a}}, \boldsymbol{v}^{\mathrm{p}}, \boldsymbol{v}^{\mathrm{n}})$，画像を基準として得られる 3 つ組 $(\boldsymbol{v}^{\mathrm{a}}, \boldsymbol{s}^{\mathrm{p}}, \boldsymbol{s}^{\mathrm{n}})$ を用いて，それぞれの基準での損失 L_{s2v}, L_{v2s} を計算する．そして，2 つの損失の和 L_{triplet} を最小化するようにモデルを学習する.

3.2　提案手法

提案手法のモデルもベースラインモデルと同様の 3 つの処理からなる．ベースラインモデルからの拡張点は，手順構造を考慮するために，(i) において LSTM を Tree-LSTM [9] に変更したことである．(ii) と (iii) に関してはベースラインモデルと同様であるため，以下で (i) の処理に焦点を当てて説明をする.

(i) の拡張箇所は 1 層の Tree-LSTM と 2 層の全結合層からなる．まず，手順構造の節点となる手順書の各手順ベクトルをベースラインモデルと同様に計算する．また，手順構造の葉となる材料ベクトル $\boldsymbol{G} = (\boldsymbol{g_1}, \boldsymbol{g_2}, ..., \boldsymbol{g_n}, ..., \boldsymbol{g_N})$ も同様の処理で計算する．こうして得られる手順，材料ベクトルおよび手順構

表 1　Cookpad Image Dataset の統計情報.

	訓練	検証	評価
レシピ数	163,525	18,051	20,193
平均手順数	6.24	6.15	6.26
レシピあたりの単語数	148.52	147.02	148.50
レシピあたりの材料数	7.85	7.79	7.86

造を Tree-LSTM へ入力する．k 番目の手順ベクトル $\boldsymbol{t_k}$ と k 番目の節に接続している子の集合 $C(k)$ を用いて，以下のように k 番目の隠れ層 $\boldsymbol{h_k}$ とメモリセル $\boldsymbol{c_k}$ を更新する．

$$\hat{\boldsymbol{h}}_k = \sum_{j \in C(k)} \boldsymbol{h}_j^k \tag{8}$$

$$\boldsymbol{i}_k = \sigma(\boldsymbol{W}^{(i)}\boldsymbol{t}_k + \boldsymbol{U}^{(i)}\hat{\boldsymbol{h}}_k + \boldsymbol{b}^{(i)}) \tag{9}$$

$$\boldsymbol{f}_{kj} = \sigma(\boldsymbol{W}^{(f)}\boldsymbol{t}_k + \boldsymbol{U}^{(f)}\hat{\boldsymbol{h}}_k + \boldsymbol{b}^{(f)}) \tag{10}$$

$$\boldsymbol{o}_k = \sigma(\boldsymbol{W}^{(o)}\boldsymbol{t}_k + \boldsymbol{U}^{(o)}\hat{\boldsymbol{h}}_k + \boldsymbol{b}^{(o)}) \tag{11}$$

$$\boldsymbol{u}_k = \tanh(\boldsymbol{W}^{(u)}\boldsymbol{t}_k + \boldsymbol{U}^{(u)}\hat{\boldsymbol{h}}_k + \boldsymbol{b}^{(u)}) \tag{12}$$

$$\boldsymbol{c}_k = \boldsymbol{i}_k \odot \boldsymbol{u}_k + \sum_{j \in C(k)} \boldsymbol{f}_{kj} \odot \boldsymbol{c}_j \tag{13}$$

$$\boldsymbol{h}_k = \boldsymbol{o}_k \odot \tanh(\boldsymbol{c}_k) \tag{14}$$

4　実験

まず，ベースラインモデル，提案手法を用いて画像検索の性能を定量的に評価した．次に，手順構造を考慮することの有用性を示すために，実際の検索結果を用いて定性的評価を行なった．

