(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7661610号
(P7661610)

(45)発行日　令和7年4月14日(2025.4.14)

(24)登録日　令和7年4月4日(2025.4.4)

(51)国際特許分類

Ｇ０６Ｆ　40/151 (2020.01)		Ｇ０６Ｆ	40/151	
Ｇ０６Ｆ　21/62　(2013.01)		Ｇ０６Ｆ	21/62	３４５

請求項の数　15　（全22頁）

(21)出願番号　特願2024-504506(P2024-504506)
(86)(22)出願日　令和4年7月29日(2022.7.29)
(65)公表番号　特表2024-530889(P2024-530889A)
(43)公表日　令和6年8月27日(2024.8.27)
(86)国際出願番号　PCT/EP2022/071384
(87)国際公開番号　WO2023/012069
(87)国際公開日　令和5年2月9日(2023.2.9)
審査請求日　令和6年11月15日(2024.11.15)
(31)優先権主張番号　21189837.4
(32)優先日　令和3年8月5日(2021.8.5)
(33)優先権主張国・地域又は機関
欧州特許庁(EP)
早期審査対象出願

(73)特許権者　516005083
ブルー プリズム リミテッド
イギリス国 ダブリュエー２ ０エックス ピー ウォリントン，ファーンヘッド，クラブ レーン，シナモン パーク ２
(74)代理人　110002572
弁理士法人平木国際特許事務所
(72)発明者　ジャン，デ
イギリス国 シーオー２ ７エフエイチ コルチェスター，ケンジントン ロード ７８
(72)発明者　ダッバ，クリシュナ サンディープ レディ
イギリス国 シーエム６ １ジービー リトル キャンフィールド，ワーウィック ロード １８

最終頁に続く

(54)【発明の名称】　データ難読化

(57)【特許請求の範囲】

【請求項１】

リモートアクセスアプリケーションを介して受信された機密データがオペレータに対して出力されることを防止するコンピュータ実装方法であって、

リモートアクセスアプリケーションを介してリモートサーバから、前記リモートサーバ上で実行されているソフトウェアアプリケーションのグラフィカルユーザインタフェースを受信するステップ；

前記グラフィカルユーザインタフェースを変更して機密データを削除するステップであって、

プロセッサによって、非構造化画像データを取得するステップ；

前記プロセッサによって、前記非構造化画像データから構造化データを抽出するステップであって、前記構造化データは、機密データであり、定義された機能的フォーマットおよび定義された視覚的フォーマットを有する、ステップ；

前記プロセッサによって、前記構造化データとは異なる人工データを生成するステップであって、前記人工データは、前記構造化データと同じ機能的フォーマットを有する、ステップ；

前記プロセッサによって、前記構造化データが前記人工データに置換された前記非構造化画像データに基づいて人工非構造化画像データを生成するステップであって、前記人工データは、前記構造化データの前記視覚的フォーマットに基づいている、ステップ；

前記プロセッサによって、前記人工非構造化画像データを出力するステップ；

によって、前記グラフィカルユーザインタフェースを変更して機密データを削除するステップ；
前記変更されたグラフィカルユーザインタフェースを、オペレータが受信するための、コンピュータの１つまたは複数の出力周辺機器に対して出力するステップ；
を有する方法。
【請求項２】
請求項１に記載の方法において、前記人工非構造化画像データは、前記構造化データと同じ視覚的フォーマットを有する方法。
【請求項３】
請求項１または請求項２に記載の方法において、前記非構造化画像データから構造化データを抽出するステップは、
前記非構造化画像データに対して光学式文字認識を実行し、前記非構造化画像データ内のテキストを識別するステップ；
前記テキスト内の構造化データを識別するステップ；
前記構造化データに対応する前記非構造化画像データ内の１つまたは複数のバウンディングボックスを決定するステップ；
前記１つまたは複数のバウンディングボックスを使用して、前記非構造化画像データから１つまたは複数の画像部分を抽出するステップ；
を有する方法。
【請求項４】
請求項３に記載の方法において、前記人工非構造化画像データを生成するステップは、
前記１つまたは複数の画像部分における前記構造化データの前記視覚的フォーマットを識別するステップ；
前記１つまたは複数の画像部分における前記構造化データの前記視覚的フォーマットに基づいて、前記１つまたは複数の画像部分に対応する１つまたは複数の人工画像部分を生成するステップ；
前記非構造化画像データを変更して、前記１つまたは複数の画像部分を前記１つまたは複数の人工画像部分によって置換するステップ；
を有する方法。
【請求項５】
請求項１記載の方法において、前記定義された機能的フォーマットは、
エンティティタイプ、および
エンティティタイプフォーマットのうちの１つまたは複数を含む方法。
【請求項６】
請求項１記載の方法において、前記定義された視覚的フォーマットは、
テキスト長、
テキストフォント、
テキスト色、および
背景色のうちの１つまたは複数を含む方法。
【請求項７】
請求項１記載の方法において、前記人工データを生成するステップは、
前記構造化データの機能的フォーマットを識別するステップ；
前記非構造化画像データから抽出された前記構造化データを一覧化するステップ；
前記構造化データの前記機能的フォーマットに基づいて人工データを生成するステップ；
前記人工データを、対応する前記構造化データと共に一覧化するステップ；
を有する方法。
【請求項８】
請求項７に記載の方法において、更に、前記一覧化された構造化データおよび前記一覧化された人工データを編集可能な表として出力することを含む方法。
【請求項９】

請求項８に記載の方法において、更に前記人工非構造化画像データを生成する前に、前記人工データを確認するためのプロンプトをオペレータに送信することを含む方法。
【請求項１０】
請求項７記載の方法において、更に、第２の非構造化画像データを取得するステップ；前記第２の非構造化画像データから第２の構造化データを抽出するステップ；前記第２の構造化データとは異なる第２の人工データを生成するステップを含み、前記第２の構造化データが前記構造化データと同一である場合、前記第２の人工データは前記人工データと同一である方法。
【請求項１１】
請求項１記載の方法において、前記人工非構造化画像データは、前記機密データを使用するソフトウェアの開発プロセス中にディスプレイに出力される方法。
【請求項１２】
請求項１１に記載の方法において、前記ディスプレイは、前記プロセッサとは異なるコンピュータに属する方法。
【請求項１３】
請求項１１記載の方法において、前記ソフトウェアは、ロボットプロセス自動化（ＲＰＡ）プロセスである方法。
【請求項１４】
プロセッサによって実行されると、前記プロセッサに請求項１記載の方法を実行させる命令を含む、コンピュータストレージデバイス上に格納されたコンピュータプログラム。
【請求項１５】
請求項１記載の方法を実行するように構成されたプロセッサを含むコンピューティングシステム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、機密データを不明瞭化するためのコンピュータ実装方法、コンピュータプログラム及びコンピュータシステムに関する。
【背景技術】
【０００２】
データの再編集（redaction）は、データの機密性が高い内容（sensitive content）、すなわち認知的内容（cognitive content）を露出させないように、機密データを視覚的に削除又は置換する処理である。例えば、図１Ａに示す画像には、日付及び／又は時刻、ドメイン名、電子メールアドレス、氏名、電話番号等、様々な種類の機密データが含まれている。このデータの機密内容は、特定の日付及び／又は時刻（Mon 22/02/2021 10:28; 1.5 hrs）、特定のドメイン名（www.blueprism.com）、特定の電子メールアドレス（Ben.Carter@blueprism.com）等である。データ再編集とは、このデータを視覚的に除去又は置換する処理である。
【０００３】
データ再編集を行う理由は、再編集されたデータが共有されるとき、再編集されたデータを受け取った当事者が機密性が高い内容を収集できないようにするためである。機密データの侵害は、機密データの偶発的又は違法な破壊、紛失、改ざん、機密データの不正な開示、又は機密データへの不正なアクセスを引き起こし、重大な人的被害をもたらす可能性があるため、この処理は重要である。更に、受信側当事者による機密データへのアクセスは、様々な法域の規制により違法となる可能性もある。
【０００４】
データの再編集には幾つかの手法が知られている。これらの手法は、一般に、例えば、不透明な（通常は、黒色の）矩形、固定文字列（パスワードの「******」やクレジットカード番号の「xxxx xxxx xxxx 6789」等）、エンティティタイプに依存する文字列（人名の「〈PERSON〉」等）、又はピクセル化によって、検出された機密データを覆ったり

