社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

大規模日本語情景文字データセット OPU-JST-1

佐藤　瞳†　松田　崇宏†　池田　佑輝†　岩村　雅一†　黄瀬　浩一†

† 大阪府立大学 大学院工学研究科　〒 599–8531 大阪府堺市中区学園町 1–1
E-mail: †{hitomi,matsuda,ikeda}@m.cs.osakafu-u.ac.jp,{masa,kise}@cs.osakafu-u.ac.jp

あらまし　スマートフォンやタブレットなどの小型端末の普及が進んでいることから，それらを用いた様々な情景中の文字を認識するサービスやアプリケーションの需要が高まっている．情景中の文字は多種多様な形状を持ち，撮影時に射影歪みや照明環境の影響を受ける．解像度が低下する場合もある．そのため，これらの文字の検出・認識は，スキャナで撮像した場合など，撮像環境をコントロールした場合よりも困難である．この問題に対処する有効な方法と考えられているのが，種々の劣化を受けた文字データを大量に集めて，識別器の学習に使用することである．本研究では全方位カメラを用いて商店街などを撮影し，得られた画像からなる情景中の日本語文字データセット，OPU-JST-1 を構築したので，それを報告する．このデータセットは，時系列情報を持つ情景画像 31,410 枚と，790,257 個の文字領域のラベル，2,764,230 文字を収録している．

キーワード　パターン認識　大規模データベース　情景中文字　日本語　動画

1. はじめに

近年，小型カメラが搭載されたスマートフォンやタブレットなどの小型端末の普及が進んでいる．それに伴い，そのような端末を用いた様々な情景中の文字を認識するサービスやアプリケーションの需要が高まっている．例えば，カメラで撮影した文字を認識し，翻訳して出力する，うつして翻訳(注1)や，撮影した看板の文字を認識し，関連するウェブサイトへのアドレスを表示する Google Goggles(注2)などがある．情景中の文字は多種多様な形状を持ち，撮影時に射影歪みや照明環境の影響を受ける．解像度が低下する場合もある．そのため，これらの文字の検出・認識は，スキャナで撮像した場合など，撮像環境をコントロールした場合よりも困難である．この問題に対処する有効な方法と考えられているのが，種々の劣化を受けた文字データを大量に集めて，識別器の学習に使用することである(注3)．

本研究では，全方位カメラを用いて商店街などを撮影し，得られた画像からなる情景中の日本語文字データセット OPU-JST-1 を構築した．このデータセットは，図 1 のようにカメラで撮影した画像と，その画像中に含まれる文字領域のラベルと位置を示したファイルで構成されている．文字領域を切り出さず，元の画像を提供するため，文字検出タスクのクエリとして用いることも可能である．このデータセットは，1,200 × 1,600[pixel] の情景画像 31,410 枚からなり，790,257 個の文字領域のラベルと，2,764,230 文字を収録している．

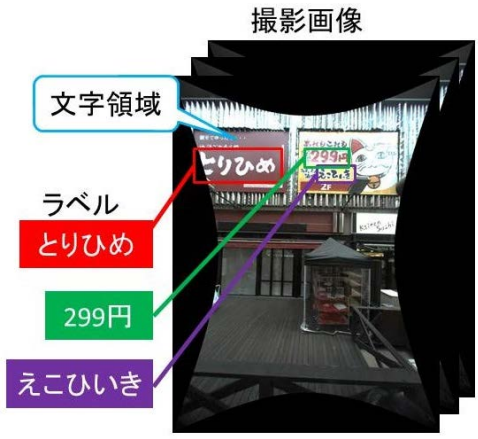


図 1　OPU-JST-1 の構成

比較のため，現在公開されている主要な文字のデータセットを表 1 に示す．OPU-JST-1 の特徴として，大阪の商店街に存在する日本語，アルファベットなどの多様な文字画像をありのままで収録してること，動画データとして画像を取得しているため，時間的に連続した画像で構成されていることが挙げられる．収録文字数は，情景文字を扱ったデータセットの中でも最大である．大規模なデータベースに対する需要は高いものの，正解ラベル付けを人手で行う必要があるため，大規模なデータセットの構築は容易でない．そのため，このデータベースが情景中文字認識研究の発展に大いに役立つと考えている．

