画像認識　特定物体認識　Open API

Technology Reports

高速大規模画像認識エンジンの開発とAPIの提供

写真のあらゆる被写体（モノ）を認識可能な世界を目指し，写ったモノが何か，500万枚以上の画像が登録されている大規模なデータベースから瞬時に一致する画像を探索し，モノを特定する画像認識システムを開発した．リアルタイムな画像認識を実現することで，文字入力・音声入力に続く新しい入力インタフェースを実現する．本稿ではドコモが開発した高速かつ認識対象が大規模な画像認識エンジンと，本画像認識エンジンをベースとして2014年10月にサービス提供を開始した画像認識APIのサービス概要について解説する．

サービスイノベーション部　赤塚 隼（あかつか はやと）　猪俣 哲平（いのまた てっぺい）　酒井 俊樹（さかい としき）

1. まえがき

画像認識とは，画像データの内容を分析することで画像に何が写っているかを識別する技術全般を指す．

画像認識の歴史において技術開発でまず先行したのは文字認識技術であり，商工業分野における業務効率化に役立てられた．日本では郵便番号制度が発足した1968年にあわせて，株式会社東芝が開発した世界初の手書き文字を認識する郵便物自動処理装置[1]が実用化され，これまでの手作業による郵便番号別の区分が機械化された．近年ではコンピューティングパワーの増大，カメラの小型化・低価格化，画像認識アルゴリズムの発展により，一般消費者も画像認識技術の恩恵を体感できるようになってきている．たとえば交通事故を起こさない車社会の実現に向けてトヨタ自動車株式会社が提供するナイトビューシステム[2]では，画像処理を用いてリアルタイムに歩行者を検出し，ドライバーに歩行者の存在を伝えることで，夜間運転時の安全性を改善した．ゲーム業界では2010年にMicrosoftがXbox360®*1向けに，物理的なコントローラを用いずジェスチャーによって直観的かつ自然にゲームプレイが可能なゲームシステムのKinect™*2 [3]を開発した．Eコマース（電子商取引分野）では，2014年にAmazon.comが物体認識を応用し映像からリアルタイムに商品を識別し，同社のオンラインショッピングサイトへユーザを誘導するAmazon Firefly*3 [4]を開発した．前記はごく一部の事例ではあるが，これらのように画像認識は登場から半世紀で我々の生活に浸透している．そして今後，スマートフォンや眼鏡型ウェアラブルデバイスなど携帯無線通信環境下のモバイル機器においても，さまざまなモノを映像や写真を通して瞬時に認識するニーズはより一層高まると考えられる．

これに向け，ドコモでは独自の画像認識技術の開発・強化に努めている．すでに文字認識においては，カメラをかざすだけで日本語に海外の言語を翻訳する，うつして翻訳™*4を提供しているが，文字以外のさらに複雑な物体の認識が可能な画像認識エンジンの開発も行っている．複雑な対象をとらえる画像認識では，事前に認識させたいモノの画像を登録しておく必要があり，市販の商品をリアルタイムに認識させるには数百万オーダーの大規模なモノの画像

＊1 Xbox360®：米国Microsoft Corp. および／またはその関連会社の商標．
＊2 Kinect™：米国Microsoft Corp. の米国およびその他の国における登録商標または商標．
＊3 Amazon Firefly：アマゾンの米国およびその他の国における商標．
＊4 うつして翻訳™：NTTドコモの商標．

のデータベースから瞬時に高精度で，類似度の高い画像を識別することが課題であった．これは認識対象物の数が増えることで，類似画像の照合時間がかかることと，似たような画像特徴を持ったアイテムが増えるため精度が低下するからである．ドコモでは画像認識アルゴリズムを改善することで上記課題を解決し，数百万オーダーの大規模なデータベースから1秒以内に類似画像を高い精度で識別する技術を開発した．