置換したりすることによって、機密データを再編集することを含む。例えば、Ｇｏｏｇｌｅ Ｃｌｏｕｄでは、Ｃｌｏｕｄ Ｄａｔａ Ｌｏｓｓ Ｐｒｅｖｅｎｔｉｏｎは、ｂａｓｅ６４エンコードされた画像のテキストを検査し、テキスト内の機密データを検出し、照合された機密データが不透明な矩形によって不明瞭化されたｂａｓｅ６４エンコードされた画像を返す。図１Ｂは、図１Ａの画像にこの先行技術を適用したものである。

【発明の概要】

【発明が解決しようとする課題】

【０００５】

既存のデータ再編集技術の問題点は、このような技術では、再編集された機密データの視覚的フォーマットのみでなく、機能的フォーマットを理解することも困難になるということである。これは、機密データがロボットプロセスオートメーションアプリケーション（robotic process automation application）等のソフトウェアアプリケーションで使用され、ソフトウェアアプリケーションの開発及びテストが必要な場合に特に問題となる。この理由は、ソフトウェア開発者は、通常、ソフトウェアアプリケーションが正しくセットアップされ及び機能することを保証するために、機密データの機能的フォーマット、場合によっては、視覚的フォーマットを理解する必要があるからである。

【課題を解決するための手段】

【０００６】

本発明は、独立請求項によって定義され、更なる任意の特徴は、従属請求項によって定義される。

【０００７】

本発明の第１の側面では、機密データを不明瞭化するためのコンピュータ実装方法が提供され、この方法は、プロセッサによって、画像データを取得することと、プロセッサによって、画像データから構造化データを抽出することであって、構造化データは、機密データであり、定義された機能的フォーマット及び定義された視覚的フォーマットを有することと、プロセッサによって、構造化データとは異なる人工データを生成することであって、人工データは、構造化データと同じ機能的フォーマットを有することと、プロセッサによって、構造化データが人工データに置換された画像データに基づいて人工画像データを生成することであって、人工データは、構造化データの視覚的フォーマットに基づいていることと、プロセッサによって、人工画像データを出力することとを含む。これにより、データの機密性の高い認知的内容（sensitive cognitive content）を明らかにすることなく、機密データの機能的フォーマットを維持しながら、機密データを不明瞭化できる。これは、（例えば、機能的フォーマットがロボットプロセス自動化（ＲＰＡ）アプリケーションのような別のソフトウェアアプリケーションとインタラクションするため）機密データの機能的フォーマットを理解する必要があるが、その内容の機密性のために機密データにアクセスできない人間であるオペレータ（human operator：以下、単に「オペレータ」という）にとって有用である。更に、人工データの視覚的フォーマットが機密データの視覚的フォーマットに基づいていることは、人工画像データが画像データの代わりに使用できることを保証するために有用である。

【０００８】

幾つかの実施形態では、人工画像データは、構造化データと同じ視覚的フォーマットを有する。これは、オペレータによって開発される必要がある機密データとインタラクションするソフトウェアアプリケーションが、例えば、ＲＰＡアプリケーションにおいて、機密データの視覚的フォーマットに何らかの形で依存する場合に有用である。

【０００９】

ある実施形態では、画像データから構造化データを抽出することは、画像データに対して光学式文字認識を実行し、画像データ内のテキストを識別することと、テキスト内の構造化データを識別することと、構造化データに対応する画像データ内の１つ又は複数のバウンディングボックスを決定することと、１つ又は複数のバウンディングボックスを使用して、画像データから１つ又は複数の画像部分を抽出することとを含む。これにより、機

密データを置換する目的で、機密データを含む画像データの部分を識別し、残りの画像データを無視でき、後続のステップの処理要件を軽減できる。

【００１０】

オプションとして、定義された機能的フォーマットは、エンティティタイプ、及びエンティティタイプフォーマットのうちの１つ又は複数を含む。エンティティタイプは、例えば、氏名、日付及び／又は時刻、電子メールアドレス、住所等、機密データが何を示しているかについての情報を提供する。これにより、使用可能なエンティティタイプフォーマットの種類を絞り込むことができる。エンティティタイプフォーマットは、所与のエンティティタイプのフォーマットである。エンティティタイプフォーマットを判定することにより、機密データと同じエンティティタイプフォーマットを使用して人工データを生成できる。例えば、機密データが英国の郵便番号の場合、人工データも英国の郵便番号フォーマットになる。これにより、他のソフトウェアアプリケーションが機密データの機能的フォーマットをどのように扱うか、また、これがそのソフトウェアアプリケーションの実行にどのように影響するかを理解しやすくなる。

【００１１】

オプションとして、定義された視覚的フォーマットは、テキスト長、テキストフォント、テキスト色、及び背景色のうちの１つ又は複数を含む。テキスト長を使用することにより、生成される人工画像データが、同じ長さ又は類似の長さを使用して、人工データが機密データを囲むバウンディングボックス内に収まるようにできる。テキストフォントを使用することによって、フォントの視覚的フォーマットに関する潜在的な問題を認識できる。例えば、あるソフトウェアアプリケーションが画像上で光学式文字認識を使用しているが、使用されている特定のフォントではＩ（大文字の「i」）とｌ（小文字の「L」）の見た目がよく似ている場合、人工画像データ内のテキストフォントを維持することで、そのような問題が識別可能になる。テキスト色と背景色により、人工画像データは、機密データと同じ視覚的フォーマットを持つことができる。

【００１２】

ある実施形態では、人工データを生成することは、構造化データの機能的フォーマットを識別することと、画像から抽出された構造化データを一覧化することと、構造化データの機能的フォーマットに基づいて人工データを生成することと、人工データを、対応する構造化データと共に一覧化することとを含む。これらの実施形態において、本方法は、オプションとして、一覧化された構造化データ及び一覧化された人工データを編集可能な表として出力することを含んでもよい。これにより、オペレータが人工データを検証し、承認できる。更に重要な点として、オペレータは、編集可能な表を変更して、特定された機密データに修正を加えることができる。具体的には、ユーザは、表に行を追加することによって見逃された機密データを追加し、表の行を削除することによって誤検出された機密データエンティティを削除し、表の対応するセルを編集することによって機密データの誤りを修正し、又は生成された人工データを単に置換することができる。これらの実施形態において、本方法は、更にオプションとして、人工画像データを生成する前に、人工データを確認するためのプロンプトをオペレータに送信することを含んでもよい。これにより、オペレータは、機密データから人工データへのマッピングがどのように実行されるかを透過的に見ることができるため、アプローチの堅牢性が向上する。

【００１３】

更なる実施形態において、本方法は、第２の画像データを取得することと、第２の画像データから第２の構造化データを抽出することと、第２の構造化データとは異なる第２の人工データを生成することとを含んでもよく、第２の構造化データが先の構造化データと同一である場合、第２の人工データは先の人工データと同一である。これは、同じ機密データの複数回の出現が、本方法を用いて常に同様に難読化されることを意味する。このように人工データを異なる画像データ間で一貫させることにより、人工データを使用するオペレータに混乱を生じさせることが回避される。更に、データの完全性が維持されるため、機密データを使用するソフトウェアアプリケーションのテスト／デバッグが可能になる。