2 節では，現在公開されている文字データセットについて述べる．3 節以降では，データセット構築手順を順に説明する．まず，(1) 全方位カメラで撮影した動画データをフレームごと

(注1)：https://www.nttdocomo.co.jp/service/information/utsushite_honyaku/
(注2)：https://play.google.com/store/apps/details?id=com.google.android.apps.unveil
(注3)：実際、ICDAR2013 Robust Reading Competition [1] の単語認識タスクで最も性能が良かった Google の PhotoOCR [2] はユーザが撮影した実データを使用して学習している．

表 1　データセットの比較

名前	画像 [枚]	ラベル数	元画像	言語	対象	動画
SVHN [3]	248,823	630,420	あり	数字	情景	×
NEOCR [4]	659	5,238	あり	英独	情景	×
Ahmed [5]	-	1,000,000	なし	英	文書	×
ICDAR2003 [6]	499	2,263	あり	英	情景	×
ICDAR2011(情景) [7]	483	2,524	あり	英	情景	×
ICDAR2013(動画) [1]	15,277	93,598	あり	英仏西	情景	○
SVT [8]	349	904	あり	英	情景	×
OPU-JST-1	31,410	790,257	あり	日	情景	○

に分割し，(2) 文字領域のラベル付けをする．そして，(3) プライバシー保護のために顔領域の特定を行い，得られた顔領域にぼかし処理をする．

2.　既存のデータセット

本節では，表 1 に示した主要な文字データセットについて述べる．Google Street View House Numbers (SVHN) Dataset [3] は，Google Street View から収集した約 25 万枚の情景画像と，63 万個の文字領域のデータを収録しており，扱っているのは数字のみである．Natural Environment OCR Dataset (NEOCR) [4] は，複数人で撮影した情景画像を集めた 659 枚の情景画像に，5,238 個の文字領域のデータを収録しており，ラベルは英語とドイツ語が主となっている．Ahmed らのデータセット [5] は，カメラで撮影した文書画像を認識するためのものである．印刷した英語の文書をさまざまな角度からカメラで撮影し，画像中の文字・単語領域を自動的にラベル付けする．100 万個の文字画像を収録している．ICDAR 2003 Robust Reading Competition [6] のために用意されたデータセットは，性能評価によく用いられる．このデータセットは，複数人で撮影した商品パッケージやポスター，看板などの 499 枚の情景画像と，2,263 個のラベルを収録している．ICDAR 2011 Robust Reading Competition [7] の Challenge 2，情景文字検出・認識用のデータセットは，看板やロゴなどの 483 枚の情景画像中に，2,524 の文字領域を収録している．ICDAR 2013 Robust Reading Competition [1] の Challenge 3，動画中の文字検出・認識のためのデータセットは，28 動画 (計 15,277 フレーム) と，93,598 個の文字領域を収録している．ただし，この文字領域数は文字であるが判読できない Don't care ラベルのものも含んでいる．The Street View Text (SVT) Dataset [8] は，SVHN Detaset と同様に Google Street View の画像を収集したものであり，349 枚の情景画像中に，904 個の文字領域を収録している．

これらのデータセットは数字やラテン系の文字を対象としたのものであり，日本語を対象としたものはない．また，ICDAR 2013 Robust Reading Competition [1] の動画タスクのデータセットを除き，時系列情報を持っていない．最近は動画撮影可能なスマートフォンやウェラブルデバイスが普及しており，動画の利用が容易になっている．動画中の文字検出・認識は，動画の各フレームに対して独立に行われるのが一般的である．そのため，同じ文字列を何度も処理する無駄が生じたり，あるフレームにだけ隠れが発生した場合には，前後のフレームから類推できるにもかかわらず，本来あるべき文字列の情報が得られないといった不都合が生じる．時系列情報を含むデータセットの公開により，この状況を改善する新たな技術開発を促す可能性がある．