本稿では，ドコモが開発した，写真の被写体（モノ）を特定する画像認識技術（特定物体認識技術）について解説し，現状の画像認識の精度と処理速度について紹介する．また，開発者への支援を行いオープンイノベーションを創出することを目的としたdocomo Developer support [5]で，2014年10月にサービス提供を開始した画像認識API（Application Programming Interface）のサービス概要について解説する．

2. 画像認識概要

2.1 画像認識アルゴリズム

ドコモが開発した画像認識エンジンの画像認識アルゴリズム（以下，本アルゴリズムと呼ぶ）における認識処理は，平面のモノの認識が主である．画像に写っているモノが何であるか（例えば書籍が写っていれば，何というタイトルの本か）を特定する．処理は大まかに3段階のフェーズに分かれる（**図1**）．

①キーポイントの検出

ユーザから入力された画像（質問画像）から，モノの特徴を現す点（キーポイント）をリアルタイムに検出する．データベースにあらかじめ登録されているモノの画像（参照画像）についても同様にキーポイントをオフラインで事前に検出してお

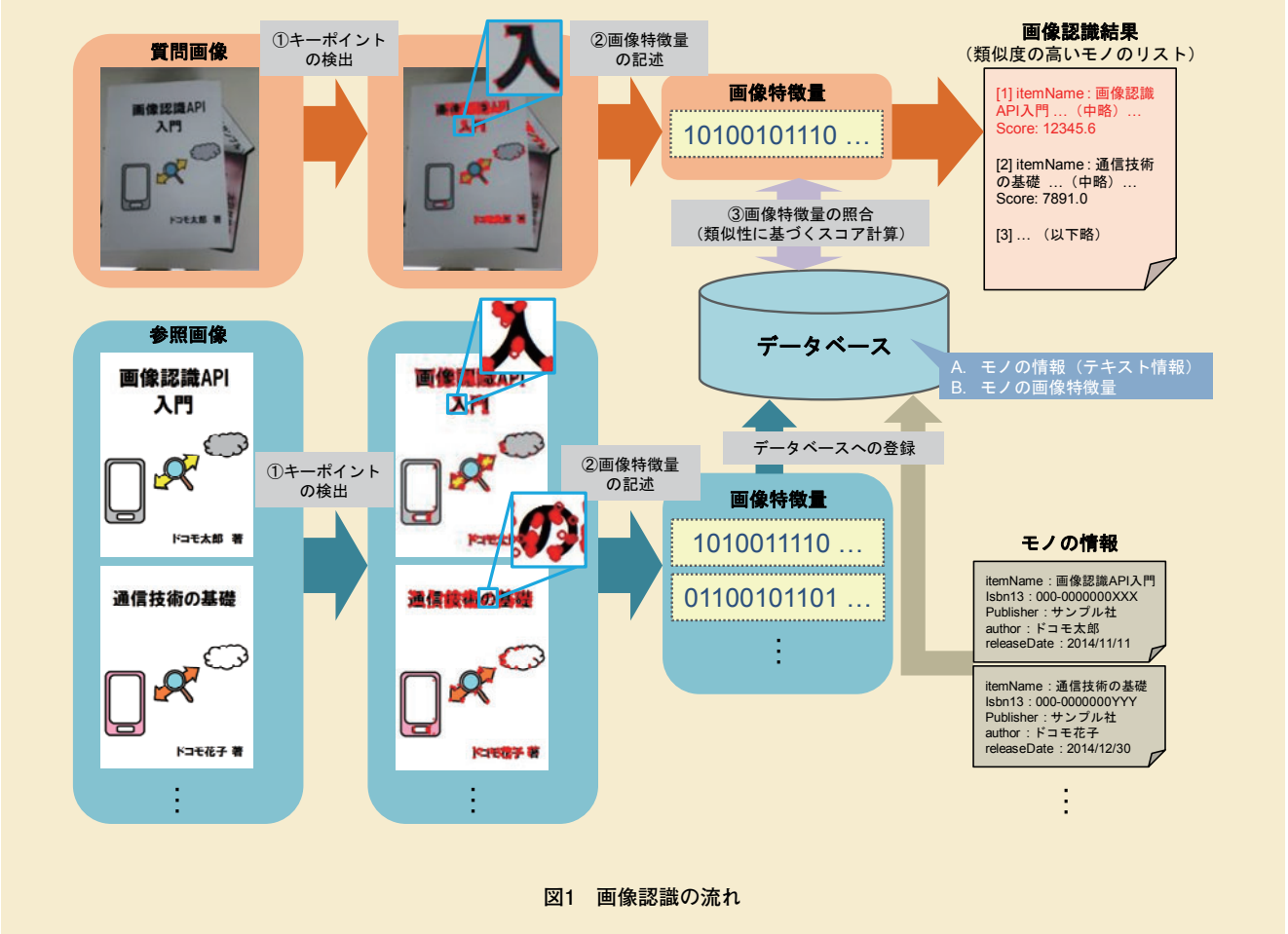

図1　画像認識の流れ

く．

②画像特徴量[*5]の記述

①で検出された質問画像・参照画像の各キーポイントについて，周辺の輝度分布などの情報からそのキーポイントを特徴づける量（画像特徴量）を数値化して記述する．質問画像についてはリアルタイムに処理を行い，参照画像についてはオフラインで事前処理を行う．

③画像特徴量の照合

質問画像と参照画像の画像特徴量を比較し，質問画像と最も類似度の高い参照画像を選択する．

以下，各フェーズについて詳しく解説する．

⑴キーポイントの検出

画像認識では，画像に写るモノを特定するために，そのモノが持つ固有の情報を画像データから抽出する必要がある．このフェーズでは撮影状況が異なっていても，そのモノが持つ固有の情報として，安定して抽出可能なキーポイントを検出する．