【００１４】
特定の実施形態では、人工画像データを生成することは、１つ又は複数の画像部分における構造化データの視覚的フォーマットを識別することと、１つ又は複数の画像部分における構造化データの視覚的フォーマットに基づいて、１つ又は複数の画像部分に対応する１つ又は複数の人工画像部分を生成することと、画像データを変更して、１つ又は複数の画像部分を１つ又は複数の人工画像部分に置換することとを含んでもよい。このように画像部分を考慮することにより、機密データを不明瞭化するために異なる画像部分における異なる視覚的フォーマットを考慮に入れることができる。
【００１５】
幾つかの実施形態では、人工画像データは、機密データを使用するソフトウェアの開発プロセス中にディスプレイに出力される。データの機能的フォーマットと、データの視覚的フォーマットは、いずれも、データのこれらの側面を何らかの方法で使用する様々なソフトウェアアプリケーションにとって重要であり得る。機能的フォーマットと視覚的フォーマットを維持することにより、ソフトウェア開発者は、そのソフトウェアアプリケーションが正しく設定され、機能していることを確認できる。オプションとして、ディスプレイは、プロセッサとは別のコンピュータに属する。通常、ソフトウェア開発を行う当事者は、機密データを見ることを許可されない。これにより、一方の当事者（すなわち、機密データを見ることを許可された当事者、クライアントコンピュータ）がそのデータの不明瞭化と人工画像データの生成を実行し、他方の当事者（開発者コンピュータ）がその人工画像データを使用することが保証される。
【００１６】
特定の実施形態では、ソフトウェアは、ロボットプロセス自動化（robotic process automation：ＲＰＡ）プロセスである。ＲＰＡは、自動化されたプロセス及びワークフローを使用して、グラフィカルユーザインタフェース（ＧＵＩ）又はドキュメント等の画像データから構造化データ、典型的には、機密データを抽出することを含む。自動化されたプロセスやワークフローが正しく機能しているかどうかを判断できるようにするためには、人工データ内の機密データの機能的フォーマットを維持することが重要である。ＲＰＡは、機密データを抽出するためにソフトウェアアプリケーションのＧＵＩに依存することが多いため、機密データの視覚的フォーマットを維持することも重要である。すなわち、機密データの視覚的フォーマットを維持することで、抽出に伴う潜在的な問題を明らかにできる。
【００１７】
本発明の第２の側面では、リモートアクセスアプリケーションを介して受信された機密データがオペレータに出力されることを防止するコンピュータ実装方法が提供される。この方法は、リモートアクセスアプリケーションを介してサーバから、リモートサーバ上で実行されるソフトウェアアプリケーションのＧＵＩを受信することと、本発明の第１の側面の方法に基づいて、ＧＵＩを変更して機密データを除去することと、変更されたＧＵＩを、オペレータが受信するための、コンピュータの１つ又は複数の出力周辺機器に出力することとを含む。これにより、本発明の第１の側面の利点をリモートアクセスアプリケーションのコンテキストで使用でき、これは、特にソフトウェア開発目的に有用である。
【００１８】
本発明の第３の側面では、プロセッサによって実行されると、プロセッサに本発明の第１の側面の方法を実行させる命令を含むコンピュータプログラムが提供される。
【００１９】
本発明の第４の側面では、プロセッサによって実行されると、プロセッサに本発明の第１の側面の方法を実行させる命令を含むコンピュータ可読媒体が提供される。
【００２０】
本発明の第５の側面では、本発明の第１の側面の方法を実行するように構成されたプロセッサが提供される。
【００２１】

本発明の第６の側面では、本発明の第１の側面の方法を実行するように構成されたプロセッサを含むコンピューティングシステムが提供される。

【００２２】

以下の図面を参照して、本発明の実施形態を例示的に説明する。

【図面の簡単な説明】

【００２３】

【図１Ａ】本発明の方法で使用する画像データの例を示す図である。

【図１Ｂ】（先行技術）公知のデータ再編集技術を適用した後の図１Ａの画像データを示す図である。

【図２】本発明の方法を実施するためのシステムの例を示す図である。

【図３】本発明の方法を実施するためのシステムの例を示す図である。

【図４】本発明の方法に基づいて不明瞭化される機密データを含む画像データの例を示す図である。

【図５】本発明の方法を示す図である。

【図６Ａ】本発明の方法に基づく、図１Ａの画像データ例からの構造化データの抽出を示す図である。

【図６Ｂ】本発明の方法で使用するための、図１Ａの構造化データに対応するデータベースを示す図である。

【図６Ｃ】本発明の方法に基づいて図１Ａの構造化データが置換された人工画像データを示す図である。

【発明を実施するための形態】

【００２４】

図２は、一実施形態に基づいて本発明の方法が実施されるコンピューティングシステム１０を示している。コンピューティングシステム１０は、１人以上のオペレータ２５（human operator 25）が物理的にアクセス可能な１台以上のクライアントコンピュータ２０を備える。

【００２５】

また、コンピューティングシステム１０は、１つ又は複数のサーバ５０を備える。サーバ５０は、通常、リモートサーバであり、すなわち、サーバ５０は、オペレータ２５が物理的にアクセスできないように、クライアントコンピュータ２０とは異なる場所に配置されている。幾つかの例では、リモートサーバ５０は、仮想サーバである。図２では、サーバ５０－１、５０－２、５０－３は、クラウドコンピューティング環境６０内の仮想サーバである。クライアントコンピュータ２０及びサーバ５０は、少なくとも１つの通信ネットワーク３０を介して互いに通信可能に接続されている。この通信可能な接続により、クライアントコンピュータ２０とサーバ５０との間でデータを通信できる。少なくとも１つの通信ネットワーク３０は、通常、インターネット（すなわち、ＩＰ、ＩＰｖ４、ＩＰｖ６）を含む。インターネットに加えて又はインターネットに代えて、セルラーネットワーク（すなわち、３Ｇ、４ＧＬＴＥ、５Ｇ）、ローカルエリアネットワーク、クラウドネットワーク、無線ネットワーク、又は任意の他の既知の通信ネットワーク等の他の通信ネットワークが存在してもよい。

【００２６】

また、コンピューティングシステム１０には、ソフトウェア開発者であるオペレータ４５（human operator 45）がアクセス可能な開発者コンピュータ４０が存在する。開発者コンピュータ４０は、サーバ５０に通信可能に接続されており、これにより、オペレータ４５は、サーバ５０上で実行されているソフトウェアアプリケーション５０Ａをセットアップ、開発、構成、スケジューリング、又は監視できる。これに代えて又はこれに加えて、開発者コンピュータ４０は、クライアントコンピュータ２０に通信可能に接続されており、これにより、オペレータ４５は、クライアントコンピュータ２０上で実行されているソフトウェアアプリケーション２０Ａをセットアップ、開発、構成、スケジューリング、又は監視できる。