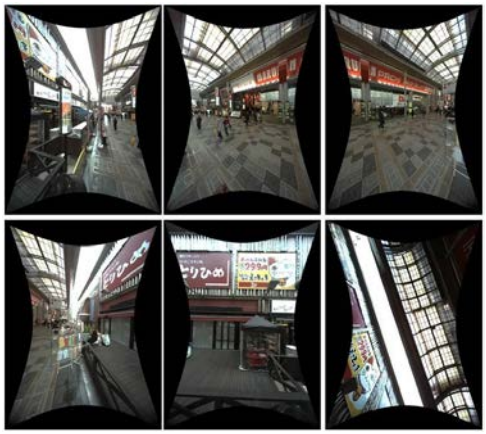


図 2　全方位カメラで得られる画像例

図 3　全方位カメラを用いた撮影セット

3.　全方位カメラによる画像撮影

本研究では，画像撮影に Point Grey Research の全方位カメラ Ladybug 3 を用いた．このカメラは，6 個のカメラからなり，5 個が横方向を，1 個が上方向を向いている．図 2 の画像は，それぞれのカメラで同時に撮影された画像である．なお，カメラのフレームレートは 6.5fps である．

このカメラを用いて，図 3 のような撮影セットを作成した．台車と三脚を用いて，カメラを固定したまま移動・撮影することができる．台車には車用のバッテリーを積んでおり，カメラに電源を供給している．このセットを用いて，難波，心斎橋，あびこなどの商店街を歩きながら動画撮影した．その詳細を表 2 に示す．撮影時間は計 5.6 時間で，全てのカメラから得られた動画をフレームごとに分割して約 78 万枚の画像を得た．

表 2　撮影場所・時間・枚数

撮影場所	撮影時間 [h]	画像枚数
堺東	0.73	101,874
あびこ	0.50	70,614
天王寺	0.38	53,754
難波	3.71	521,988
アメリカ村	0.25	35,100
合計	5.57	783,150

4.　文字のラベル付け

学生アルバイトを雇い，延べ約 2,600 時間かけて画像中のテキストに手動でラベル付けを行った．以下でその詳細を述べる．

4.1　ラベル付けソフト

ラベル付け作業を効率化するために，図 4 に示すラベル付けソフトを作成した．ラベル入力者は，写真上で 4 点を順に指定することにより，四角形の文字領域を決定し，その中に含まれる文字のラベルを入力する．その際，文字の言語，ID，反転しているかどうかを入力することが出来る．このソフトの最大の特徴は，あるフレームの画像に付与したラベルを時間的に連続した別のフレームに伝搬させることができる点である．これにより，ラベル入力者は新しいフレームのラベルを一から入力せずとも，既に入力したラベルの位置を微調整することのみで新たなフレームのラベルを付与することが出来る．

ここで，ラベルの伝搬システムの詳細について述べる．このラベル伝播システムは，文字領域を図 5 のように次のフレームに射影し，そのラベルを伝搬させることができる．このシステムは，RANSAC [9] による射影変換行列の算出と，テンプレートマッチングによる誤差修正によって実現する．まず，ラベルをつけた n フレーム目の画像と，伝播先となる $n+1$ フレーム目の画像から局所特徴量を抽出し，それらの比較を行う．局所特徴量とは，画像の局所的な領域から抽出される特徴量であり，このシステムでは SIFT [10] を用いている．これを用いることで，n フレーム目と $n+1$ フレーム目の画像の特徴点の対応を得ることが出来るので，RANSAC によって n フレーム目から $n+1$ フレーム目への変換行列を算出する．フレーム間の変換を射影変換と仮定しているので，対応する特徴点が最低 4 組あれば，射影変換行列を求めることが出来る．最後に，RANSAC の誤差を修正するため，局所的にテンプレートマッチングを行い，相関が最も高くなる位置を射影位置とする．このとき，RANSAC で求めた射影変換行列を用いて $n+1$ フレーム目に射影した四角形上を，元の n フレーム目の四角形で走査する．

4.2　ラベル付け規則

本節では，ラベル付けの規則について述べる．文字領域は，図 6(a) のようにスペースで区切る．日本語の場合，原則として文節単位で区切る．例えば「あれもこれも売ります」という文であれば,「あれも」「これも」「売ります」と区切る．アルファベットの場合は単語単位で区切る．また，原則として記号を無視し,「06-001」のように記号を含んだ文字列の場合,「06」「001」と記述する．特殊な場合として，図 6(b) のように，領域の一部の文字が見切れや隠れなどで判別出来ない場合は,「　e」のように，その文字部分にスペースを入力する．また，図 6(c) のように文字が見えない場合や，図 6(d) のように判読できず文字を特定することが出来ない場合は,「#Don't care#」のラベルを付与する．これは，文字検出タスクにおいて，検出が非常に難しい領域は検出できなくてもペナルティを与えないが，検出できなくても誤検出とみなさないことへの対処である．図 6(e) のようなアルファベットの単語中のアクセントなどは，日本語で表現できないため,「Veterin*a*ria」のように，アクセントのついた文字をアスタリスクで囲んで記述する．