検出されるキーポイントは，モノが傾いていたりスケールが変動していたりと，さまざまな写り方をしていても安定して同じ点を検出可能であることが望ましい．つまり，モノの向きや傾き，スケールなどの変動に頑健であることが要求される．一般にこのような性質を持つのは線と線が交差するコーナー点であるとされ，本アルゴリズムにおいては複数のコーナー点検出手法を組み合わせることによって安定したキーポイント検出を実現している．

質問画像からはリアルタイムにキーポイント検出を行い，参照画像については事前にオフラインでキーポイントを検出する．質問画像と参照画像で同じキーポイントが検出されることが望ましいが，撮影状況による多少のずれは想定される．

⑵画像特徴量の記述

ここでは，前フェーズにおいて検出されたキーポイントの画像特徴量をキーポイントの周辺輝度に基づいて算出する．質問画像と参照画像に写っているモノが同一である場合，質問画像のキーポイントの多くが参照画像のキーポイントと対応関係を持つことが期待される．各キーポイントに固有の画像特徴量を記述することによって，特徴量の比較により質問画像と参照画像のキーポイントの対応関係を知ることができる．

本アルゴリズムにおける画像特徴量とは，具体的には局所画像特徴と呼ばれる，キーポイントとそのキーポイント周辺の領域（近傍領域）でどのように輝度が分布しているかを記述したベクトルである．本アルゴリズムでは，同一のキーポイントならば撮影時のスケールの変動やモノの面内の回転（ロール）による画像の変化に関係なく一定の値を維持する性質（不変性）を持つように画像特徴量を定義している．これにより，後述する画像特徴量の比較において，モノの向きやスケールが変わっても安定的に特徴点のマッチングが可能となる．

スケール不変性や回転不変性を持つ画像特徴量としてSIFT（Scale-Invariant Feature Transform）[6]やSURF（Speeded Up Robust Features）[7]などがすでに存在しているが，本アルゴリズムではSIFTやSURFと同じくスケール不変・回転不変性を持ちながら，より高速に抽出・マッチング可能なバイナリ[*6]ベクトルで記述された特徴量を使用している．