【００２７】
ここに説明したコンピューティングシステム１０は、例示的なものに過ぎず、システム構成要素の削除又は追加を含む変更が可能である。
【００２８】
図３は、図２に示すコンピューティングシステム１０の選択された側面を示している。具体的には、図３は、通信ネットワーク３０を介してサーバ５０と通信するクライアントコンピュータ２０を示している。クライアントコンピュータ２０は、１つ又は複数のソフトウェアアプリケーション２０Ａ、プロセッサ２０Ｂ、メモリ２０Ｃ、１つ又は複数の入力周辺機器２０Ｄ、及び１つ又は複数の出力周辺機器２０Ｅを備える。プロセッサ２０Ｂは、中央処理装置（central processing unit：ＣＰＵ）及び／又はグラフィック処理装置（graphical processing unit：ＧＰＵ）を含む。メモリ２０Ｃは、データ記憶装置及び／又は半導体メモリを含む。データ記憶装置は、ハードディスクドライブ、ソリッドステートドライブ、外部ドライブ、リムーバブル光ディスク、及び／又はメモリカードの形態をとる。半導体メモリは、データを一時的に記憶するための揮発性メモリ、例えば、ランダムアクセスメモリ（ＲＡＭ）、及びデータを長期的に記憶するための不揮発性メモリ、例えば、リードオンリメモリ（ＲＯＭ）、フラッシュメモリの形態をとる。
【００２９】
１つ又は複数のソフトウェアアプリケーション２０Ａは、メモリ２０Ｃにコンピュータプログラムとして格納され、クライアントコンピュータ２０上でプロセッサ２０Ｂによって実行される。入力周辺機器２０Ｄ及び出力周辺機器２０Ｅを介してオペレータ２５との直接的なインタラクションを実現するこれらのソフトウェアアプリケーションは、オペレーティングシステム（ＯＳ）及びデスクトップアプリケーションを含む。既知のオペレーティングシステムの例には、マイクロソフトウィンドウズ(登録商標)、ＭａｃＯＳ、及びＬｉｎｕｘ(登録商標)が含まれる。クライアントコンピュータ２０用の既知のデスクトップアプリケーションの例には、Ｇｏｏｇｌｅ　Ｃｈｒｏｍｅ等のウェブブラウザ、Ｍｉｃｒｏｓｏｆｔ　Ｗｏｒｄ等の文書作成アプリケーション、Ｍｉｃｒｏｓｏｆｔによるリモートデスクトッププロトコル（remote desktop protocol：ＲＤＰ）やリモートフレームバッファ（remote framebuffer：ＲＦＢ）プロトコル等のリモートアクセスアプリケーションが含まれる。但し、本発明は、ここで述べた特定のアプリケーションと共に使用することに限定されるものではない。
【００３０】
上述したように、クライアントコンピュータ２０は、１つ又は複数の入力周辺機器２０Ｄを備える。入力周辺機器２０Ｄの目的は、オペレータ２５がクライアントコンピュータ２０に指示を送ることを可能にすることである。入力周辺機器２０Ｄの例には、マウス、キーボード、タッチスクリーン、イメージスキャナ、バーコードリーダ、ゲームコントローラ、マイク、デジタルカメラ、ウェブカメラ等が含まれる。
【００３１】
また、クライアントコンピュータ２０は、１つ又は複数の出力周辺機器２０Ｅを備える。出力周辺機器２０Ｅの目的は、オペレータ２５がクライアントコンピュータ２０から情報を受け取ることを可能にすることである。出力周辺機器２０Ｅの例には、ディスプレイ装置（例えば、コンピュータモニタやプロジェクタ）、プリンタ、ヘッドフォン、コンピュータスピーカ等が含まれる。入力周辺機器２０Ｄと同様に、出力周辺機器２０Ｅは、クライアントコンピュータ２０と一体化されていてもよく、クライアントコンピュータ２０の外部に設けられていてもよい。オペレータ２５は、視覚や聴覚等の感覚を用いてアプリケーション２１のＵＩを解釈することにより、出力周辺機器２０Ｅを用いてクライアントコンピュータ２０から情報を受け取る。
【００３２】
クライアントコンピュータ２０には、（図３には示していない）他の構成要素も存在する。例えば、コンピュータ２０は、通信ネットワーク３０を介した通信を可能にするネットワークアダプタカード、電源、マザーボード、サウンドカード等の１つ又は複数を備え

る。

【００３３】

開発者コンピュータ４０は、図３に示すクライアントコンピュータ２０と同じ構成要素を有する。開発者コンピュータ４０とクライアントコンピュータ２０の違いとして、開発者コンピュータ４０は、サーバ５０において、サーバ５０上で実行されている１つ又は複数のソフトウェアアプリケーション５０Ａをセットアップ、開発、設定、スケジューリング、及び監視できるアクセス権を有しており、クライアントコンピュータ２０は、このようなアクセス権を有していない。これに加えて又はこれに代えて、開発者コンピュータは、クライアントコンピュータ２０において、クライアントコンピュータ２０上で実行されている１つ又は複数のソフトウェアアプリケーション２０Ａをセットアップ、開発、構成、スケジューリング、及び監視できるアクセス権を有しており、クライアントコンピュータ２０は、開発者コンピュータ４０と同等の機能を有していない。したがって、開発者コンピュータ４０は、オペレータ４５がこのセットアップ、開発、構成、スケジューリング、及び監視を実行できるように、追加のソフトウェアアプリケーション２０Ａを有していてもよい。例えば、マイクロソフト社のリモートデスクトッププロトコル（remote desktop protocol：ＲＤＰ）やリモートフレームバッファ（remote framebuffer：ＲＦＢ）プロトコル等のリモートアクセスアプリケーションをこの目的に使用できる。

【００３４】

図３に示すように、サーバ５０は、１つ又は複数のソフトウェアアプリケーション５０Ａに加えて、プロセッサ５０Ｂ、メモリ５０Ｃ、及びマシンインタフェース５０Ｄを備える。１つ又は複数のアプリケーション５０Ａは、メモリ５０Ｃにコンピュータプログラムとして格納され、プロセッサ５０Ｂによってリモートサーバ５０上で実行される。

【００３５】

サーバ５０は、単一のサーバ（例えば、図２に示すサーバ５０）又は複数のサーバ（例えば、図２に示すサーバ５０－１、５０－２、５０－３）の形態をとることができ、あるいは、分散サーバの形態をとることもできる。分散サーバは、構成要素であるコンポーネントに処理とデータを分散して動作する。サーバ５０は、物理サーバでも仮想サーバでもよい。サーバ５０が仮想サーバである場合、ソフトウェアアプリケーション５０Ａ、プロセッサ５０Ｂ、メモリ５０Ｃ及びマシンインタフェース５０Ｄは全て、コンピュータシステム１０のクラウドコンピューティング環境６０においてホストされる仮想エンティティである。

【００３６】

サーバ５０上の１つ又は複数のソフトウェアアプリケーション５０Ａは、必ずしも、入力周辺機器２０Ｄ及び出力周辺機器２０Ｅを介してオペレータ２５又はオペレータ４５と直接インタラクションするとは限らない。これに代えて、１つ又は複数のソフトウェアアプリケーション５０Ａは、通信ネットワーク３０及びマシンインタフェース５０Ｄを介してクライアントコンピュータ２０又は開発者コンピュータ４０と直接インタラクトするアプリケーションであってもよい。幾つかの例では、サーバ５０上のソフトウェアアプリケーション５０Ａは、オプションとして、開発者コンピュータ４０上のソフトウェアアプリケーション２０Ａを介して、ソフトウェアアプリケーション５０Ａのセットアップ、開発、構成、スケジューリング、及び監視を実行する際にオペレータ４５を支援するための開発者インタフェースを開発者コンピュータ４０に提供できる。この開発者インタフェースは、クライアントコンピュータ２０を操作するオペレータ２５には提供されない。

【００３７】

１つ又は複数のソフトウェアアプリケーション５０Ａは、クライアントコンピュータ２０からのデータ又はクライアントコンピュータ２０に関連するデータを使用できる。このデータは、本明細書で更に説明する「機密データ」である可能性がある。クライアントコンピュータ２０からのデータ又はクライアントコンピュータ２０に関連するデータを使用する例示的なソフトウェアアプリケーション５０Ａは、米国特許出願第１４／０５３３１９号及び米国特許第１０，４６９，５７２号に記載されているようなロボットプロセス自

動化（robotic process automation：ＲＰＡ）アプリケーションである。ＲＰＡアプリケーションでは、クライアントコンピュータ２０からのデータ又はクライアントコンピュータ２０に関連するデータを使用して、自動化されたプロセスが実行される。

