視覚と言語の共有潜在表現獲得による モノづくり手順理解

橋本 敦史

組立作業や創薬実験、生化学実験を始め、複数のものを組み合わせ、様々な工程を経て価値の高いプロダクトを作り上げる作業は、わかりやすい価値創出の方法であろう。このような作業を機械に理解させることができれば、該当のシーン検索から手順書の生成、果てはロボットによる作業の代替まで、機械との多様な協働方法が実現できる。このような理解の 1 つの形として、作業を観測した視覚情報データを手順書中の教示テキストと対応付ける視覚言語統合技術がある。特に視覚情報データとして画像列を入力として教示テキストを生成する課題において、筆者らが提案した一連の手法は State-of-the-Art となっている。本稿ではこれらの成果を中心に視覚情報と言語情報の統合による作業理解のための機械学習処理を解説する。

Cross-modal Representation Learning for Understanding Manufacturing Procedures

HASHIMOTO Atsushi

Creating high-value products through combining materials and processing them, including assembly work, drug discovery experiments, and biochemical experiments, would be one of the primary methods of value creation. If computers can understand such activities, we can collaborate with machines in various directions, such as relevant video scene retrieval, procedural text generation, and eventually substituting work by robots. As one form of such understanding, there are the vision and language arts to extract shared latent representation between visual observation and textural description of manufacturing. In particular, a series of our methods for procedural text generation from image sequence are state-of-the-art. With the central focus on these methods, this paper explains the recently-developed vision and language technologies for understanding manufacturing procedures.

1. はじめに

本稿では機械に手順を理解させるための視覚言語統合（Vision&Language）技術について解説する。まず「手順を理解する」とはどういう状態だろうか？ 1980 年代に提起された「中国人の部屋問題[*1]」は記号論理の世界に閉じた情報処理が「理解」として十分であるかを哲学的な側面から問いかけた。例えば、もし、中国語の部屋の中の人に言葉で回答できる質問ではなく、行動することを要求したらどうだろうか？辞書だけでは行動を起こすことはできないだろう。行動をするということは、言葉に対応する行為を観測し、対応関係を得る必要がある。視覚言語統合はこのような対応関係を得るための技術である。我々は、この技術を用いることで「手順の理解」、すなわち、実世界で生じる事象と自然言語で記述される記号世界を結びつけることを目指している。自然言語と視覚情報の対応付けの試みとしては、言語指示をアニメーションでシミュレートするというものが 1988 年に既に行われていたようである[9)]。残念ながらこの時代のシステムは基本的にルールベースであり、言語や視覚情報に多くの制約を課す箱庭的なシステムとなっている。このような制約なしに機械が視覚情報と言語情報を対応付けられるようになると何ができるようになるだろうか？　まず、作業を観測したデータの中から、言語で表現された手順書の各工程に対応する箇所を自由に検索することができるようになる。多様な作業者に対して、ある工程がどのように行われるかを横断的に解析することができれば、誤った方法を取っている作業者に働きかけて効率的に教育をすることができるようになる。また、作業を観測したデータから、その自然言語表現や明確なフォーマットに従ったスクリプトを生成できるようになる。これにより、個人差なく作業手順を一定の品質で書き

Contact : *HASHIMOTO Atsushi* atsushi.hashimoto@sinicx.com

出すことができるようになるだろう。例えば、現在研究ノートに手作業で書き留められるような生命科学分野の実験手順を正規化された手順書として書き出せれば、それは最新研究の再現性の問題に寄与する。同様に、個々の製品の組立状況を書き出すことは新たなトレーサビリティにもなり、品質保証の向上にも資するだろう。究極的には、言葉だけで誰でもロボットに作業を依頼できるようになるかもしれない。

このような未来を実現するには、作業と、その作業内容に対応する言語表現が同じ点に射影されるような共通潜在特徴空間を得ることが重要になる（図 1）。このような空間を得ることができれば、空間中の近傍点を探すことで、テキストをクエリにして映像を（または映像をクエリにしてテキストを）検索できる。また、映像を射影した点から、その点に対応するであろうテキストを復号（デコード）することで手順書を生成できる（またはテキストから映像を生成できる）。本稿では、このような共有潜在空間の獲得と、この空間を用いた応用研究について解説する。

