

QuIC-360°: 360° 画像に対する クエリ指向画像説明文生成のためのデータセット構築

前田 航希^{1,3} 栗田 修平¹ 宮西大樹^{2,1}

¹ 理化学研究所 AIP, ² 国際電気通信基礎技術研究所 (ATR), ³ 東京工業大学
{koki.maeda, shuhei.kurita}@riken.jp, miyanishi@atr.jp

概要

360° 画像は一般的な画像と比較して、撮影者による情報の取捨選択が行われないため、多くのコンテキストを同時に含む。既存の画像説明文生成では、コンテキストを画像情報のみから読み取るが、360° 画像に対しては、画像に加えて補助的な情報を付加することで、記述するコンテキストを指定することが必要になる。本研究では、画像に加えて言語情報(クエリ)を与えることで説明文生成を制御する**クエリ指向説明文生成**を提案し、そのためのデータセットとして 5,800 枚の 360° 画像と 22,956 文の説明文からなる **QuIC-360°** を構築した。QuIC-360° による再学習で、360° 画像に対してクエリを用いることで説明文生成の制御性・多様性が高まることが確認された。

1 はじめに

画像説明文生成は、画像に含まれる内容を自然言語で自動的に記述する課題である [1]。近年の画像説明文生成では、Attention 機構を導入するなどコンテキスト、すなわち画像の中で記述されるべき特徴への注目を行う手法が提案されてきた [2]。Cornia ら [3] は画像の注目すべき領域の集合や順序を補助的な情報として与えることで生成される説明文を制御することを提案している。これまでの画像説明文生成では、主にモデル作成側の設計により画像情報からコンテキストへの注目が行われている。Microsoft COCO Captions [4] などの既存のデータセットに含まれる画像では、撮影時に視野角の制約から対象の選択が行われている。撮影された画像のコンテキストは限定的で、大きく異なるコンテキストが複数含まれることは想定されてこなかった。

一方で、全方位カメラによって撮影された 360° 画像では、通常画像にある視野角の制約がなく、対

象物の選択が行われないため撮影された情景(シーン)全方向への情報を含む。360° 画像を利用したシーンからの説明文生成は、ライブカメラや監視カメラといったシーンカメラを用いたテキスト実況や、自動運転車や自動配達ロボット等に付属したカメラが撮影した画像を用いたテキスト伝達 [5] などに応用できる。画像の説明文生成で 360° 画像の特性を活かすには、通常画像と比較して注目するコンテキストやユーザーの需要に応じて説明文生成を制御することがより重要になる。

本研究では、画像に加えて単語や短いフレーズといった言語情報をクエリをとして入力することで説明文生成を制御する**クエリ指向画像説明文生成**を提案し、データセットとして **QuIC-360° (Query-based Image Captioning for multi-context 360° images)** を構築した。図 1 に課題の概要を示す。天気や花壇、周囲の往来など複数のコンテキストを持つ画像に対して、注目する対象に応じて異なる画像説明文が記述されることが期待される。しかし、一般的な説明文生成では、画像全体を言及するが目的に合わせた詳細な記述が行われない。提案課題は「人々」や「風景」といった具体的な記述対象を事前に与え、より具体的な説明文を作成することを目的としている。

我々は 5,800 枚の 360° 画像に対して、クエリに基づいた 22,956 文の説明文を収集した。既存の最先端の画像説明文生成モデルを QuIC-360° を用いて再学習させることで、(i) 説明文生成の制御性が高くなること、また、(ii) 生成された説明文はより多様であったことが確認された。クエリ指向説明文生成は、既存手法が抽象的な説明文を生成する問題を解決できる。構築したデータセットは一般に公開されている¹⁾。

