6ZC-09

キーワードによる画像処理システム

東 夢太† 久保田 光一‡
中央大学大学院 理工学研究科† 中央大学 理工学部‡

1 はじめに

本研究では, 画像から特定の物体を消去し, 消去した空間を新たに補間するシステムの開発を行った. 画像加工では多くの場合, 処理領域をユーザーがマウス等で細かく指定する必要がある. これは複数の画像中の同一カテゴリの物体に対して同一の処理を一括で行いたい時などに手間である. そこで, 本研究ではユーザー入力の軽減および単純化を目的として, 自然言語による画像処理を最終目標と定め, 今回はその足がかりとしてキーワードによる処理領域の指定を行った. また, 画像補間の際, 画像からその説明文および Scene Graph を生成し, それらを利用することによって, より違和感のない補間の実現を目指した.

2 提案システム概要

提案システムは, 画像とキーワードを入力として受け取り, キーワードによって指定された領域に何らかの処理を施したものを最終結果として出力する. 今回はその処理の一例として, 指定領域を削除し補間を行う. 以下に処理の流れを示す.

(a) 画像から物体のカテゴリおよびその位置を検出 (YOLOv3 [1])
(b) キーワードによる処理領域の指定
(c) 指定領域と他の物体領域との干渉判定
(d) 干渉領域を含む画像領域の切り出し
(e) d. の画像からキャプションを生成 (OBJ2TEXT [2])
(f) キャプションから Scene Graph [3] を生成
(g) Scene Graph から得た位置関係を利用し処理領域を決定
(h) 処理領域内を削除し補間

3 提案システムで用いた手法

3.1 物体検出 (a)

まず, キーワードによる処理領域の指定を行うために, 入力画像から物体のカテゴリおよびその位置の検出を行う. 提案システムではその手法として YOLOv3 [1] を採用した. 図 1 は検出結果の一例である.

3.1.1 YOLO [4]

YOLO とは物体検出アルゴリズムの一種で, その構造はひとつの畳み込みニューラルネットワーク (CNN) で完結しており, 画像から物体領域の推定と, その物体のクラス分類を同時に行えることが大きな特徴となっている. その結果として得られる物体領域は, バウンディングボックス (矩形領域) で表現される. 今回用いた YOLOv3 は, 高速かつ高精度で物体検出を行うことが可能である. また, 本システムにおいて, ネットワークのパラメータは MS COCO 学習済みモデルを利用した.

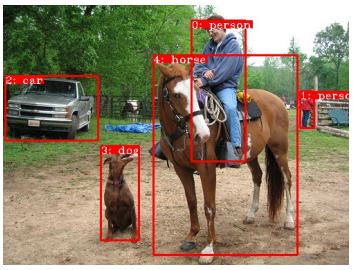

図 1 YOLOv3 による物体検出例

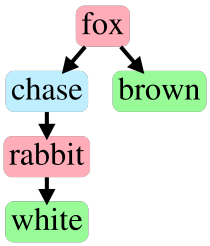

図 2 Scene Graph 例

3.2 キーワードによる処理領域の指定 (b ～ d)

次に, 入力されたキーワードから処理領域の決定を行う. まず, 物体カテゴリが入力キーワードと一致する矩形領域を初期領域として設定する. 初期領域が他の物体領域と干渉していない場合は, そのまま処理領域として決定され, 以降の処理は行わない. 干渉する物体領域が存在する場合は, 初期領域および干渉領域全てを含む画像領域を切りだし, その画像に対して以降の処理 (e ～ g) を行い, 処理領域の更新を行う.

3.3 画像キャプション生成 (e)

キーワードで指定された処理領域が他の物体領域と干渉していた場合, 物体同士の関係性を取得するため, まず画像からキャプションを生成する. ここで, 本研究における「画像キャプション」とは, 「その画像が表す場面や状況を文章化したもの」とする. キャプション生成には, OBJ2TEXT [2] を用いた.

3.3.1 OBJ2TEXT

OBJ2TEXT は, 画像から文章を生成する sequence to sequence モデルである. 一般的な対話モデルでは発話文をベクトルに変換し, それを入力として応答文を生成する. この OBJ2TEXT では発話文のかわりに, CNN を用いて画像から抽出した特徴ベクトルを入力として文章を生成する. 文章の生成には, リカレントニューラルネットワークの一種である LSTM が用いられる. こちらも, パラメータには MS COCO 学習済みモデルを利用した.

3.4 Scene Graph の生成 (f)

物体の関係性を解析するため, 節 3.3 で得られたキャプションから Scene Graph [3] の生成を行う. Scene Graph 生成の道具として, Stanford Scene Graph Parser を利用した.

3.4.1 Scene Graph

Scene Graph とは, ある場面を表現するグラフ構造であり, 各ノードは object, attribute, relationship, いずれかの属性をもつ. 図 2 は, “A brown fox chases a white rabbit.” という文章の Scene Graph である. 各ノードの色は赤が object, 緑が attribute, 青が relationship に対応している.