【００３８】

ある特定の実施形態において、コンピュータシステム１０を使用してＲＰＡアプリケーションを実行する場合、図２に示すように、クラウドコンピューティング環境６０内の複数の仮想サーバ５０－１、５０－２、５０－３に加えて、物理サーバ５０が存在してもよい。本実施形態において、物理サーバ５０は、プロセス定義、ログ、監査、及びユーザ情報を保持する集中リポジトリであるデータベースサーバである。データベースサーバは、複数の仮想サーバのうちの第１の仮想サーバ５０－１と通信する。第１の仮想サーバ５０－１は、データベースサーバと、第２及び第３の仮想サーバ５０－２、５０－３との間の接続を制御するアプリケーションサーバである。アプリケーションサーバは、仮想Ｗｉｎｄｏｗｓサーバとしてプロビジョニングされ、セキュアなクレデンシャル管理（secure credential management）、データベース接続マーシャリング（database connection marshalling）、データ暗号化、スケジューリングされたプロセス実行等の機能を含んでいてもよい。第２の仮想サーバ５０－２は、ロボットプロセス自動化のための自動化されたプロセスの実行を担当する、通常、標準化されたエンドユーザデスクトップの仮想化されたインスタンスをホストする。第３の仮想サーバ５０－３は、自動化プロセスのセットアップ、開発、設定、スケジューリング、及び監視を実現するエンドユーザデスクトップ構築である。第３の仮想サーバ５０－３は、専用のソフトウェアアプリケーション２０Ａを介して、開発者コンピュータ４０にアクセス可能である。

【００３９】

機密データ

多くの場合、図２に示す例示的なコンピュータシステム１０のようなコンピュータシステムは、機密データを扱うことが要求される。例えば、開発者コンピュータ４０は、ソフトウェアアプリケーション５０Ａのセットアップ、開発、設定、スケジューリング、又は監視を行うために、サーバ５０上で実行されているソフトウェアアプリケーション５０Ａ（例えば、ＲＰＡアプリケーション）にアクセスする必要がある場合がある。しかしながら、ソフトウェアアプリケーション５０Ａは、機密データを使用している可能性があり、すなわち、開発者コンピュータ４０を操作するオペレータ４５によって閲覧されるべきではないクライアントコンピュータ２０からのデータ又はクライアントコンピュータ２０に関連するデータを使用している可能性がある。別の例では、開発者コンピュータ４０は、機密データを使用するクライアントコンピュータ２０上のソフトウェアアプリケーション２０Ａを設定、開発、構成、スケジューリング、又は監視するために、クライアントコンピュータ２０にアクセスする必要がある場合がある。

【００４０】

ここで言う機密データとは、その認知的内容により、高度なセキュリティ配慮を必要とする特別なタイプのデータである。機密データの侵害は、機密データの偶発的又は違法な破壊、紛失、改ざん、機密データの不正な開示、又は機密データへのアクセスを引き起こし、重大な人的被害をもたらす可能性がある。例えば、ある個人の医療記録が永久的に削除されると、その個人の健康に重大かつ長期的な影響が及ぶ可能性がある。このため、欧州連合（ＥＵ）では、一般データ保護規則（General Data Protection Regulation：ＧＤＰＲ）、英国では、データ保護法（Data Protection Act 2018）等、様々な法域で機密データの保存と処理が規制されている。

【００４１】

機密データは、テキストの形態をとる。機密データには、個人情報、すなわち、特定又は識別可能な自然人に関連する情報が含まれる場合がある。例えば、機密データには、氏名、住所、生年月日、電話番号等が含まれる。他の種類の機密データには、個人の位置データ、オンライン識別情報、個人の身体的、生理的、遺伝的、精神的、経済的、文化的若しくは、社会的アイデンティティに固有の１つ又は複数の要素が含まれる。これに加えて

又はこれに代えて、機密データには、クレジットカード番号や銀行番号等の金融情報が含まれる場合がある。更に、別の例として、機密データには、医療情報が含まれる場合がある。

【００４２】

本発明の目的のための機密データは、構造化データの形態をとる。本明細書で使用する「構造化データ」という用語は、行（レコード）と列（フィールド）で構造化された、リレーショナルデータベース等の電子ファイル内に格納できるデータを意味する。例えば、日付を取得するには、日付フィールド（日付列）にアクセスする。これに対し、「構造化されていない」自由フォーマットのテキストから意味を導き出すには、テキストを順次スキャンして比較する必要がある。

【００４３】

機密データは、定義された視覚的フォーマットのみでなく、定義された機能的フォーマットを有する。本明細書で使用する「機能的フォーマット」という用語は、ソフトウェアアプリケーション２０Ａを介してクライアントコンピュータ２０の動作を制御するような技術システムにおいて技術的機能を有する機密データの部分、特にその部分のフォーマットを指す。ここで使用する「視覚的フォーマット」という用語は、機密データの提示の手法を指す。

【００４４】

機能的フォーマットは、エンティティタイプ及び／又はエンティティタイプフォーマットのうちの１つ又は複数を含むことができる。エンティティタイプは、例えば、氏名、日付及び／又は時刻、電子メールアドレス、住所等、機密データが何を示しているかについての情報を提供する。電子メールアドレスの機能的フォーマットは、電話番号の機能的フォーマットとは異なるため、使用可能なエンティティタイプフォーマットの種類を絞り込むことができる。エンティティタイプフォーマットは、所与のエンティティタイプのフォーマットである。日付及び／又はや時刻のエンティティタイプは、様々なエンティティタイプフォーマットを有することができる。例えば、日付は、「DD/MM/YYY」（欧州標準フォーマット）、「MM/DD/YYYY」（米国標準フォーマット）、「YYYY/MM/DD」（日本標準フォーマット）で表すことができる、「D」は、日、「M」は、月、「Y」は、年を表す。時間は「hh:mm」、「hh:mm:ss」で表すことができ、「h」は、２４時間制の時、「m」は、分、「s」は、秒である。１２時間制（ＡＭ、ＰＭ）、異なるタイムゾーン（ＧＭＴ、ＥＳＴ等）等、その他の時間フォーマットもある。別の例において、英国の郵便番号の場合、エンティティタイプのフォーマットは、「A9AA 9AA」、「A9A 9AA」、「A9 9AA」、「A99 9AA」、「AA9 9AA」、又は「AA99 9AA」（ここで、「A」は文字、「9」は、数字を表す）といった設定された数のフォーマットのいずれかでなければならない。

【００４５】

視覚的フォーマットは、テキスト長、テキストフォント、テキスト色、及び／又は背景色の１つ又は複数を含んでいてもよい。テキスト長は、テキストの水平方向の長さを表す。テキスト長は、テキストの文字数に基づいて測定してもよく（例えば、「Ben Carter」は、スペースを含めて１０文字）、ピクセル数を使用して測定してもよい。テキストフォントには、テキストの書体（例えば、Ａｒｉａｌ、Ｔｉｍｅｓ　ＮｅｗＲｏｍａｎ、Ｃｏｕｒｉｅｒ　Ｎｅｗ）、テキストのサイズ（例えば、１２ｐｔ）、及び特殊なスタイル特性（太字、斜体、下線、取り消し線、下付き文字、上付き文字等）が含まれる。テキスト色は、テキストの主要な色であり、通常はＲＧＢスケール又はＨＳＬスケールで測定される。背景色は、テキストを囲む背景の主な色であり、これも通常はＲＧＢスケール又はＨＳＬスケールで測定される。

【００４６】

図４は、機密データ４１０を含むグラフィカルユーザインタフェース（ＧＵＩ）４００の例である。この例では、機密データ４１０は、日付及び／又は時刻４１０Ａ、ドメイン名４１０Ｂ、電子メールアドレス４１０Ｃ、氏名４１０Ｄ、及び電話番号４１０Ｅといっ

たエンティティタイプを含む。日付及び／又は時刻４１０Ａとして示されるフォーマットには、「ddd DD/MM/YYYY hh:mm」及び「小数時間」の２つのエンティティタイプフォーマットがある。視覚的フォーマットとして、機密データは、同じフォント（Ｃａｌｉｂｒｉ）を有し、グレーのテキスト色と白の背景色を有する。この例については、本発明の方法を説明する文脈で、図６を参照して更に詳しく後述する。このＧＵＩは、ＲＰＡアプリケーションに見られる典型的なＧＵＩである。