⑶画像特徴量の照合

このフェーズでは，質問画像と参照画像の特徴量の類似度を比較し，データベースより類似度の高いモノを特定する．本アルゴリズムの特徴は，この特定が高速に行えることである．

データベースには参照画像に写っているモノの商品IDなどの情報と，参照画像で検出されたキーポイントについて算出された画像特徴量の集合が格納されている．画像そのものではなく画像特徴量を比較することで，質問画像に対して最も類似している参照画像を見つけ，その質問画像に写っているモノが何であるかを認識する．

具体的には，質問画像の各キーポイントに対し，データベース内の全キーポイントの特徴量との類似度を計算し，類似するキーポイントのマッチングを行う．質問画像と参照画像に写っているのが同一のモノであれば十分な数の対応する特徴点ペアが見つかるので，質問画像と当該参照画像との間でモノの姿勢がどう変化しているか推定を行い，推定さ

＊5 **特徴量**：データから抽出される，そのデータを特徴づける量（数値）のこと．本稿における特徴量とは特に画像特徴量とも呼ばれ，画像から検出された特徴点（コーナー点）においてその周辺の輝度分布を特徴づける量である．

＊6 **バイナリ**：0または1の数字列で記述される2進数での数値表現形式．

れた姿勢変化に沿わないペアを除去することで最終的なキーポイントペアを決定する．この対応するキーポイントペアの数とその類似度に基づいて，各参照画像の類似スコアを算出することでモノを特定する．

ところが参照画像が大量にある場合，総当たりで類似度を計算していたのでは，計算時間に数十秒，あるいはそれ以上を要してしまい，ユーザ体験の質的低下を招く．

そこで本アルゴリズムでは局所性鋭敏型ハッシュ*7（LSH：Locality Sensitive Hashing）を応用した高速な探索手法を開発した．LSHを用いることで特徴量をより低次元なハッシュ空間にまとめることができるため，類似しているデータが効率的に探索可能となる．確率的な探索手法であるため，理論的には必ず最適解が求まる保証はないが，実際にはほとんどの場合最適解が見つかる．また，数百万オーダーの参照画像に対しても1秒以内に照合が完了する．

2.2　認識性能

本アルゴリズムの性能を確認するために評価実験を実施した．ここでは，本アルゴリズムの性能と評価実験の結果について，認識精度と処理速度の両面から説明する．

⑴認識精度

約100万件の参照画像に対し，モノを正面から大きくはっきりと写した画像，（正面の画像に対して）拡大および縮小を加えた画像，以下同様にロール・パン・チルトを加えた画像，ノイズ・ぼかしを加えた全8種の質問画像について精度評価実験を行った（**図2**）．

ここでは，類似スコアが高い順に上位3位までの認識結果に正解が含まれる割合を認識成功率とする（**図3**）．正面の画像やノイズ・ぼかしなど，画像の「見え」が大きく変化しない画像の場合，90％以上の高精度を達成できている．一方でパン・チルトなどのモノの見えが変化する画像は，正面から写したものやロールなどと異なり特徴量の不変性が無いため，相対的に認識成功率が低くなった．また拡大縮小については不変性を持っているものの，実際には拡大の場合は図2の例のように被写体が画面からはみだしてしまう，縮小の場合には解像度が低くなり情報量が不足してしまうため，対応するキーポイントが少なくなり，認識成功率が下がる結果となった．

ここまでの評価実験の結果，認識

図2　評価用の質問画像8種

*7 **ハッシュ**：もとのデータからある一定範囲の数値を生成する処理．主にデータの高速な検索や照合などに使われる．

成功率を低下させる2つの要因が判明した．1つめはスケール変化や傾きなどによる「見え」の変化である．モノの見え方が変化することによってキーポイントの検出やその特徴量が不安定になり，マッチングが正しく行われないことが増えるため，認識成功率が低下する．2つめは似たような特徴による参照画像の取り違えである．新書など商品の見た目としてタイトル名と著者名以外は違いがないモノなどの認識は，そうでない場合に比べて認識成功率が落ちる．現在，ドコモではさらなる認識成功率改善へ向け，以上2つの問題の解決に取り組んでいる．