図 4　ラベル付けソフト

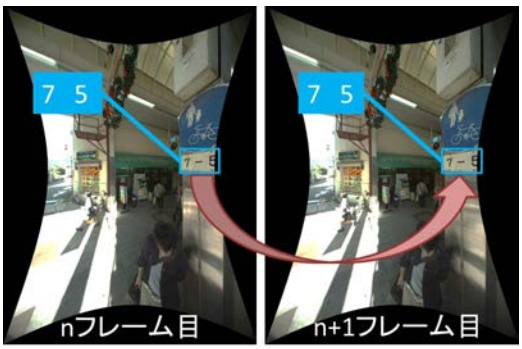


図 5　ラベルの射影システム

(a) 領域の区切り方

(b) 文字に抜けのある場合

(c) Don't care 領域 (1)

(d) Don't care 領域 (2)

(e) アクセントの処理

図 6　ラベル付け規則

4.3　文字ラベルの集計結果

ラベル付け作業の結果，31,410 枚の画像中に，913,916 個の

文字領域が得られた．そのうち 123,659 個が Don't care 領域であり，残る 790,257 個がそれ以外の文字領域であった．その中には，2,764,230 文字が含まれていた．図 7 の画像は，ラベル付けされた領域とそのラベルの一例である．作業時間はのべ約 2600 時間であり，単純に 24 時間/日で換算すると 3 か月半に相当する．時間的に連続している画像群の数は 32，連続フレームの最大は 5,001 フレーム，最小は 8 フレーム，平均は 964 フレームであった．

ラベル付けした文字の集計結果を図 8， 9 に示す．図 8 は，字種ごとの集計結果を示している．図 8 より，今回撮影した場所では漢字，カタカナ，アルファベットが多かったことが分かる．このような実世界に存在する文字領域の統計データを得ることにより，文字の出現頻度を事前確率とした識別器の生成も可能となると考えられる．

以下に，図 9 で示す集計結果の詳細について述べる．漢字は 50 位まで，数字はすべての結果を，それ以外は上位 30 位までの結果を表示したものである．漢字は，計 811 種の文字が得られた．常用漢字は約 2,000 字あるので，大半は現れていないことが分かった．図 9(a) より，得られた文字の中で一番多かったのは「円」であった．これは，撮影場所が商店街であることから，店先の値札の文字が多く得られたためだと考えられる．2 位から 5 位では，日本橋，大阪など，主に撮影場所の地名に含まれる文字が多く含まれていた．これも，看板などの店名が由来であると考えれる．ひらがなでは，図 9(b) のように「の」が最も多く得られていた．これは,「○○の××」など，助詞としての使用が多いためと考えられる．カタカナでは，図 9(c) のように「ン」と「ー」が突出して多く得られた．これは,「パソコン」や「セール」などの，頻出する単語に含まれていたためと考えられる．アルファベットでは，図 9(e) のように「A」「O」が多く含まれていた.「SALE」「NAMBA」「OPEN」などの店舗に関わる文字が多く見られた．その他，数字では図 9(d) のように「0」が多く含まれていた．これは，大阪の市外局番「06」から始まる店舗の電話番号や，値札の表記が多く含まれていたためと考えられる．

以上より，商店街で撮影したため，それにまつわる文字が多く得られていることが分かった．また，ひらがなと数字はすべての字種が現れていたが，カタカナでは「ゾ」，アルファベットでは「j」，そして大半の漢字など，全く出現しない文字もあった．残りの画像についてもラベル付けを行い検証をする必要はあるが，文字の出現頻度は非常に偏っていることがわかった．より多くの文字種を得るために，様々な場所での撮影が必要となると考えられる．