1) <https://github.com/Silviase/QuIC-360>

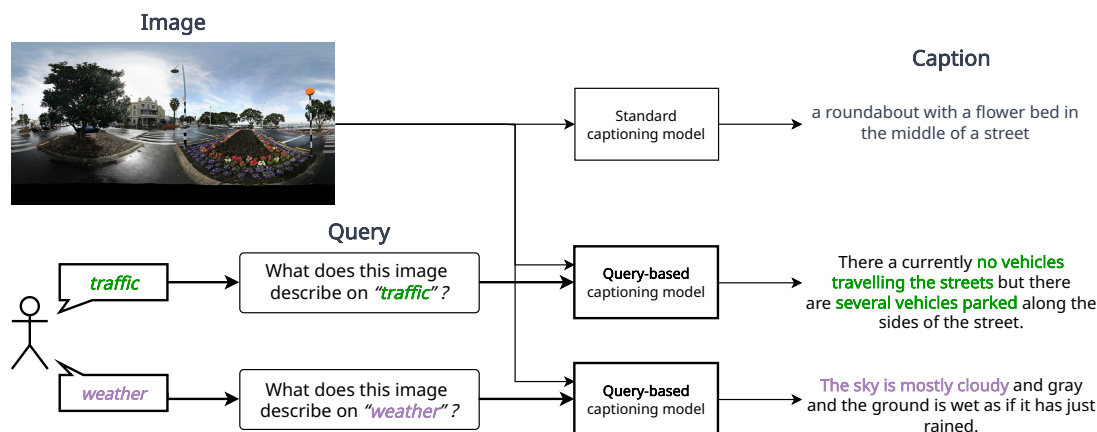


図1 クエリ指向画像説明文生成 (QuIC) の概要.

2 関連研究

2.1 画像説明文生成の制御

説明文生成も他の多くの言語課題と同じように、Attention 機構の利用によって研究が大きく進展した [2]. モデルの進歩につれて、ユーザが利用する用途や文脈によって、同じ画像について異なる説明文が求められるようになった. 近年では与えられた画像に記述される対象の優先順位を付与して説明文生成を制御する研究が行われている. 付与された情報は、生成過程でモデルに補助となる特徴として利用される. 補助的な情報の例として、画像に付与された領域 [3, 6] やマウスの軌跡 [7], 抽象シーングラフ [8] などがある.

QuIC も画像の補助的な情報として言語情報を付与するという点で制御可能な説明文生成手法の一つである. しかしながら、QuIC は画像に関する情報ではなく、自然言語を用いた言語情報をシグナルとして与える点で既存の研究と異なる.

2.2 視覚的質問応答

視覚的質問応答は、画像と質問を入力として、それに対する回答を自動的に生成する課題である [9]. 画像の局所的な部分に着目し適切な記述を行うという点では本研究に類似しており、360° 画像を用いた研究も行われている [10, 11]. しかしながら、視覚的質問応答は質問に対応する回答の長さが短い傾向があり [12], 実際に視覚的質問応答は事前に定められた限られた回答語彙の中からの選択モデルとして動作させる事が多い. 本研究は、視覚的質問応答における回答が限られていることを念頭に置き、単語

および数語の自然言語で与えられたクエリに応じて画像の説明文を制御する手法を提案する.

2.3 360° 動画画像データセット

全方位カメラで撮影された動画画像を収集したデータセットは数多く構築されてきた [13, 14, 15, 16]. しかし、既存のデータセットは撮影環境が室内に限られたもの [17, 18] や、車載カメラや HMD によって撮影されたもの [19, 20, 5, 21] など画像の対象や環境に偏りがみられる. 本研究に近いデータセットとして、60,000 枚以上の屋内・屋外で撮影された 360° 画像からなる SUN360° [22] があるものの、現在では公開が停止されている²⁾. 本研究では、新たにインターネット上から多様な画像を収集し、画像の取得方法と付与された説明文を公開することでデータセットを一般に利用可能な状態とする.

3 QuIC-360° データセット

3.1 360° 画像の収集

これまで作成されてきたデータセットは、そのシーンや目的に応じて画像のドメインを制限してきた. 一般的に、屋内画像は物体が密集しており、屋外画像は撮影者を中心とした特定の活動を捉えた画像が多く、物体は疎である [23]. しかし、どのシーンにおいても全方位を撮影した画像について、コンテキストは一つに限られることは少ない. したがって、我々は撮影地点を区別せず画像を収集した. また画像の可用性を確保するために、これまで開発されてきたデータセットに含まれる 360 度画像を利用

2) 後述のように Refer360° として公開されている一部の画像以外は公開されておらず、著者はメールに反応しなかった.