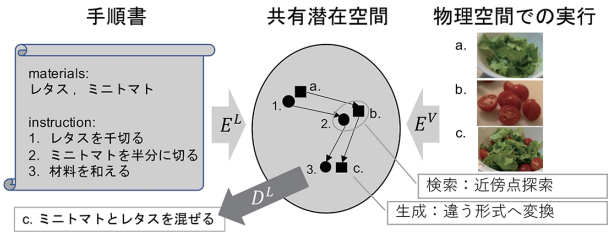


図 1　手順書と観測データの共有潜在空間への埋め込み

2. 関連研究

2.1 視覚言語統合処理

静止画や動画と言語の対応を取る研究は深層学習の登場により急速に発展した分野であり、一般には Vision & Language と呼ばれる。まだ定着した邦訳は存在しないが、本稿では筆者の過去の著作[25]に合わせ、これを視覚言語統合* と呼ぶ*2。視覚言語統合は深層学習に伴い発展した分野ではあるが、その源流は深層学習以前から連綿と研究されてきた Web 上での画像検索課題に遡ることができる。

画像検索課題はテキスト x^L を D 次元潜在特徴 z^L へ写像する関数 $E^L: x^L \rightarrow z^L$（言語符号化器）と、画像（または映像）x^V を D 次元潜在特徴 z^V へ写像する関数 $E^V: x^V \rightarrow z^V$（視覚符号化器）を得ることで実現される（図 1）。学習は同じ内容を示す $\{x^L, x^V\}$ の対からなるデータセットを用いて、対照性損失 *（Contrastive Loss）や三重項損失*（Triplet Loss）といった距離学習用の損失関数を最小化することで実現される。

検索課題は言語や画像といったモダリティを超えた共通の情報（key feature）を見つけることができれば解くことができた。これに対して、生成課題は共通部分だけでなく、生成対象のモダリティに含まれる情報すべてを予測しなければならないという点で検索よりも挑戦的な課題であり、特に画像からテキストを生成する「キャプショニング」は深層学習による視覚言語統合の代名詞ともなっている。キャプショニングの学習データは検索課題同様、$\{x^L, x^V\}$ の対となるが、E^V に加えて、潜在特徴からテキストを復号する関数 $D^L: z^V \rightarrow x^L$ を学習する（図 1）。単純にはテキスト中の単語をカテゴリとして、カテゴリ系列の一致度最大化により学習を行う。

上述の課題において、潜在特徴 $z^{L/V}$ は x^L と x^V のいずれとも対応づくことから、（複数のモダリティの）共有潜在特徴と呼ばれる（図 1 の●や■の点に相当する）。また、このような $z^{L/V}$ の集合が張る空間が図 1 に示した共有潜在空間となる。

2.2 手順理解に関する研究

手順とは目的を達成するために実行すべき工程の内容だけでなく、その実行順序の依存関係を記述したものとなる。したがって、手順は前節で述べたような単一の点ではなく、点と点の関係を示す枝を含む構造で表現される。

テキストや動画は基本的には同一軸上で順序付けられた、つまりシリアライズされたデータである。このため、これらは単純には図 1 に示したように直線的な遷移により表現される。

視覚情報は伴わないモノモーダルな自然言語においては、このような遷移を追跡し、工程ごとに引き起こされる材料の状態変化を潜在空間中の点の移動として表現し、それを追跡可能にする Neural process network（NPN）[1] という研究が知られている。この研究では、例えば「トマトを洗って、串切りにする」といった指示に対して「トマト」、「洗ったトマト」、「串切りにされたトマト」のそれぞれを潜在空間中の異なる点として表現する。また、「トマト」に対して「洗う」という工程を行うことで、潜在空間中のどこへ移動するかも予測する。これは、いわば潜在空間中での作業のシミュレーションとみなすことができる。

NPN はあくまでも材料の特徴変化を追跡する、つまり手順書に示された順序に従った図 1 のような直線的な遷移を予測するものである。一方で、手順書は本質的には手順の入れ替えが一部許された半順序構造を持つ。例えば図中のレシピではトマトとレタスのどちらから取り組んでも差し支えない（工程 1 と工程 2 には順序の依存性がない）。一方で、工程 3 は工程 1、2 の結果得られるものを混ぜる工程であり、両方終了していないと取りかかれない。このような「材料を混ぜる」ことによって生じる順序の依存関係は図 2 のようなグラフで表現される。自然言語処理による手順の理解には、このようなグラフ構造を抽出するものも存在する。