4.1　データセット

本研究では，Cookpad Image Dataset，vSIMMR の 2 つのデータセットを用いて実験を行なった．Cookpad Image Dataset は，レシピ，材料のリスト，レシピの各手順に対応する画像からなるデータセットである．Cookpad Image Dataset のレシピの中には，全ての手順に写真が付与されていないものも存在する．本研究では，そうしたレシピを除き，全ての手順に写真が付与しているレシピのみを用いた．表 1 に Cookpad Image Dataset の統計情報を示す．

vSIMMR は，Cookpad Image Dataset の一部に手順構造を人手でアノテーションしたデータセットである．このデータセットは，Cookpad Image Dataset のレシピから，全ての手順に写真がついており，かつ材料数と手順数が 3 以上のものを取り出し，その一部に木構造のアノテーションが行われたものである．表 2 に vSIMMR の統計情報を示す．

表 2　vSIMMR の統計情報.

	訓練	検証	評価
レシピ数	1,603	250	250
平均手順数	6.78	6.74	6.85
レシピあたりの単語数	118.23	113.91	114.68
レシピあたりの材料数	6.58	6.37	6.64

4.2　実験設定

本研究では，word2vec モデルは Cookpad Image Dataset と vSIMMR の訓練データの全レシピを用いて学習した．なお，word2vec の出力の次元数は 300 である．画像ベクトルは学習済みの ResNet50 の最終層を取り除いたモデルから得られ，出力の次元数は 2,048 である．共有潜在空間の次元数は 512 に設定した．また，バッチサイズは 128 とし，最適化手法には Adam [10] を用いた．

また，比較手法としてベースラインモデルの材料リストを用いない場合 (材料なし) に加え，提案手法との公平な比較のために，材料リストも用いた場合 (材料あり) も考えた．(材料なし) では，手順側の入力として手順列のみを用いて学習しており，(材料あり) では，手順側の入力として手順書の各手順ベクトルに材料ベクトルの平均を連結し学習している．

4.3　定量的評価

4.3.1　評価尺度

画像検索の性能を評価するために，1,000 の手順に対する Recall@K (R@K) と Median rank (medR) を計算した．R@K は手順を入力し，得られた画像集合を cos 類似度で降順にソートした時に手順に対応する画像が上位 K 番以内に現れる割合，medR は各手順に対応する画像の順位の中央値を示す．

4.3.2　結果

結果を表 3 に示す．ベースラインモデルの (材料あり) は用いられる材料を考慮することで検索精度の上昇を期待したが，(2) と (3)，(4) と (5) を比較した結果，手順ベクトルに材料ベクトルを単に連結するだけでは検索精度が必ずしも良くなるわけではないことが分かった．ベースラインモデルの (5) と提案手法の (6) を比較すると (6) の結果の方が medR の

表 3　画像検索の評価結果.

	medR	R@1	R@5	R@10
ベースラインモデル				
(1) ランダム	500	0.001	0.005	0.010
(2)Cookpad Image Dataset(材料なし)	10	0.111	0.342	0.492
(3)Cookpad Image Dataset(材料あり)	8	0.129	0.383	0.536
(4)vSIMMR(材料なし)	42.5	0.034	0.113	0.210
(5)vSIMMR(材料あり)	56	0.030	0.108	0.177
提案手法				
(6)vSIMMR	29.5	0.047	0.167	0.271