5. プライバシーの保護

全方位カメラから得られた画像は，通行人の多い昼間の商店街で撮影しているため，多くの顔領域を含んでいる．データセットとして公開するためには，画像中に含まれる顔領域を特定し，プライバシーの保護のため，個人が特定されないようにぼかしなどの処理を加える必要がある．

図 7　ラベル付けされた文字とラベルの例．見やすさのため，ラベル中のスペースは"_"で表記している．

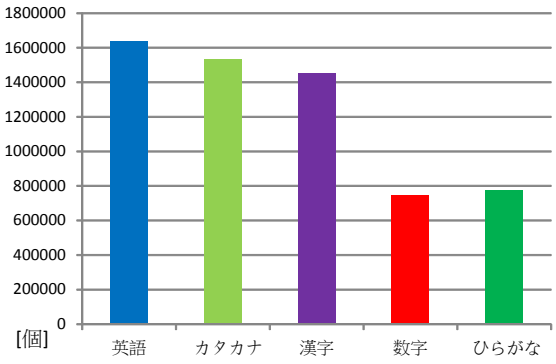


図 8　字種別の集計結果

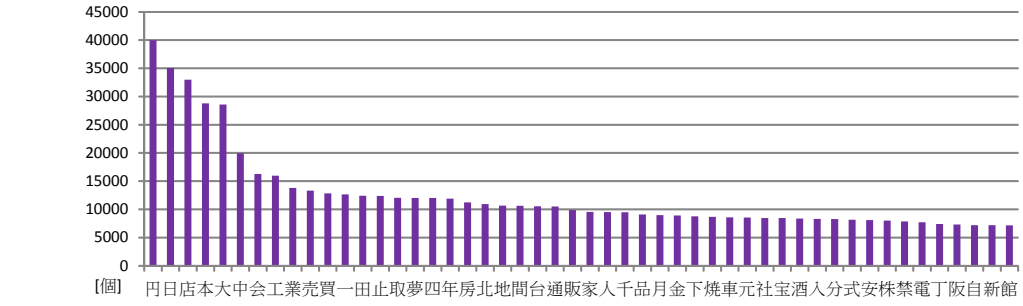
5.1 領域特定の自動化の試み

文字のラベル付けが終了している画像は，31,410 枚である．これらの顔領域の特定を全て手作業で行うのは，時間的，コスト的にも避けたい．そこで，既存の顔検出手法による特定精度を検証するため，正解データとして使用する画像の顔領域の特定を手作業で行う．なお，検証のため，顔領域には向きと解像度についてのラベルを付与する．3,700 枚の画像を調べた結果，2,460 枚の画像に 4,267 個の処理の必要な顔領域が存在した．顔領域を特定した結果を表 3 に，特定した顔領域の一例を図 10 に示す．このように，様々な顔向きや照明条件があることが分かった．そのため，そのような条件に頑健な手法を用いる必要がある．

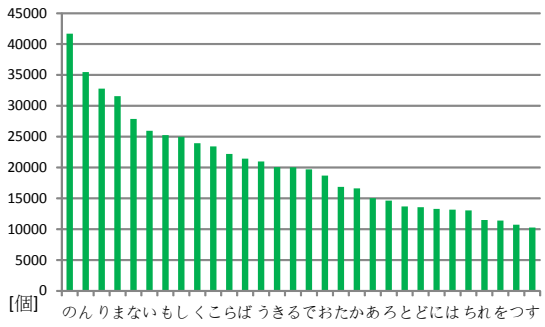
顔領域の自動検出手法として，そのような条件に頑健な，Viola-Jones の顔検出手法 [11]，Hog 特徴による人物検出，Zhu らの face alignment 手法 [12] を用いて実験を行った．なお，パラメータはデフォルトまま用いた．その結果を，表 4 に示す．再現率は，検出したい領域のうち，どれだけの領域を正しく検出できたか，適合率は，検出できた領域のうち，どれだけが正しい顔領域であったかを示している．実験の結果，どの手法でも再現率，適合率ともに非常に低くなることがわかった．そのため，既存の手法による顔領域の自動検出は非常に困難であるといえる．