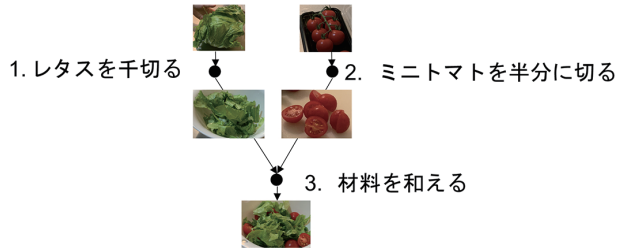


図 2　順序の依存関係のグラフ表記

このような手順に関する研究は、Web 上で大量に入手可能な調理レシピを対象に行われるのが主流である。これは調理レシピのユーザ数が、化学実験の手順書のユーザに対して圧倒的に多いこと、および、工業製品の組立手順書に対して機密性が低いことに起因している。

例えば、グラフ構造の抽出については、その構造の定義が粗い順から、SIMMR[8]、Action Graph[10]、Flow Graph Corpus[12,23] がある。いずれも、構造の定義だけでなく、そのような構造がアノテーションされたレシピのコーパスが公開されているが、このうち、特に Flow Graph Corpus は筆者を含む日本の研究グループによるもので日本語[12] と英語[23] のコーパスがそれぞれ公開されている。

2.3 クロスモーダル処理による手順理解研究

クロスモーダル処理による手順理解研究の先駆けは、深層学習が登場する以前の 2011 年に行われた道満らによるレシピと調理映像の整合*（alignment）[26] であろう。この研究では物体認識や動作認識結果を駆使してテキストに書かれたレシピの各工程に対応する映像区間を特定している。

深層学習以前であるため、共有潜在空間は図 1 のような連続的なものではなく、物体認識結果や動作認識結果の one-hot-vector による離散空間が用いられている。

これに対して、料理に関連したクロスモーダル処理で近年もっとも活発に研究されているのは im2recipe と呼ばれる、完成した料理の画像からそのレシピを特定する研究であろう。この研究の端緒は画像から対応するレシピを検索するもの[18] であったが、その精度向上を主な目的として、レシピに応じた料理画像の生成[24] を行う手法などが登場している。さらに Transformer の登場によって、これらの研究はさらにレシピ検索からレシピ生成へと拡張されてきている[17]。この中では、画像のみから前節で述べたようなグラフ構造の推定[22] を行うことにより精度向上を目指した手法なども提案されている。

しかし、完成した料理を見るだけでレシピを推定する問題は人間にも難しく、実応用を想定するにはやや不良設定な問題に思える。これらの研究に対して、我々は調理の過程を撮影した画像列や動画を入力としたレシピ生成に取り組んでおり、本稿では、それら 2 つの入力を対象とした最近の研究を紹介する。

3. 画像列からの手順書生成

一般的な視覚言語統合において、画像列を入力とした文章生成は視覚的叙述生成*（Visual Storytelling）[7] として知られている[11]。しかし、視覚的叙述生成は、対象とする画像列が物語的であり、どこまで逸脱した内容を正解として良いか定めることが難しい。また、複数の材料を追跡するモノづくりと異なり、物語の軸となる登場人物は絞られている。これに対して Chandu らは調理を対象とし、手順書と画像列からそれぞれ作業の状態遷移を有限オートマトン上でシミュレートし、その状態が一致するような損失を導入することでレシピ生成の精度向上を測る手法を提案した[2]。これは状態の遷移をモデル化し、その一致度を高めることで工程の進行度を反映した共有潜在空間を得ることを目指したものと解釈できる。

これとは独立に筆者を含む研究グループは大量のレシピ・画像列の対がある Cookpad Image Dataset[4] を利用し、検索課題による事前学習を利用する Retrieval Attention（RetAttn）[14] という手法を提案した（図 3）。この手法では、工程の違いまで含めて正確に検索できるようモデルを予め訓練することで、工程の進行度を反映した共有潜在空間を得る。同時に、得られた検索機能を利用し、画像に対応する文書の特徴を参照することで、画像と文書という異なる入力に起因する潜在特徴分布の偏り（モダリティギャップ）による悪影響を低減することで高い性能を達成している（詳細は次に紹介する手法と併せて表 1 で述べる）。