⑵処理速度

総当たり的なマッチング手法であれば，参照画像数の件数が増大すればマッチングに必要な処理時間もそれに比例して増大する．しかし本アルゴリズムのLSHによる高速マッチング手法ではハッシュ関数を用いた低次元なベクトル空間への写像により特徴量を次元圧縮するため，その増大幅は総当たり的マッチングに比べてずっと緩やかである．本評価実験の結果，10万件の参照画像について質問画像1枚を認識するために必要な処理時間は平均0.24秒であったのに対し，100万件の参照画像については平均0.64秒であった．10倍の件数増加に比して処理時間の増加は約2.7倍に留まっている．現在，眼鏡型ウェアラブルデバイスなどの普及も視野に入れ，リアルタイム映像からのシームレスな認識を実現するためさらなる高速化に取り組んでいる．

3. 画像認識API サービス概要

本画像認識APIは，前述の画像認識機能をdocomo Developer support [8]においてアプリ／サービスの開発者向けに提供しているREST（REpresentational State Transfer）API*8であり，docomo Developer supportへの会員登録および利用申請を行うことで誰でも利用することができる．

3.1 画像認識APIの特性

画像認識APIでは，日本国内で販売されている商品を，パッケージの画像を基に認識する機能を提供している．対応する商品のカテゴリは書籍，DVD，CD，PCソフト，ゲームソフト，食品である．ドコモが所有するデータベース上には500万件以上の市販商品の画像および商品情報が蓄積されており，入力された質問画像の特徴量とデータベース上の画像特徴量を照合し，質問画像と類似する参照画像に紐づいている商品情報を返却する．

これまでの画像認識サービスでは，

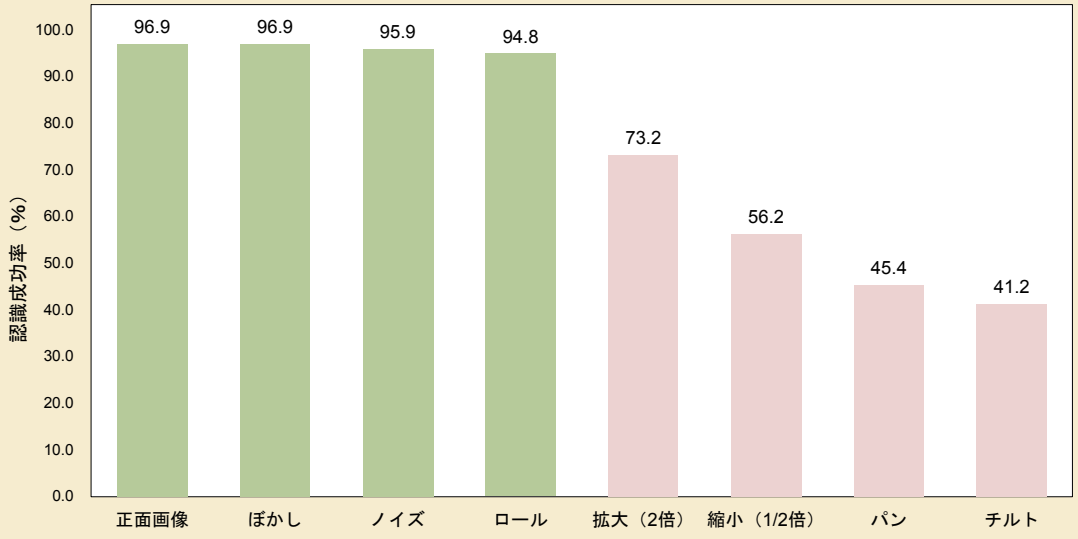

図3　認識精度（上位3位までの認識結果に正解が含まれる割合）

*8 **REST API**：RESTの制約に従ったAPI．RESTはRoy Fielding氏が2000年に提唱した設計原則を基に発展した，ソフトウェアアーキテクチャのスタイル．