5.2 Amazon mechanical turk を利用した顔領域の特定

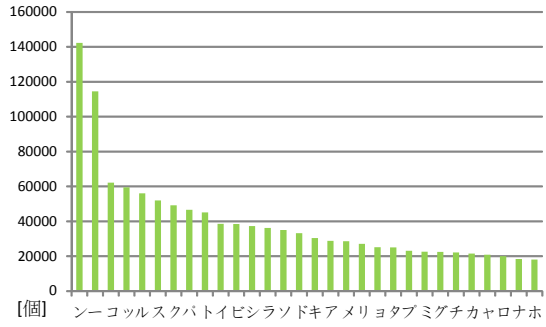
前節における実験の結果から，残りの画像についても人の手



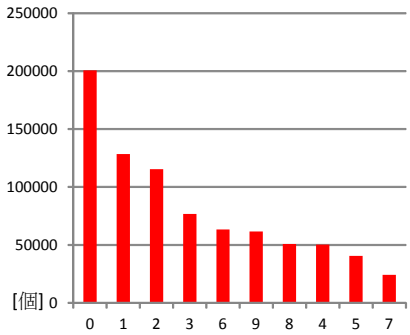
(a) 漢字



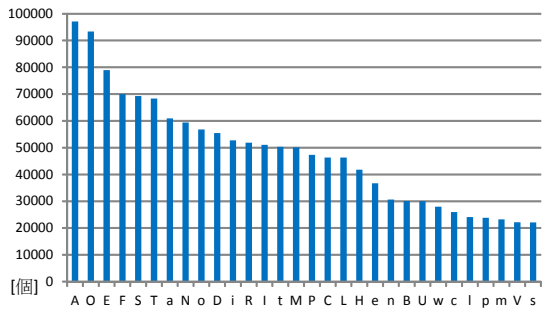
(b) ひらがな



(c) カタカナ



(d) 数字



(e) アルファベット

図 9　文字ラベルの集計結果

表 3　手動で特定した顔領域の分類

合計	解像度:高	解像度:低	正面	右向き	左向き
4,267	2,146	2,121	2,584	472	1,211

図 10　抽出された顔領域の例

で顔領域の特定作業を行うことにした．3,700 枚の領域特定の作業にかかった時間は，のべ 15 時間であった．文字ラベルが付与されている残りの画像は約 29,000 枚あるため，残り全ての画像にも人力で特定を行う場合，のべ約 120 時間が必要となる．そこで，コスト削減のため，Amazon Mechanical Turk(注4)にてラベル付け作業の作業者を募集することとした．作業者は英語圏やインドの人々が主であるため，文字領域のラベル付けは困難であるが，顔領域の特定であれば言語に関係なく委託することが可能である．ただ，全ての作業者が正しい仕事をするわけではない．精度を確保するために，同一のタスクを複数人に割り当てる．

表 4　顔領域の特定実験の結果．Viola-Jones の手法では，パラメータを変化させ，再現率，適合率のどちらかが最も高くなった結果を示している．

手法	再現率 [%]	適合率 [%]
Viola-Jones の手法 (1)	0.45	10.67
Viola-Jones の手法 (2)	8.20	0.20
Zhu らの手法	1.06	34.09
Hog 特徴による手法	4.75	1.80

今回は，残りの画像から保護の必要な顔領域を含む画像を抽

(注4)：https://www.mturk.com/

表 5　タスクにかかった費用

	タスク数	単価	割当人数	合計
抽出タスク	973	$0.04	3	$131.355
特定タスク	18,713	$0.01	2	$561.39

図 11　顔領域のぼかし処理

出するタスク (抽出タスク) と，得られた画像の顔領域の特定をする (特定タスク) の，2 つのタスクを募集した．これは，事前の特定作業から顔領域を持たない画像があることがわかっており，それらすべてに特定タスクを行うよりも，事前に抽出タスクを行うことで，コストを削減可能であるからである．各タスクの見積もりを，表 5 に示す．合計は，作業者への支払額と，Amazon Mechanical Turk への支払額の合計となり，本研究での Amazon Mechanical Turk への支払額は，タスク数×割り当て数×$0.005 で求めることが出来る．

5.2.1　抽出タスク

抽出タスクは，30 枚の画像から顔を含んでいる画像にチェックを入れる作業である．抽出タスクのタスク数は，画像枚数 29,187 枚を 30 で割った 2,919(余りは繰り上げ) となる．抽出タスクを募集した結果，約 2/3 にあたる 18,713 枚の画像に顔領域が含まれていた．このタスクは，264 人の作業のもと約 12 時間で終了した．