さらに、同グループは中間状態としてグラフ構造を推定し、そのグラフ構造をテキスト生成に積極的に用いる手法[15] を提案し、さらなる精度向上を実現した（図 4）。このモデルでは画像列から図 2 に示す依存関係を推定する。この推定結果を二値化して木構造を得たのち、デコーダに TreeLSTM[21] を採用し、推定された木構造を組み込んだテキスト生成を計算する。木構造の学習は半教師あり学習で行った。具体的には SIMMR を参考に、そのクロスモーダル版である Visual SIMMR データセットを作成し、全体の 1%程度のデータには木構造の正解を付与することで学習を行った。正解が存在しないデータに対しては、画像列から推定された木構造と生成されたテキストから推定される木構造が一致するような自己教示学習（Tree Re-prediction）を適用した。

表 1 に、これらの手法の定量的評価を掲載する。テキスト生成課題では、明確に 1 つの正解が定まらず、また全てのありえる正解の列挙も難しいことから単語の重複などを考慮した複数の指標（ワードオーバーラップ指標）によって近似的な評価が行われており、この表もその方法に従って得られたものとなっている。なお、実際に生成されたテキストを含む定性的評価を含む詳細な議論について我々の論文[14,15] を参照いただきたい。

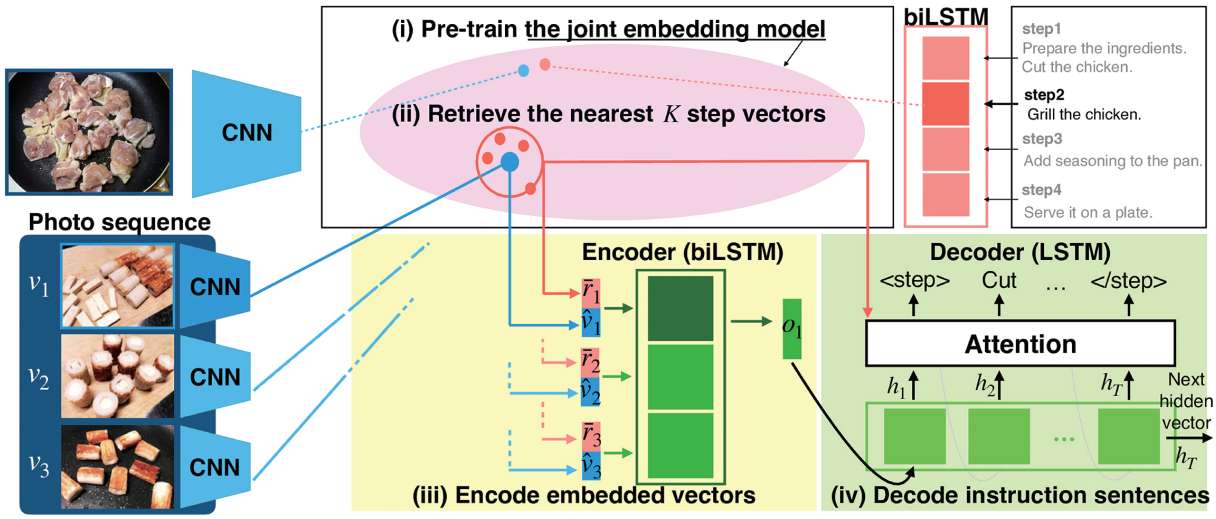


図 3　Retrieval Attention の概要．各画像から得られる潜在特徴と手順書から得られる各工程の潜在特徴が近くなるような距離学習を行い，共有潜在空間を学習する．テキスト生成では，各画像の潜在特徴に加えて，その画像の近傍にある（学習データ中の）手順書の特徴を平均したものを入力としてテキストを生成する（文献14）より引用）