することを避け、新たに画像を収集した。画像はインターネット上の画像投稿サイトの Flickr を用いて 12,930 枚の画像を収集した。

データセットの品質を高めるために、以下の手順で収集した画像をフィルタした。まず、画像の形式を統一するために正距円筒図法で保存されている画像以外を除去した。次に、同一地点での類似した画像を取り除くことを目的として同一ユーザの投稿時刻の差が 1 時間以内であるものを除去した。さらに、画像の多様性と可用性を最大化するために、同一ユーザの投稿が 100 枚以下となるようランダムに画像を破棄して、合計で 3,800 枚の画像を取得した。

また、補助的なデータセットとして Refer360° [24] に含まれる 2,000 枚の画像を利用した。画像は現在では利用不可能となっている SUN360° [22] の一部である。本研究ではこれらの画像と付与された説明文は全て追加の訓練データとして扱う。

3.2 説明文の付与

収集した画像に付与される説明文は、画像の特徴を捉えて詳細に記述されていることが望ましい。クエリは詳細な記述を行うための指針として重要である。既存の画像の分類カテゴリはデータセットによって様々であり、標準的な分類が存在しない。我々は実際に収集した画像を人手で確認し、画像を記述するためのクエリを 34 種類作成した。

収集した画像への説明文の付与のために、Amazon Mechanical Turk (MTurk) を用いて評価者を雇用した。評価者は異なる 3 つのクエリを選択し、そのクエリに関連する説明文を作成する³⁾。単純な説明文を抑制するため、評価者に 8 語以上での回答を要求した。不適切な回答を自動および人手で破棄し、合計 22,956 文の説明文を収集した。

3.3 統計量

QuIC-360° は 5,800 枚の画像に、34 個のクエリに関する 22,956 文の説明文を付与したデータセットである。画像に対してそれぞれのクエリが選択された回数は表 1 の通りである。訓練・検証・評価データの分割は表 2 に示すように行った。評価データについては画像・クエリの組に対して 4 件以上の説明文が付与されるようにした。同一クエリに対する説明文数を増やすために、評価データの各画像は初め

location	2,011	art	443	happenings	167
people	1,950	plants	372	garden	147
weather	1,477	activity	343	fashion	138
building	1,454	paintings	320	corridor	124
interior	1,426	mountains	271	foods	121
furniture	1,085	street	260	animals	70
trees	988	monuments	254	traffic	57
architecture	850	rivers	210	trains	50
landscape	696	lake	204	planes	36
what they are doing	653	sea	204	accidents	13
small objects	624	houses	181		
cars	564	stores	177		

表 1 それぞれのクエリが画像に対して選択された回数。

Split	# images	# captions	# vocab.	Avg. Length
Train	3,007	9,459	8,106	21.9
Valid	400	1,251	3,097	21.9
Test	393	6,246	6,127	20.0
Refer	2,000	6,000	5,190	20.7
Total	5,800	22,956	11,804	21.1

表 2 QuIC-360° におけるデータセット分割と統計量

の評価者が選択したクエリを指定して 2 人目以降の評価者に説明文を付与させた。

4 実験

4.1 実験設定

クエリ指向説明文生成は、画像に加えて言語情報としてクエリを入力とする。そのため、通常の説明文生成モデルと入力形式が異なり、全ての手法に対して即座に利用できない。そこで、検証を行うための説明文生成モデルとして説明文生成課題で最先端の性能を持つ OFA [25] を採用した。

OFA は Transformer をもとにしたエンコーダ・デコーダモデルであり、種々の言語および画像を用いた課題を統合した機構である。OFA における説明文生成では、生成モデルへの入力として画像情報に加えて “What does this image describe?” というプロンプトを与える。提案手法ではプロンプトに追加する形でクエリを補助的な情報として与えた。具体的には、クエリが “location” であれば “What does this image describe **on location**?” とした。公開されている事前学習済みモデルを利用した。ただし、利用したモデルは 360° 画像を用いた大規模な事前学習を行っていないことは留意すべきである。