5.2.2　特定タスク

特定タスクは，LabelMe(注5)のラベル付けツールを使用し，1 枚の画像中のすべての顔を特定する作業である．作業者は，顔領域の矩形の描画と，顔向き (正面，左右) と解像度 (高低) の記述を行う．抽出タスクの結果，18,713 枚に顔が含まれていたため，それらの画像について特定タスクを行った．特定タスクは，全 37,426 タスクが，1,492 人の作業のもと，約 20 日で終了した．なお，リジェクトは全体の約 14% の約 5,000 件であった．ほとんどは，顔向きと解像度の記述が所定の形式に則っていなかったためである．

5.3　顔領域のぼかし処理

ラベル付けソフトを用いたものと，Amazon Mechanical Turk に委託した画像の両方に，特定した顔領域に対して，図 11 のようにぼかし処理を行った．

6. ま と め

文字認識・検出においては，大規模なデータセットを用いることで精度が向上することがわかっている．しかし，大規模データセットは数が少なく，それらも字種や撮影条件などが偏っており，実世界のデータを反映しているとは言い難い．そこで，我々は大規模日本語情景文字データセットを公開することを目的とし，全方位カメラで撮影した画像によるデータセット OPU-JST-1 を構築した．このデータセットは，撮影した画像と，その画像中に含まれる文字領域のラベルと位置を示したファイルで構成されている．このデータセットには，撮影した画像約 78 万枚のうち，人の手によるラベル付けが終了した 31,410 枚が含まれている．これらの画像には日本語の他にアルファベットも含まれており，790,257 個の文字領域のラベルが含まれている．更に，動画として取得したため時系列の情報も保持していることから，動画中の文字検出・認識への応用も可能であると考えられる．

謝辞　本研究の一部は JST CREST の補助を受けた．ここに記して感謝する．

文　献

[1] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. i Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan and L. P. de las Heras: “ICDAR 2013 robust reading competition”, Proc. ICDAR, pp. 1115–1124 (2013).

[2] A. Bissacco, M. Cummins, Y. Netzer and H. Neven: “Photoocr: Reading text in uncontrolled conditions”, Proc. ICCV, pp. 785–792 (2013).

[3] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu and A. Y. Ng: “Reading digits in natural images with unsupervised feature learning”, NIPS Workshop on Deep Learning and Unsupervised Feature Learning (2011).

[4] R. Nagy, A. Dicker and K. Meyer-Wegener: “NEOCR: A configurable dataset for natural image text recognition”, CBDAR, Lecture Notes in Computer Science, Springer, pp. 150–163 (2011).

[5] S. Ahmed, K. Kise, M. Iwamura, M. Liwicki and A. Dengel: “Automatic ground truth generation of camera captured documents using document image retrieval”, ICDAR 2013, pp. 528–532 (2013).

[6] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, R. Young, K. Ashida, H. Nagai, M. Okamoto, H. Yamamoto, H. Miyao, J. Zhu, W. Ou, C. Wolf, J.-M. Jolion, L. Todoran, M. Worring and X. Lin: “ICDAR 2003 robust reading competitions: Entries, results and future directions”, IJDAR, **7**, 2-3, pp. 105–122 (2005).

[7] A. Shahab, F. Shafait and A. Dengel: “ICDAR 2011 robust reading competition challenge 2: Reading text in scene images”, Proc. ICDAR, pp. 1491–1496 (2011).

[8] K. Wang, B. Babenko and S. Belongie: “End-to-end scene text recognition”, Proc. ICCV2011, pp. 1457–1464 (2011).

[9] M. A. Fischler and R. C. Bolles: “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography”, Communication of the ACM, **24**, 6, pp. 381–395 (1981).

[10] D. G. Lowe: “Distinctive image features from scale-invariant keypoints”, IJCV, **60**, 2, pp. 91–110 (2004).

[11] P. Viola and M. J. Jones: “Robust real-time face detection”, IJCV, **57**, 2, pp. 137–154 (2004).

[12] X. Zhu and D. Ramanan: “Face detection, pose estimation, and landmark localization in the wild”, Proc. CVPR, pp. 2879–2886 (2012).

(注5)：http://labelme.csail.mit.edu/Release3.0/