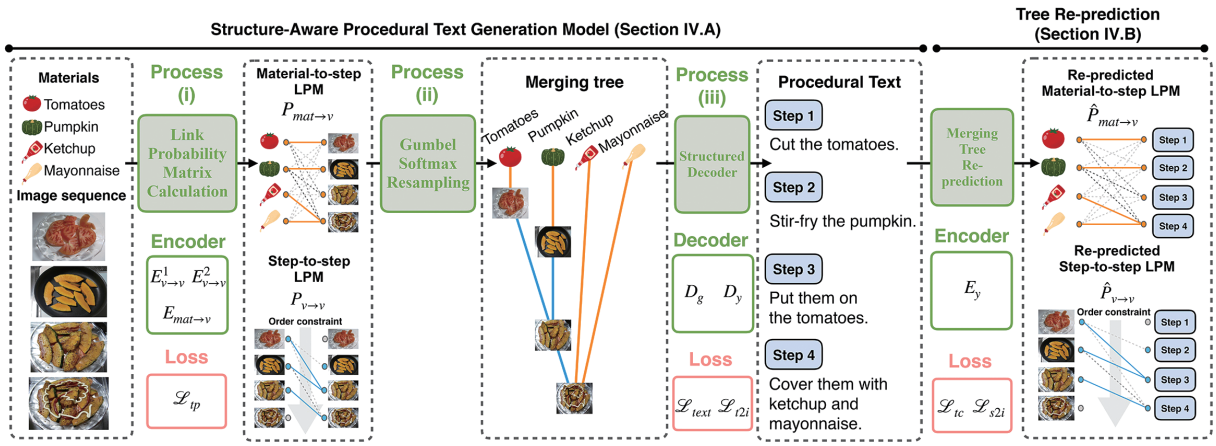


図 4　依存関係を考慮したテキスト生成モデルの概要．入力は材料リスト（記号の集合）と画像列となる．Process（i）では材料－画像，および，画像－画像間の依存関係を推定する．Process（ii）では Gumbel Softmax Resampling というテクニックを用いて微分可能な形で推定結果を離散化し，木構造を得る．Process（iii）では Tree LSTM により木構造を反映したテキスト生成計算を行う．最後に Tree Re-prediction では生成テキストからも木構造を推定し，画像から推定された木構造との一致度が高くなるような自己教示型の木構造推定を行う（文献15）より引用）

表 1　3 つのベースライン手法と，それぞれを[15] による構造考慮型モデルに拡張したモデルに対するワードオーバーラップ指標による評価（+Half と +Full は図 4 の Tree Re-prediction 部の有無の違い）．B$=$BLEU, RL$=$ROUGE-L, D$=$Distinct であり，食材や動作に相当する単語のみを抜き出した単語列による評価をさらに I$=$Ingredient, Ac$=$Action を付して表記している．また手法別の比較で最良のスコアを太字で，手法内の比較で最良のスコアを下線で表している（文献15）より抜粋）

	B1	B4	RL	D1	D2	I-B1	I-B4	I-RL	Ac-B1	Ac-B4	Ac-RL
GLAC Net［11］	28.5	5.9	21.4	46.6	69.0	9.5	0.3	13.2	21.2	2.1	22.8
+Half	28.5	5.9	21.8	46.7	68.8	10.5	0.4	15.6	21.0	2.3	23.1
+Full	28.9	6.1	21.3	47.2	69.9	11.8	0.4	16.3	21.0	2.3	22.5
SSiD［2］	28.7	6.0	20.9	45.5	66.6	8.6	0.2	13.1	20.1	2.1	21.6
+Half	30.3	6.2	20.8	43.9	65.1	11.4	0.4	16.4	19.8	2.2	22.1
+Full	31.1	6.4	21.6	48.3	71.0	12.9	0.4	16.7	21.5	2.2	23.5
RetAttn［15］（Ours）	32.2	6.5	21.6	40.2	60.3	11.2	0.3	14.5	22.1	9.0	21.2
+Half	32.2	6.5	21.8	52.4	77.8	11.9	0.3	14.8	21.8	9.4	22.9
+Full	33.2	7.1	22.1	52.7	78.6	12.1	0.3	14.8	22.3	9.5	23.1

各手法単独では RetAttn が最も性能が高く、この手法が最高性能を達成していることがわかる。また、構造考慮型モデルへの拡張がモデルを問わず性能向上に寄与することも読み取れる。

この他、同グループではデータ構造という観点では画像列中の食材矩形と、Flow Graph 中の要素（つまり調理途中状態の食材）との対応関係をアノテーションした Recipe FlowGraph Bounding Box（RFG-BB）データセット[16] も公開しており、今後の活用を検討している。

4. 動画と手順書のクロスモーダル処理

前節までで紹介してきた画像列を入力とした手順理解では「材料の変化」を追跡することにその主眼が置かれてきた。これに対して、動画があれば、「材料の変化」を引き起こす作業者の動作そのものも観測対象となり、「動作」と「材料の変化」の関係性獲得が技術的な挑戦となる。