【００４７】

図４に示すタイプのような機密データを再編集するための様々な手法が知られている。しかしながら、このような手法は、クライアントコンピュータ２０からの又はクライアントコンピュータ２０に関連する機密データがソフトウェアアプリケーション５０Ａによって使用され、ソフトウェアアプリケーション５０Ａが開発者コンピュータ４０上でオペレータ４５によって更に開発及びテストされる必要がある場合のシナリオでの使用には適していない。この理由は、既知の技術では、機密データの機能的フォーマットのみでなく視覚的フォーマットも置換又は除去され、データの機能的フォーマット及び視覚的フォーマットが使用可能にならないため、オペレータ４５がソフトウェア開発及びテストタスクを実行することが困難になるからである。

【００４８】

方法の概説

図５は、既知のデータ再編集技術と同様の欠点を有することなく、機密データを不明瞭化するために本発明が採用する方法５００を示している。図５の方法５００は、クライアントコンピュータ２０のプロセッサ２０Ｂによって実行してもよく、サーバ５０のプロセッサ５０Ｂによって実行してもよい。開発者コンピュータ４０は、機密データを含む画像データにアクセスすべきではないため、図５の方法は、開発者コンピュータ４０のプロセッサ２０Ｂによって実行されるべきではない。

【００４９】

図５に示すように、方法５００は、プロセッサによって実行される以下のステップを含む。

【００５０】

・プロセッサによって、画像データを取得する（ステップ５１０）

・プロセッサによって、機密データであり、定義された機能的フォーマット及び定義された視覚的フォーマットを有する構造化データを画像データから抽出する（ステップ５２０）

・プロセッサによって、構造化データと同じ機能的フォーマットを有する、構造化データとは異なる人工データを生成する（ステップ５３０）；

・プロセッサによって、構造化データが、構造化データと同じ視覚的フォーマットに基づいている人工データに置換された画像データに基づいて人工画像データを生成する（ステップ５４０）

・プロセッサによって、人工画像データを出力する（ステップ５５０）。

【００５１】

方法５００は、データの機密性の高い認知的内容を明らかにすることなく、機密データの機能的フォーマット、並びに視覚的フォーマットを維持しながら、機密データを不明瞭化することを可能にする。これは、開発者コンピュータ４０のオペレータ４５が、クライアントコンピュータ２０に関連する又はクライアントコンピュータ２０からの機密データを使用するサーバ５０におけるソフトウェアアプリケーション５０Ａ上のソフトウェア開発及びテストタスク、又はクライアントコンピュータ２０に関連する又はクライアントコンピュータ２０からの機密データを使用するクライアントコンピュータ２０におけるソフトウェアアプリケーション２０Ａのためのソフトウェア開発及びテストタスクを実行するために特に有用である。

【００５２】

ステップ５１０からステップ５５０については、以下のセクションで更に詳細に説明す

る。
【００５３】
画像データ取得
図５のステップ５１０において、プロセッサ２０Ｂ又はプロセッサ５０Ｂは、画像データを取得する。画像データは、クライアントコンピュータ２０に関連付けられ又はクライアントコンピュータ２０から取得される。「関連付けられる」とは、画像データがクライアントコンピュータ２０のクライアントに属するか又はクライアントコンピュータ２０のクライアントに関連することを意味する。「から」とは、クライアントコンピュータ２０から間接的に受信される画像データ及び／又はクライアントコンピュータ２０から直接的に受信される画像データの両方を意味する。
【００５４】
画像データには、クライアントコンピュータ２０から見て機密データであるデータが含まれている可能性がある。これは、クライアントコンピュータ２０を操作するオペレータ２５は、機密データを閲覧し、インタラクトする権限を有するが、開発者コンピュータ４０を操作するオペレータ４５は、その権限がないことを意味する。
【００５５】
画像データは、グラフィカルユーザインタフェース（ＧＵＩ）の形態をとることができる。例えば、画像データは、（サーバ５０がオペレータによってアクセス可能なＧＵＩを有する実施形態では）クライアントコンピュータ２０又はサーバ５０からキャプチャされたＧＵＩ画像であってもよい。ＧＵＩは、クライアントコンピュータ２０上で実行されているソフトウェアアプリケーション２０Ａ又はサーバ５０上で実行されているソフトウェアアプリケーション５０Ａのうちの１つ又は複数を示すことができる。ＧＵＩは、（オペレータに出力されるような）デスクトップインタフェース全体を含んでいてもよく、ソフトウェアアプリケーション２０Ａ又は５０Ａの特定の１つに関連するデスクトップインタフェースの部分のみを含んでいてもよい。画像データがＧＵＩの形態をとる場合、プロセッサ２０Ｂ又はプロセッサ５０Ｂは、ＧＵＩをキャプチャすることによって画像データを取得する。これに代えて、ＧＵＩは、プロセッサ２０Ｂ又はプロセッサ５０Ｂによって先にキャプチャされ、それぞれメモリ２０Ｃ又はメモリ５０Ｃに格納されていてもよい。このような例では、ＧＵＩは、メモリ２０Ｃ又はメモリ５０Ｃから読み出すことができる。幾つかの例では、ＧＵＩは、リモートアクセスアプリケーションによってキャプチャされてもよい。
【００５６】
これに代えて、画像データは、ＰＤＦ又は画像ファイル等の文書の形態をとってもよい。この場合、プロセッサ２０Ｂ又はプロセッサ５０Ｂは、画像データが格納されているメモリ２０Ｃ又はメモリ５０Ｃからそれぞれ画像データを取得する。
【００５７】
図６Ａは、画像データの一例を示す図である。具体的には、図６Ａは、Ｍｉｃｒｏｓｏｆｔ　Ｏｕｔｌｏｏｋ等の電子メールアプリケーションのＧＵＩ６００を示している。このＧＵＩには、氏名、日付、電話番号、電子メールアドレス、ドメイン名等の機密データが含まれている。
【００５８】
構造化データ抽出
図５のステップ５２０では、プロセッサ２０Ｂ又はプロセッサ５０Ｂ（ステップ５１０で画像データを取得した一方）が画像データから構造化データを抽出する。上述したように、構造化データは、定義された機能的フォーマット及び定義された視覚的フォーマットを有する機密データである。したがって、このステップの目的は、機密データである画像データ内の構造化データを特定して抽出することである。
【００５９】
画像データは生来的に構造化されていないが、構造化されていない画像データから構造化データを識別する方法は既知である。例えば、Ｇｏｏｇｌｅは、コンピュータビジョン

（光学式文字認識（ＯＣＲ）を含む）及び自然言語処理（ＮＬＰ）を使用して、文書に対して事前に訓練されたモデルを作成するＤｏｃｕｍｅｎｔ　ＡＩを使用している。

【００６０】

ステップ５２０を実行するための１つの方法は、最初に画像データに対して光学式文字認識を実行して、画像データ内のテキストを識別し、テキスト内の構造化データを識別することである。これは、Ｄｏｃｕｍｅｎｔ　ＡＩ等の既知の方法を使用して実行できる。構造化データがテキスト内で識別された後、ステップ５２０を実行することは、構造化データに対応する画像データ内の１つ又は複数のバウンディングボックスを決定することと、１つ又は複数のバウンディングボックスを使用して画像データから１つ又は複数の画像部分を抽出することとを含んでもよい。機密データであると識別された構造化データのみについて、画像データから対応する画像部分を抽出することが望ましい。

【００６１】

上述した方法を用いて抽出された、機密データに対応する構造化データをそれぞれ含む画像データからの１つ又は複数の画像部分に加えて、ステップ５２０は、第２の出力を有する。すなわち、１つ又は複数の画像部分内の機密データに対応する基底にある構造化データも出力される。このデータは、ステップ５３０の準備として一覧化（tabulated）される（表にされる）。