例えば日本電気株式会社が提供するGAZIRU®*9 [9]，PUX株式会社が提供するオブジェクト認識ソフトウェア[10]のように，認識機能のみを提供し，データベース構築に必要な画像と，その画像に写っている物体の情報は，画像認識サービスのユーザが独自に集め，登録する方式が一般的であった．本画像認識APIは，画像認識機能に加えてドコモが独自に収集した500万件以上のデータベースを合わせてユーザに提供する方式をとる．APIを利用するユーザがデータを収集する労力を低減するので，より手軽に画像認識機能を利用したマッシュアップ*10アプリ／サービスを開発・構築できるという利点がある．加えて，後述するが，本画像認識APIは画像を入力するだけで画像認識が可能な形で設計されており，開発者は内部の処理に関与しないので，画像認識のメカニズムに関して知識がなくても，簡単に画像認識アプリ／サービスの開発が可能となっている．

3.2　利用方法

図4に，画像認識APIの画像の入力と認識結果の返却の流れのイメージを示す．画像認識APIを利用するユーザは，HTTP（Hyper Text Transfer Protocol）のPOSTメソッド*11を用い，リクエストボディ*12に質問画像を添付して送信することで，認識結果を受け取ることができる．

結果は，JSON（JavaScript Object Notation）*13形式のテキストデータで返却され，質問画像に写っている商品の名称のほか，認識の確からしさを示すスコア（参照画像との類似度），商品の詳細情報が返却される．商品の詳細情報としては，例えば書籍であれば出版社や出版年月日，著者などの情報のほか，その商品が購入できるECサイトのリンクなどが返却される．

また本件のAPIはフィードバック用のエンドポイントも提供しており，認識結果の妥当性をユーザがフィードバックすることが可能である．フィードバックされた内容は，認識精度の改善およびデータベースの更新のために利用される．

3.3　サービス例

画像認識APIを利用するユーザは，画像認識APIから返却される情報と自身のアイデアを組み合わせることで，画像認識を応用した新しいサービスを開発することが可能である．例えば，商品の名称およびECサイトのリンクを基に商品のレビューや価格を取得して表示するアプリケーションや，Amazon Firefly[4]のように写真に撮った商品をすぐにECサイトで購入できるようなアプリケーションなどが考えられる．この他にも商品を撮影して商品管理に利用するなど，さまざまな応用が考えられるであろう．