再学習による効果を検証するために、事前学習済みモデルに対して QuIC-360° による (i) 再学習しない (None), (ii) 再学習したがクエリを補助的な情報として与えない (FT w/o query), (iii) 再学習を行いク

3) Appendix A にアノテーションに用いたページのスクリーンショットを示す。

Finetune	Query(Inference)	Vocab.	Avg. Length	BLEU@4	METEOR	ROUGE-L	CIDEr	SPICE
None	×	507	9.7	4.9	8.7	23.1	8.4	6.5
FT(w/o query)	×	459	16.8	10.5	12.4	29.1	14.2	7.2
FT(w/ query)	×	529	16.7	10.5	13.0	29.4	15.5	8.1
None	✓	549	10.0	5.3	8.9	23.8	9.3	6.8
FT(w/o query)	✓	626	15.0	13.9	13.9	32.3	20.3	9.2
FT(w/ query)	✓	699	16.5	18.7	17.5	38.0	30.3	12.8

表3 QuIC-360° の評価データを用いた説明文生成性能の比較.



Method	Query	Caption
FT(w/ query)	location	The image is set in the middle of a busy city street filled with tall buildings that are lit up
	street	This is a busy city street in the middle of the city
	cars	The street is filled with cars and people
	cars	There are a lot of cars that are driving down the street
FT(w/o query)	location	a crowd of people on a red carpet in a city at night
	street	a crowd of people on a red carpet in a city at night
	cars	a crowd of people on a red carpet in a city at night

表4 QuIC-360° を用いた再学習による出力結果の変化. 再学習前はクエリによらず同様の出力を行うが、再学習を行うことでクエリによって説明文生成を制御できる.

エリも与えた (FT w/ query), の3つの条件について性能を比較した. 3.3 節で構築したデータセットの訓練/検証/評価データセットの分割を用いて, 説明文に含まれる各単語に対する交差エントロピー誤差を損失関数として学習を行った. ハイパーパラメータの設定は公開されている実装⁴⁾を利用した.

4.2 定量的評価

QuIC-360° を用いた再学習によって, クエリを用いて生成される説明文を制御できるか検証した. また, 比較対象とする3つの場合において推論時にクエリを与えるかどうかで性能が変わるか検証した.

評価指標として画像説明文生成で一般的な指標である BLEU [26], ROUGE [27], METEOR [28], CIDEr [29], SPICE [30] を用いた. 加えて, 多様な説明文生成が行われているかどうかを示す指標として, 生成した文の平均単語数と語彙数を比較した.

実験結果を表3に示す. 全ての評価指標において, QuIC-360° によるクエリ指向説明文生成課題での再学習を行った場合が最も高い性能を示した. また, 平均単語数や語彙数も増加することが確認された. 一方で, 推論時にクエリを与えない場合, 提案手法は性能向上に寄与しないことがわかった.

以上から, QuIC-360° を用いた再学習によって, 説明文生成の制御性・多様性を高められることが確認された. しかしながら, 人手で付与された説明文と比較して生成された説明文の語彙数は少なく, 文長も短い. 人間のような説明文を生成することは現

時点では挑戦的な課題であるといえる.

4.3 定性的評価

クエリ指向説明文生成は一般的な説明文生成と異なり, 同じ画像が入力されたとしてもクエリに応じて説明文を制御する必要がある. そこで, QuIC-360° での再学習を行った場合とそうでない場合について, 出力した説明文を比較した.

表4に示すように, 一般的な説明文生成課題による再学習を行った場合, 推論時にどのクエリを与えてもクエリを無視して同じ文を生成した. これは画像説明文生成において, コンテキストを画像情報のみから選択していることの証左である. 提案手法では, クエリに応じて適切に出力が変化したことから, 言語情報によってコンテキストの選択を行い, 説明文の生成を制御できることが確認された.

5 おわりに

本研究では, クエリを用いて画像の説明文生成の制御を行う課題を提案し, 360° 画像を利用したデータセットとして QuIC-360° を構築した. 複数のコンテキストを含む 360° 画像に対して, ユーザが言語情報を補助として与え, より需要に即した説明文を生成するよう制御することを目的としている. 実験では, QuIC-360° によって説明文生成モデルを訓練することで, 360° 画像の説明文生成の制御性と多様性を高められることを定量的・定性的に確認した. 今後の発展として, より大規模なデータセットの構築・360° 画像に最適化した手法の作成などがある.