3.5 処理領域の選択 (g)

節 3.4 で生成した Scene Graph の relationship を使用する. 本システムでは, グラフで上位に存在する物体ほど画像に対する影響度が高いと考え, その物体がより前景に存在すると判別する. 節 3.1 で検出されたカテゴリと Scene Graph の単語が完全には一致しない場合は, それらの類似度によっ

Image processing system with keywords
† Yumeta HIGASHI, Graduate School of Science and Engineering, CHUO University
‡ Koichi KUBOTA, Faculty of Science and Engineering, CHUO University

てそれらの紐付けを行う．このとき，類似度の指標には Path Similarity を用い，計算の際にシソーラスは WordNet [5] を使用した．

3.5.1 Path Similarity

Path Similarity は，シソーラスの階層構造を利用し，あるワード w_1 のもつ意味 s_1 と，あるワード w_2 のもつ意味 s_2 間の最短パスを求めることで計算できる．一般的に，より最短パスが短いほど s_1 と s_2 は類似度が高いと考えられている．

3.6 画像の補間 (h)

本研究の主とするところは処理領域の指定までではあるが，その後の画像処理の一例として，処理領域部分を削除し，空いた空間を自然に補間するような処理の実装も行った．PatchMatch [6] を用いた類似パッチ (小領域) の探索により，これを実現している．具体的には，欠損領域を含む各パッチに対してそれぞれ類似パッチを同一画像中から探索し，欠損領域中の各ピクセルに対してそれらの色情報の平均を適応して補間を行う．図 3 および図 4 は，それぞれ一部を削除した画像，欠損領域を補間した画像である．

3.6.1 PatchMatch

ある画像内の注目パッチに対して，類似度の高いパッチを類似パッチと呼ぶ．類似度評価には，画素値の差分距離による Sum of Squared Difference (SSD) という指標が一般的に用いられる．PatchMatch は，類似パッチを探索するためのランダム近似最近傍探索アルゴリズムである．全探索と比べ，大幅に探索コストを削減することができる．

図 3　欠損画像

図 4　補間結果

4 実験

指定した領域と重なる物体が存在しない場合は，単純にその領域がそのまま処理領域となるため，その結果はここでは省略する．図 5, 図 6 はそれぞれ上から元画像，キーワードで指定された処理領域を削除した画像，欠損領域を補間した画像，である．図 5 では “a man riding a horse in field .” というキャプションが生成され，“man ride horse” という関係から “person” が “horse” より前景であると判別している．図 6 では “a man holding a dog on a leash .” というキャプションが生成され，“man hold dog” という関係から “person” が “dog” より前景であると判別しているが，実際には “dog” が前景であると判別するのが適切であると考えられる．

4.1 考察

この 2 つの結果を比較するだけでも，グラフで上位の物体が前景というのは必ずしも正しいとは限らないことがわかる．各 relationship に対して優先度や前後関係を個別に設定することで，この問題は解決可能であると考えられる．また，生成したキャプション中に指定した物体が現れない場合や，別の物体として扱われてしまう場合も多く見られた．これには，物体の一部が隠れて十分な情報が得られなかった可能性や，頻出する関係性，あまり見られない関係性，など，学習データに偏りがあった可能性が考えられる．

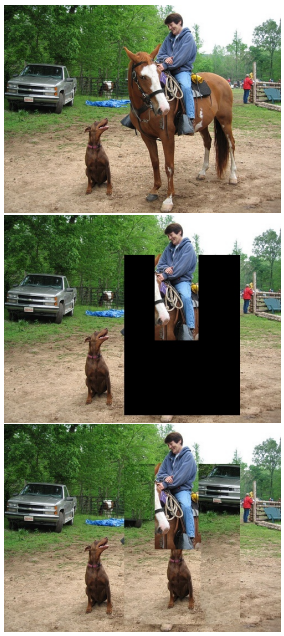

図 5　“horse” で指定

図 6　“dog” で指定

5 今後の課題

本システムは同じカテゴリの物体同士が重なる場合に対応することができない．物体の詳細な情報まで検出することができれば，Scene Graph の attribute を利用することによってこの問題の緩和が可能になると考えられる．また，物体領域を矩形ではなく，セマンティックセグメンテーションによって詳細に得ることで，より柔軟な領域指定が期待できる．

参考文献

[1] Joseph Redmon and Ali Farhadi. YOLOv3: An incremental improvement. *CoRR*, abs/1804.02767, 2018.

[2] Xuwang Yin and Vicente Ordonez. OBJ2TEXT: generating visually descriptive language from object layouts. *CoRR*, abs/1707.07102, 2017.

[3] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Image retrieval using scene graphs. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3668–3678, 2015.

[4] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *CoRR*, abs/1506.02640, 2015.

[5] George A. Miller. Wordnet: A lexical database for English. *Commun. ACM*, 38:39–41, 1995.

[6] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B. Goldman. Patchmatch: a randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3):24:1–24:11, 2009.