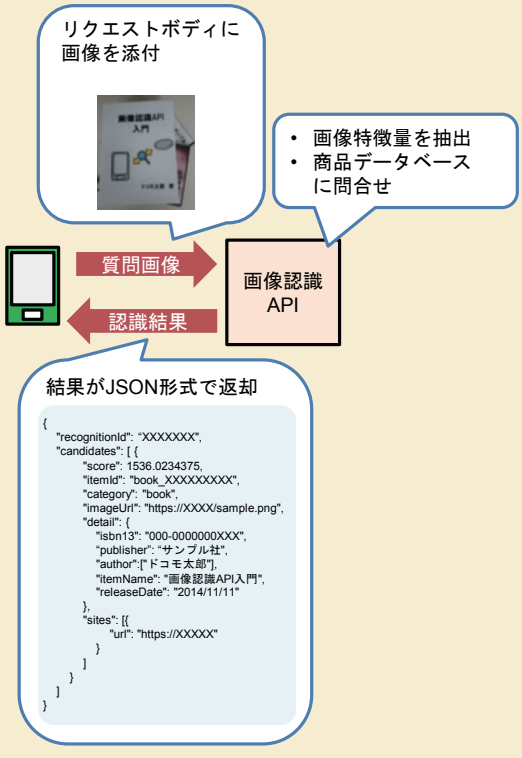

図4　画像認識APIをサービス（アプリなど）に組み込んだ際の処理の流れ

*9 **GAZIRU®**：日本電気株式会社の登録商標．
*10 **マッシュアップ**：複数の異なるサービスやコンテンツを組み合わせて，1つのサービスを作成，提供すること．
*11 **POSTメソッド**：HTTPを用いた通信において，クライアントからサーバにデータを送信するためのメソッド．
*12 **リクエストボディ**：POSTメソッドにおいて，クライアントから送信するデータが記述される部分．
*13 **JSON**：JavaScriptのオブジェクト記述に基づくデータ記述言語．

また画像認識APIは，拡張現実（AR：Augmented Reality）技術との親和性が高く，商品を認識し，認識結果を基にした情報を画像や動画に重ねて表示するなどの利用方法が可能である．特にGoogle Glassのような眼鏡型ウェアラブルデバイスとは相性が良く，付属のカメラが撮影した映像に対して画像認識を行い，情報を眼鏡上のスクリーンに表示することで，ユーザがよりシームレスにモノの情報を得ることができるようになるだろう．本画像認識APIの公開以降，実際に一般のユーザによってAR技術・ウェアラブルデバイスを用いたアプリ開発が活発に行われている．

4. あとがき

本稿では，画像認識を用いて写真の被写体を識別する技術と，公開した画像認識APIのサービス概要について解説した．

画像認識の精度については，モノの写り方によって認識精度が異なるが，正面から撮影されている場合，高精度な認識が可能であり，また処理時間に関しては100万規模の参照画像数の場合であっても高速に処理が行えることを実験により証明した．

現在ドコモが開発した画像認識アルゴリズムは平面のモノの認識が主であるが，今後は，3次元の角度によって見え方が異なる立体的なモノ（ランドマーク*14，有名人，ファッション，料理など）に対する高速大規模画像認識技術の実現に向けて，さらに取組みを進めていく．

文　献

[1] 東芝：“東芝未来科学館：世界初の郵便物自動処理装置．”
http://toshiba-mirai-kagakukan.jp/learn/history/ichigoki/1967postmatter/index_j.htm

[2] トヨタ自動車：“トヨタ｜安全技術｜ナイトビュー．”
http://www.toyota.co.jp/jpn/tech/safety/technology/technology_file/active/night_view.html

[3] Microsoft Research: “Human Pose Estimation for Kinect - Microsoft Research.”
http://research.microsoft.com/en-us/projects/vrkinect/default.aspx

[4] Amazon. com: “Understanding Firefly - Amazon Apps & Services Developer Portal.”
https://developer.amazon.com/public/solutions/devices/fire-phone/docs/understanding-firefly

[5] NTTドコモ：“docomo Developer supportにおける開発者支援の取組み，”本誌，Vol.22，No.2，pp.47-50，Jul. 2014.

[6] D. G. Lowe: “Distinctive Image Features from Scale-Invariant Keypoints,” International Journal of Computer Vision, Vol.60, No.2, pp.91-110, 2004.

[7] H. Bay, T. Tuytelaars, and L. V. Gool: “SURF: Speeded Up Robust Features,” 9th European Conference on Computer Vision, 2006.

[8] NTTドコモ：“画像認識｜docomo Developer support｜NTTドコモ．”
https://dev.smt.docomo.ne.jp/?p=docs.api.page&api_docs_id=102

[9] NEC：“サービス実現イメージ｜画像認識サービスGAZIRU（ガジル）｜NEC．”
http://jpn.nec.com/solution/cloud/gazou/service.html

[10] PUX Corp.: “PUX Developers site - technology.html.”
https://pds.polestars.jp/contents/technology.html

*14 **ランドマーク**：その土地の名所や象徴となるような建造物．