この問題に対して Naim らは化学実験のように手順が厳格に定義され、図 2 に示した順序の曖昧性は存在しない状況を仮定し、教師なしでの手順書と動画中のシーンの対応付けを目指した[13]。この研究では、把持された物体と文書中の名詞との対応を隠れマルコフモデルにより確率的に定式化し、各シーンでの名詞と物体の対応関係の一貫性と、映像全体を通した手順の再現性の双方が最尤となる対応付けを EM アルゴリズムによって推定する手法を提案している。また、調理を対象として、深層学習を用いながらも基本的な発想としては類似した取り組みが後年に Stanford 大の研究グループによって行われている[3,5,6]。残念ながら精度という観点では、これらの取り組みはあくまで実験的なものと言わざるをえない結果となっている。これは、おそらく教師情報やデータ量の不足が根本的な原因であろう。

これらは手順書と作業映像との対応付けに関するものであったが、編集済みの動画から手順書を生成する取り組みも発表されている[19,20]。ただし、これらの研究では、手順書を生成するために、映像中の音声書き起こしも入力に加えており、この書き起こしから得られる言語的な情報に大きく依存したモデルとなっている。このような外部情報を削減しながら、映像から抽出できる情報を増やしていくことが、モノづくりに対するクロスモーダルな手順理解研究にとっての次の課題となっている。

5. むすび

本稿では視覚言語統合技術を用いて、手順の共有潜在表現を抽出するための諸技術を概観した。現状ではデータセットの不足や重要シーンのみを抽出することの困難さから、人が編集し、重要シーンのみに分割されたような動画に対する技術しか研究が進んでいない。また、組み合わせによる特徴変化を追跡するモデルは構築できているものの、単独で材料の状態を変化させるような作業に対して、その状態変化を追跡する技術も未解決である。これらについては現在、問題解決に向けての取り組みをそれぞれ進めている。

参考文献

1）Bosselut, A.; Ennis, C.; Levy, O.; Holtzman, A.; Fox, D.; Choi, Y. “Simulating action dynamics with neural process networks”. ICLR. 2018.

2）Chandu, K.; Nyberg, E.; Black, A. W. “Storyboarding of recipes: grounded contextual generation”. Proc. ACL. 2019, p.6040–6046.

3）Chang, C.-Y.; Huang, D.-A.; Sui, Y.; Li, F.-F.; Niebles, J. C. “D3TW: Discriminative differentiable dynamic time warping for weakly supervised action alignment and segmentation”. CVPR. 2019.

4）Harashima, J.; Someya, Y.; Kikuta, Y. “Cookpad image dataset: An image collection as infrastructure for food research”. SIGIR. 2017, p.1229–1232.

5）Huang, D.-A.; Buch, S.; Dery, L.; Garg, A.; Li, F.-F.; Niebles, J.C. “Finding “it”: Weakly-supervised reference-aware visual grounding in instructional videos”. CVPR. 2018.

6）Huang, D.-A.; Lim, J. J.; Li, F.-F.; Niebles, J. C. “Unsupervised visual-linguistic reference resolution in instructional videos”. CVPR. 2017.

7）Huang, T.-H. K.; Ferraro, F.; Mostafazadeh, N.; Misra, I.; Agrawal, A.; Devlin, J.; Girshick, R.; He, X.; Kohli, P.; Batra, D.; Zitnick, C. L.; Parikh, D.; Vanderwende, L.; Galley, M.; Mitchell, M. “Visual storytelling”. NAACL-HLT. 2016, p.1233–1239.

8）Jermsurawong, J.; Habash, N. “Predicting the structure of cooking recipes”. EMNLP. 2015, p.781–786.

9）R. F. Karlin. “Defining the semantics of verbal modifiers in the domain of cooking tasks.” ACL. 1988, p.61–67.

10）Kiddon, C.; Ponnuraj, G. T.; Zettlemoyer, L.; Choi, Y. “Mise en place: Unsupervised interpretation of instructional recipes”. EMNLP. 2015, p.982–992.

11）Kim, T.; Heo, M.-O.; Son, S.; Park, K.-W.; Zhang, B.-T. GLAC Net: Glocal attention cascading networks for multi-image cued story generation. arXiv preprint arXiv:1805.10973, 2018,（参照 2019–2–13）.