【００６２】

図６Ａの画像データ例であるＧＵＩ６００は、構造化データを含む複数のバウンディングボックス６１０を示している。複数のバウンディングボックス６１０は、機密データを含む複数の画像部分を形成するように抽出される。更に、画像部分内の構造化データが抽出される。ＧＵＩ６００の場合、構造化データには、「Ben Benjamin Carter」、「Ben Carter」、「De Zhang」、「Eric Tyree」、「John Reid」、「Krishna Dubba」、及び「+44 785 407 9884」が含まれる。この構造化データは、図６Ｂの表６２０の列６１５に示すように、一覧化できる。

【００６３】

人工データ生成

図５のステップ５３０において、プロセッサ２０Ｂ又はプロセッサ５０Ｂ（ステップ５１０において画像データを取得した一方）は、構造化データとは異なるが、構造化データと同じ機能的フォーマットを有する人工データを生成する。このステップの目的は、機密性の高い内容を含むことなく機能的フォーマットを維持した、構造化データの代替データを提供することである。

【００６４】

構造化データとは異なるが、構造化データと同じ機能的フォーマットを有する人工データを生成できるようにするには、データの機能的フォーマットを識別できるようにする必要がある。これは、機械学習モデルやヒューリスティックルールを利用することで達成される。機械学習モデルは、ニューラルネットワーク等の機械学習技術を使用して事前に訓練される。具体的には、テキスト中のエンティティタイプを認識するためのニューラルネットワークモデルは、ＲｏＢＥＲＴａ等のトランスフォーマアーキテクチャに基づいていてもよい。例えば、Ｈｕｇｇｉｎｇ　Ｆａｃｅ社（https://huggingface.co/roberta-base参照）が提供する事前訓練済みＲｏＢＥＲＴａモデルの基本バージョン（「ｒｏｂｅｒｔａ－ｂａｓｅ」）は、７６８の隠れベクトルサイズを持つ１２個のエンコーダ層、１２個のアテンションヘッド、１２５Ｍのパラメータで構成されている。このモデルは、強化されたＢＥＲＴ（Bidirectional Encoder Representations from Transformers）アルゴリズムを用いて、様々なサイズとドメインの５つの一般公開コーパスから得た１６０ＧＢ以上の英語テキストデータで学習されている。エンティティ認識の目的のために、事前に学習されたｒｏｂｅｒｔａ－ｂａｓｅモデルは、よく見られる１８のエンティティタイプを含むＯｎｔｏＮｏｔｅｓ５データセット（https://deepai.org/dataset/ontonotes-v5-english等参照）等のラベル付きＮＥＲ（Named Entity Recognition）データセットによって微調整できる。ヒューリスティックルールには、（重み付けされた）正規表

現のテキストパターンや、周囲の単語等の文脈上の手がかりを含めることができる。例えば、正規表現「＼b([0-9]{10})＼b」を使用して、米国の電話番号又は米国の銀行口座番号を表すと思われる１０桁のシーケンスをテキストから抽出できる。次に、ローカルコンテキスト（例えば、１０桁のシーケンスの前後５語）に「mobile」や「call」等の単語があれば、その１０桁のシーケンスは、米国の電話番号である可能性が高い。逆に、ローカルコンテキストに「savings」や「debit」といった単語があれば、１０桁の数字列が米国の銀行口座番号である可能性が高い。更に、正規表現を使用してデータの機能的フォーマットを判定することもできる。例えば、正規表現「(＼+[0-9]{1,3})?([0-9]{10})＼b」を使用すると、１０桁の電話番号の前に対応する国の通話コード（プラス記号の後に１～３桁の数字が続き、その後にスペース文字が続くフォーマット）を有するかどうかを判断できる。

【００６５】

エンティティタイプ及びエンティティタイプフォーマットの表示は、メモリ２０Ｂ又はメモリ５０Ｂに出力又は格納できる。例えば、図６Ａでは、バウンディングボックスの１つに「Mon22/02/2021 10:28」という構造化データが含まれている。上述した方法を用いると、この構造化データエンティティタイプは「日付及び／又は時刻」であるとみなされ、エンティティタイプのフォーマットは「ddd DD/MM/YYYY hh:mm」であるとみなされる。

【００６６】

次に、人工データを生成するために、オペレータ２５が編集可能なテーブルを使用することが好ましい。具体的には、構造化データの機能的フォーマットを特定した後、画像から抽出された構造化データを一覧化する。この一覧化は、識別された機能的フォーマットの補助によって行うことができる。更に、機能的フォーマット自体を一覧化することもできる。図６Ｂは一例を示しており、表６２０は、識別された機能的フォーマットエンティティタイプを使用する第１の列６２５と、対応する構造化データを示す第２の列６１５とを有する。この例では、エンティティタイプは、アルファベット順にソートされ、これにより、エンティティタイプがグループ化されている。

【００６７】

その後、テーブルの各行に対して、構造化データと同じ機能的フォーマットを有するが、異なる認知的内容を有し、したがって機密データではない人工データが生成される。例えば、日付及び／又は時刻のエンティティタイプである構造化データ「Mon 22/02/2021 10:28」に対して、これを同じ機能的フォーマットの人工的な日付及び／又は時刻、例えば「Fri 24/06/1987 19:03」に変更できる。その後、人工データは、構造化データ及びオプションとして機能的フォーマットとともに一覧化される。図６Ｂは、生成された人工データを含む表６２０の第３列６６０を示している。

【００６８】

構造化データと同じ機能的フォーマットを有する人工データの生成は、エンティティタイプとエンティティタイプフォーマットを条件とするエンティティの確率分布からのランダムサンプリングによって実行される。例えば、人工的な人名を生成するには、姓（surname）のリストから姓を、名（first-name）のリストから名をランダムにサンプリングする。そして、姓と名は、構造化データの同じフォーマット（例えば、「SURNAME, First-name」）にまとめられる。別の例として、人工的な日付を生成するために、元の日付から５年以内の距離にあるカレンダーの日付がランダムにサンプリングされ、その後、構造化データの同じフォーマット（例えば、「ddd DD/MM/YYYY」）を使用して、日付のテキスト表現が生成される。オプションとして、生成された人工データが、対応するソフトウェアアプリケーションによって要求される有効範囲内にあることを保証するために、ポストフィルタリングを実施してもよい。

【００６９】

一覧化された構造化データ及び一覧化された人工データは、オペレータ２５が編集できるように、編集可能な表としてクライアントコンピュータ２０上に出力してもよい。これ

により、オペレータは、構造化データと人工データとのマッピングを適切と思われるように追加、削除、又は編集できる。例えば、オペレータ２５は、画像データ内の構造化データ検出に修正を加えるように編集可能テーブルを編集できる。具体的には、ユーザは、表に行を追加することによって、見逃された機密データエンティティを追加してもよく、表の行を削除することによって、誤検出された機密データを削除してもよく、表の対応するセルを編集することによって、機密データ検出の誤りを修正してもよく（例えば、図６Ｂにおいて、誤って識別された「Ben Benjamin Carter」を「Benjamin Carter」に変更してもよく）、又は自動的に生成された偽のデータエンティティを、ユーザが好むものに単に置換してもよい（例えば、図６Ｂにおいて、Krishna Dubbaの偽の氏名を「Michele Oneal」を「Dr Strangelove」に変更してもよい）。

【００７０】

ステップ５４０において人工画像データを生成する前にオペレータ２５に人工データを承認させることは有用である。したがって、プロセッサ２０Ｂは、人工データを確認するためのプロンプトをオペレータ２５に送信してもよい。

【００７１】

編集可能テーブルは、異なる（例えば、第２、第３等の）画像データからの構造化データ及び人工データを含むことができる。このような例では、後続の構造化データ（例えば、第２の構造化データ）が先の構造化データと同一である場合、後続の画像データ（例えば、第２の人工データ）に基づいて生成される人工データは先の人工データと同一であることが好ましい。例えば、図６ＡのＧＵＩ６００を参照すると、第２の画像データが、Ben Carterからの電子メールを含む更なるＧＵＩに関連する場合、構造化データ「Ben Carter」は、元の画像データ及び後続の第２の画像データの両方に現れることになる。そして、構造化データ「Ben Carter」が元の画像データに由来するか第２の画像データに由来するかにかかわらず、生成される人工データは、Anne Wellsとなる（図６Ｂ参照）。編集可能テーブルは、第２の画像データの「Ben Carter」のインスタンスのために新たな行を生成する必要はなく、元の画像データから既に生成されている行を利用できる。このように、編集可能テーブルは、１つ又は複数の画像データにわたる構造化データと人工データとのマッピングのグローバルテーブルであるとみなすことができる。