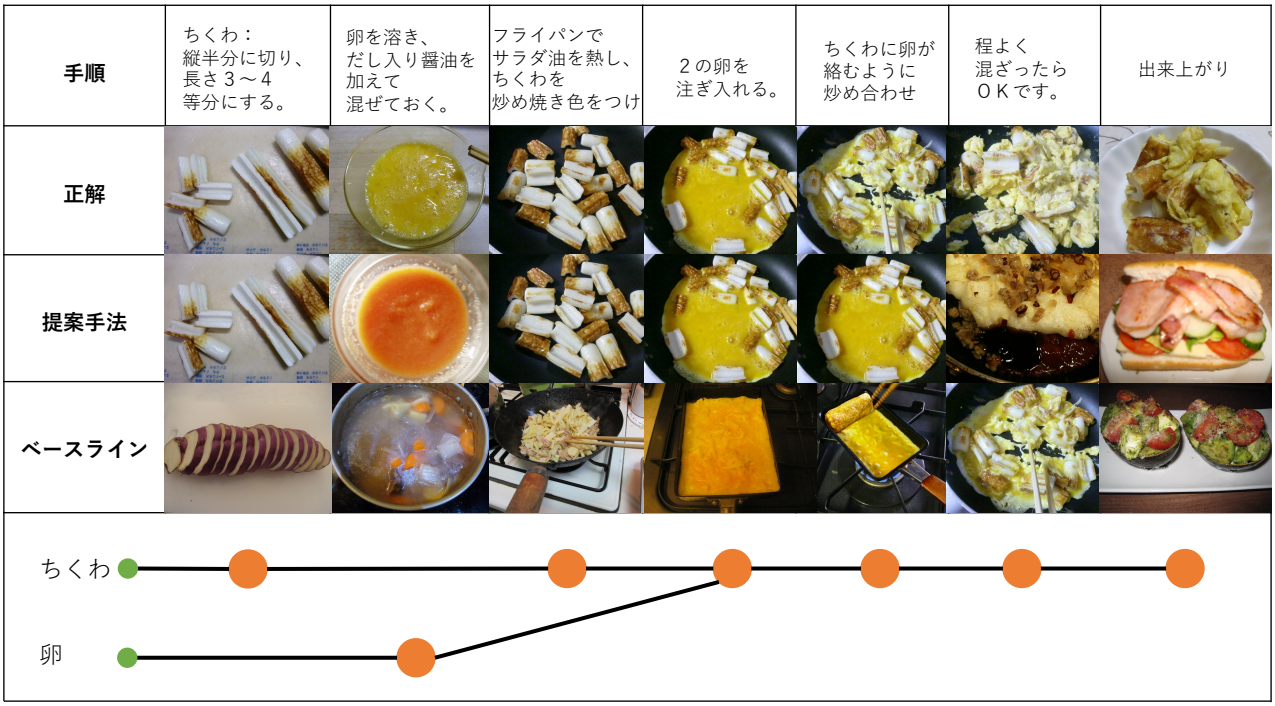

図 4　画像検索結果の一例.

値が低く，R@K の値が高かったことから，手順構造を考慮することで検索精度が良くなることが示され，提案手法の有用性を確認できた．一方で、大量のデータセットで学習されたベースラインモデル (2)，(3) と提案手法の (6) の結果を比較すると学習に用いたデータセットが多ければ手順構造を考慮しなくとも検索精度が良くなるなることが分かった．

4.4　定性的評価

レシピを入力して作業画像を検索した結果を図 4 に示す．このレシピでは 2 番目の手順で示されている溶き卵と 3 番目の手順で炒められたちくわが 4 番目の手順で混ぜ合わされるような構造となっている．提案手法ではこの混ぜ合わされた結果の画像を正確に検索することができており，手順構造を考慮することが有効であると確認できた．

5　おわりに

本研究では，手順書から作業画像を検索する課題に取り組んだ．手順に対応する画像を正しく検索するためには過去の手順で使用した材料や動作との因果関係をモデルが捉える必要がある．本研究ではこうした関係を表現している手順構造を考慮し，画像検索を行う手法を提案した．実験では，定性的，定量的に評価を行い，ベースラインモデルに比べて提案手法が有用であることを確認した．

今後の課題として，手順構造がアノテーションされていない場合でも，手順構造を予測し，それを考慮しながら画像検索を行える手法へと拡張することを検討する．

参考文献

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

[2]Jermsak Jermsurawong and Nizar Habash. Predicting the structure of cooking recipes. In *EMNLP*, 2015.

[3]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.

[4]Jun Harashima, Yuichiro Someya, and Yohei Kikuta. Cookpad image dataset: An image collection as infrastructure for food research. In *SIGIR*, 2017.

[5]Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. Learning cross-modal embeddings for cooking recipes and food images. In *CVPR*, 2017.

[6]Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[7]Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

[8]Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *CVPR*, 2014.

[9]Kai Sheng Tai, Richard Socher, and Christopher D. Manning. Improved semantic representations from tree-structured long short-term memory networks. In *ACL*, 2015.

[10]Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.