4) <https://github.com/OFA-Sys/OFA>

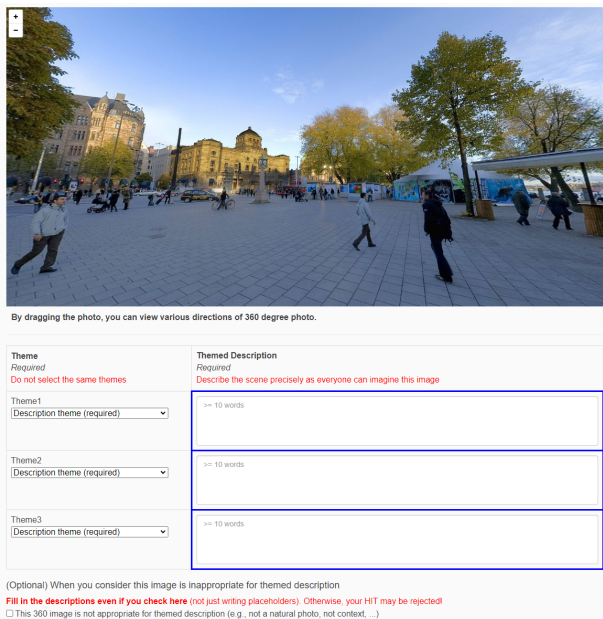
謝辞

本研究は JST さきがけ JPMJPR20C2, JSPS 科研費 22K17983, JSPS 科研費 JP20269633 の助成を受けたものです。

参考文献

- [1] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In **CVPR**, pp. 3156–3164, 2015.
- [2] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In **ICML**, pp. 2048–2057, 2015.
- [3] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions. In **CVPR**, 2019.
- [4] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. **arXiv:1504.00325**, 2015.
- [5] Y. Liao, J. Xie, and A. Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. **TPAMI**, 2021.
- [6] Ali Furkan Biten, Lluís Gómez, Marçal Rusiñol, and Dimosthenis Karatzas. Good news, everyone! context driven entity-aware captioning for news images. In **CVPR**, pp. 12466–12475, 2019.
- [7] Zihang Meng, Licheng Yu, Ning Zhang, Tamara L. Berg, Babak Damavandi, Vikas Singh, and Amy Bearman. Connecting what to say with where to look by modeling human attention traces. In **CVPR**, pp. 12679–12688, 2021.
- [8] Shizhe Chen, Qin Jin, Peng Wang, and Qi Wu. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In **CVPR**, pp. 9962–9971, 2020.
- [9] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In **ICCV**, 2015.
- [10] Shih-Han Chou, Wei-Lun Chao, Wei-Sheng Lai, Min Sun, and Ming-Hsuan Yang. Visual question answering on 360° images. In **WACV**, pp. 1596–1605, 2020.
- [11] Heeseung Yun, Youngjae Yu, Wonsuk Yang, Kangil Lee, and Gunhee Kim. Pano-avqa: Grounded audio-visual question answering on 360° videos. In **ICCV**, 2021.
- [12] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. **IJCV**, Vol. 123, No. 1, pp. 32–73, 2017.
- [13] Jianxiong Xiao, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Recognizing scene viewpoint using panoramic place representation. In **CVPR**, 2012.
- [14] Yu-Chuan Su, Dinesh Jayaraman, and Kristen Grauman. Pano2vid: Automatic cinematography for watching 360 videos. In **ACCV**, 2016.
- [15] Erik Wijmans and Yasutaka Furukawa. Exploiting 2d floorplan for building-scale panorama rgb-d alignment. In **CVPR**, pp. 1427–1435, 2017.
- [16] Hao Yang and Hui Zhang. Efficient 3d room shape recovery from a single panorama. In **CVPR**, 2016.
- [17] Steve Cruz, Will Hutchcroft, Yuguang Li, Naji Khosravan, Ivaylo Boyadzhiev, and Bing Kang. Zillow indoor dataset: Annotated floor plans with 360° panoramas and 3d room layouts. In **CVPR**, pp. 2133–2143, 2021.
- [18] Shih-Han Chou, Cheng Sun, Wen-Yen Chang, Wan Ting Hsu, Min Sun, and Jianlong Fu. 360-indoor: Towards learning real-world objects in 360° indoor equirectangular images. **WACV**, pp. 834–842, 2019.
- [19] Ahmed Rida Sekkat, Yohan Dupuis, Pascal Vasseur, and Paul Honeine. The omniscene dataset. In **ICRA**, pp. 1603–1608. IEEE, 2020.