12）Mori, S.; Maeta, H.; Yamakata, Y.; Sasada, T. “Flow graph corpus from recipe texts”. LREC. 2014, p.2370–2377.

13）Naim, I.; Song, Y. C.; Liu, Q.; Kautz, H.; Luo, J.; Gildea, D. “Unsupervised alignment of natural language instructions with video segments”. AAAI. 2014, p.164–174.

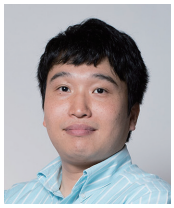
14）Nishimura, T.; Hashimoto, A.; Mori, S. “Procedural text generation from a photo sequence”. INLG. 2019, p.409–414.

15）Nishimura, T.; Hashimoto, A.; Ushiku, Y.; Kameko, H.; Yamakata, Y.; Mori, S. Structure-aware procedural text generation from an image sequence. IEEE Access. 2020, Vol.9, p.2125–2141.

16）Nishimura, T.; Tomori, S.; Hashimoto, H.; Hashimoto, A.; Yamakata, Y.; Harashima, J.; Ushiku, Y.; Mori, S. “Visual grounding annotation of recipe flow graph”. Proceedings of the 12th Language Resources and Evaluation Conference. European Language

Resources Association, 2020, pp.4275–4284.
17）Salvador, A.; Drozdzal, M.; Nieto, X. G.; Romero, A. “Inverse cooking: Recipe generation from food images”. CVPR. 2019.
18）Salvador, A.; Hynes, N.; Aytar, Y.; Marin, J.; Ofli, F.; Weber, I.; Torralba, A. “Learning cross-modal embeddings for cooking recipes and food images”. CVPR. 2017
19）Shi, B.; Ji, L.; Liang, Y.; Duan, N.; Chen, P.; Niu, Z.; Zhou, M. “Dense procedure captioning in narrated instructional videos”. ACL. 2019, p.6382–6391.
20）Shi, B.; Ji, L.; Niu, Z.; Duan, N.; Zhou, M.; Chen, X. “Learning semantic concepts and temporal alignment for narrated video procedural captioning”. ACMMM. 2020, p.4355–4363.
21）Tai, K. S.; Socher, R.; Christopher D. “Manning. Improved semantic representations from treestructured long short-term memory networks”. Proc. ACL-IJCNLP. 2015, p.1556–1566.
22）Wang, H.; Lin, G.; Hoi, S. C.; Miao, C. “Structure-aware generation network for recipe generation from images”. ECCV. Springer, 2020, p.359–374.
23）Yamakata, Y.; Mori, S.; Carroll, J. “English recipe flow graph corpus”. Proceedings of the 12th Language Resources and Evaluation Conference. European Language Resources Association, 2020, p.5187–5194..
24）Zhu, B.; Ngo, C.-W.; Chen, J.; Hao, Y. “R2GAN: Cross-modal recipe retrieval with generative adversarial network”. CVPR. 2019.
25）原島純，橋本敦史．キッチン・インフォマティクス―料理を支える自然言語処理と画像処理―．オーム社，March 2021.
26）道満恵介，高橋友和，井手一郎，村瀬洋ほか．マルチメディア料理レシピ作成のための料理レシピテキストと料理番組映像との対応付け．電子情報通信学会論文誌 A. 2011, Vol.94, No.7, p.540–543.

執筆者紹介



橋本 敦史　HASHIMOTO Atsushi
オムロン サイニックエックス株式会社
リサーチアドミニストレイティブディビジョン
専門：画像処理、機械学習、HCI
所属学会：IEEE、電子情報通信学会、情報処理学会
博士（情報学）

本文に掲載の商品の名称は、各社が商標としている場合があります。

*1 漢字の意味を全く知らない人が中国語の辞書が山のように積まれた部屋の中にいるとする（字が同じかどうかはなんとか判定できる）。
部屋には手紙をやり取りするための穴があり、そこには中国語で書かれた質問が書かれている。
部屋の中の人は言葉の意味がわからないながらも辞書を参照して質問に答えることができるかも知れない。
部屋の中の人が質問に答えられるとして、これは果たして質問の意味を理解したといえるだろうか?という問い。

*2 この分野はまだ邦訳が定まっていないものが多いため、本稿では25）での訳語を用いる際には初出時に * を付与する。