- [20] Senthil Yogamani, Ciaran Hughes, Jonathan Horgan, Ganesh Sistu, Padraig Varley, Derek O’Dea, Michal Uricar, Stefan Milz, Martin Simon, Karl Amende, Christian Witt, Hazem Rashed, Sumanth Chennupati, Sanjaya Nayak, Saquib Mansoor, Xavier Perrotton, and Patrick Perez. Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving. In **ICCV**, 2019.
- [21] Stephan Fremerey, Ashutosh Singla, Kay Meseberg, and Alexander Raake. Avtrack360: An open dataset and software recording people’s head rotations watching 360° videos on an hmd. In **MMSys. ACM**, p. 403–408, 2018.
- [22] Jianxiong Xiao, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Recognizing scene viewpoint using panoramic place representation. In **CVPR**, pp. 2695–2702, 2012.
- [23] Shih-Han Chou, Wei-Lun Chao, Wei-Sheng Lai, Min Sun, and Ming-Hsuan Yang. Visual question answering on 360° images. In **WACV**, pp. 1596–1605, 2020.
- [24] Volkan Cirik, Taylor Berg-Kirkpatrick, and Louis-Philippe Morency. Refer360°: A referring expression recognition dataset in 360° images. In **ACL**, pp. 7189–7202, 2020.
- [25] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In **ICML**, pp. 23318–23340, 2022.
- [26] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **ACL**, pp. 311–318, 2002.
- [27] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Text Summarization Branches Out**, pp. 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics.
- [28] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In **WMT**, pp. 376–380, 2014.
- [29] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In **CVPR**, pp. 4566–4575, 2015.
- [30] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: semantic propositional image caption evaluation. In **ECCV**, pp. 382–398, 2016.

A MTurk を用いた説明文の収集



By dragging the photo, you can view various directions of 360 degree photo.

Theme
Required
Do not select the same themes

Theme1
Description theme (required)

Theme2
Description theme (required)

Theme3
Description theme (required)

Themed Description
Required
Describe the scene precisely as everyone can imagine this image

>= 10 words

>= 10 words

>= 10 words

(Optional) When you consider this image is inappropriate for themed description
Fill in the descriptions even if you check here (not just writing placeholders). Otherwise, your HIT may be rejected!
☐ This 360 image is not appropriate for themed description (e.g., not a natural photo, not context, ...)

図2 MTurk の作業画面のスクリーンショット。画像をドラッグすることでそれぞれの方向の歪みの無い画像を見ることができる。

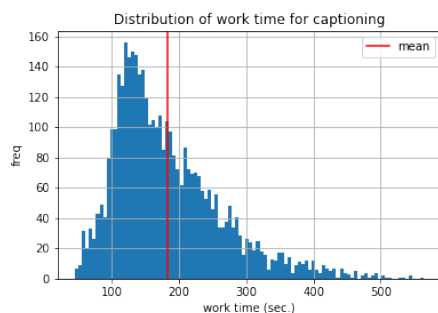


図3 作業者が説明文の付与にかかった時間の分布。

MTurk の作業者の雇用は、品質の担保のため、US 在住者のうち直近の課題承認率が 99%を上回る評価者に限定して雇用した。図3 は作業者が説明文を付与するためにかかった実際の作業時間の分布を示す。3つの説明文を付与するのににかかった作業時間が45秒未満であるものは不適切とみなして自動的に除去した。そのうえで計測した作業時間の平均は182秒であり、時給が10\$以上となるように報酬を支払った。