【００７２】

人工画像データの生成

図５のステップ５４０において、プロセッサ２０Ｂ又はプロセッサ５０Ｂは、構造化データが人工データに置換された画像データに基づいて人工画像データを生成し、人工データは、構造化データと同様の視覚的フォーマットに基づいている。

【００７３】

ステップ５４０を実行するために、プロセッサ２０Ｂ又はプロセッサ５０Ｂは、まず、１つ又は複数の画像部分における構造化データの視覚的フォーマットを識別し、次に、１つ又は複数の画像部分における構造化データの視覚的フォーマットに基づいて、１つ又は複数の画像部分に対応する１つ又は複数の人工画像部分を生成し、次に、画像データを変更して１つ又は複数の画像部分を１つ又は複数の人工画像部分に置換する。

【００７４】

視覚的フォーマットの識別は、Ｐｙｔｈｏｎ Ｉｍａｇｅ Ｌｉｂｒａｒｙ（Ｐｉｌｌｏｗ）に実装されているようなデジタル画像処理技術や、ＯｐｅｎＣＶライブラリによって提供されているようなコンピュータビジョン技術を利用することによって実行される。テキスト長、テキスト色、背景色を検出するための既知の技術が存在する。テキストフォントの認識は、フォントの書体、フォントサイズ、及びフォントスタイルの可能な構成に対してグリッド検索又はベイズ最適化を実行し、どの構成が元の画像部分に最も類似した構造化データの画像を生成するかを見つけることによって達成できる。

【００７５】

１つ又は複数の人工画像部分の生成は、Ｐｙｔｈｏｎ Ｉｍａｇｅ Ｌｉｂｒａｒｙ（Ｐｉｌｌｏｗ）に実装されているようなデジタル画像処理技術を利用して実行される。例え

ば、Ｐｙｔｈｏｎ Ｉｍａｇｅ Ｌｉｂｒａｒｙ（Ｐｉｌｌｏｗ）は、新しい画像を作成したり、既存の画像に注釈を付けたり、レタッチしたりするために使用できるＩｍａｇｅＤｒａｗモジュールを提供する。具体的には、ＩｍａｇｅＤｒａｗ．ｒｅｃｔａｎｇｌｅ関数を使用すると、指定された位置に、指定されたサイズで、指定された塗りつぶし色を背景とする矩形ボックスを描画でき、ＩｍａｇｅＤｒａｗ．ｔｅｘｔ関数を使用すると、指定された位置に、指定されたフォントと色で、指定されたテキスト（すなわち、元の機能的フォーマットで生成された人工データ）の一部を描画できる。ここで、テキストフォント、テキスト色、及び背景色は、全て、上述の方法を介して識別された視覚的フォーマットによって与えられる。

【００７６】

幾つかの例では、人工画像データは、画像データ中の構造化データと同じ視覚的フォーマットを有するように生成される。これにより、構造化データの視覚的フォーマットに依存するソフトウェアアプリケーション２０Ａ上でソフトウェア開発を行うオペレータ４５が視覚的フォーマットが何であるかを高い精度で理解できるようになる。あるいは、人工画像データは、人工データに関係する部分と画像データに関係する部分とが区別されるように、画像データの構造化データとは１つ又は複数の相違点を有する視覚的フォーマットを有するように生成できる。例えば、図６Ｃの人工ＧＵＩ６５０の人工画像データにおいて、人工画像部分は、これらの部分が人工的であることを示すために、強調表示された背景を有する。この場合でも、人工画像データは、構造化データと同様の視覚的フォーマットに基づくものとみなされる。

【００７７】

人工画像データの出力

図５のステップ５５０において、プロセッサ２０Ｂ又はプロセッサ５０Ｂ（画像データを受信した一方）は、人工画像データを出力する。

【００７８】

人工画像データは、メモリ２０Ｃ又はメモリ５０Ｃに格納されるファイルに出力してもよい。このファイルは、後日、開発者コンピュータ４０において、オペレータ４５がソフトウェア開発の目的でアクセスできる。ファイルは、人工画像データのみを含んでいてもよい。このような場合、ファイルは、元の画像データと同じファイルフォーマット（例えば、ｐｄｆ、ｊｐｅｇ）であってもよい。あるいは、ファイルは、人工画像データと、ソフトウェア開発をガイドしサポートするための他の関連情報とを含む包括的なドキュメント（comprehensive document）であってもよい。このようなファイルの例としては、ロボットプロセス自動化（robotic process automation：ＲＰＡ）内で開発されるビジネスプロセスのフローをキャプチャするプロセス設計文書（process design document：ＰＤＤ）が挙げられる。

【００７９】

これに代えて又はこれに加えて、人工画像データは、ディスプレイに出力してもよい。例えば、オペレータ２５がチェックするために、人工画像データをクライアントコンピュータ２０に出力してもよい。他の例では、人工画像データは、プロセッサ２０Ｂ又はプロセッサ５０Ｂとは異なるコンピュータに属するディスプレイに出力してもよい。具体的には、開発者コンピュータ４０のディスプレイに人工画像データを出力してもよく、これにより、機密データを使用するソフトウェアのソフトウェア開発プロセス中に、オペレータ４５が画像データ内の機密データを閲覧できないようにできる。

【００８０】

幾つかの例では、人工画像データを使用して、機密データを使用するロボットプロセス自動化（ＲＰＡ）アプリケーションを開発してもよい。ＲＰＡでは、自動化されたプロセスやワークフローを使用して、ＧＵＩやドキュメント等の画像データから、通常は、機密データである構造化データを抽出する。自動化されたプロセスやワークフローが正しく機能しているかどうかを判断できるようにするためには、人工データ内の機密データの機能的フォーマットを維持することが重要である。ＲＰＡは、機密データを抽出するためにソ

フトウェアアプリケーションのＧＵＩに依存することが多いため、機密データの視覚的フォーマットを維持することも重要である。すなわち、機密データの視覚的フォーマットを維持することで、抽出に関する潜在的な問題を明らかにできる。

【００８１】

幾つかの例では、人工画像データは、Ｍｉｃｒｏｓｏｆｔによるリモートデスクトッププロトコル（remote desktop protocol：ＲＤＰ）やリモートフレームバッファ（remote framebuffer：ＲＦＢ）プロトコル等のリモートアクセスアプリケーションで使用されることがある。例えば、米国特許出願ＵＳ１７／１４４，６４０号（Method of Remote Access）で使用されているＧＵＩ画像変更技術の代わりに、図５の方法５００と出力された人工画像データを使用してもよい。これにより、リモートアクセスプロトコルを介して受信された機密データがオペレータ（例えば、オペレータ４５）に出力されることを防止するコンピュータ実装方法が提供される。このような例では、この方法は、リモートアクセスアプリケーションを介してサーバ５０から、リモートサーバ上で実行されているソフトウェアアプリケーション５０ＡのＧＵＩを受信することと、本発明の方法に基づいて、機密データを不明瞭化するためにＧＵＩを変更することと、変更されたＧＵＩを、オペレータ４５が受信するための、開発者コンピュータ４０の１つ又は複数の出力周辺機器４０Ｅに出力することとを含む。

【００８２】

総括

ソフトウェアで実施する場合、本発明は、コンピュータプログラムの形態をとることができる。コンピュータプログラムは、プロセッサによって使用され又はプロセッサに関連して使用されるコンピュータ実行可能コードを有するコンピュータ可読媒体として具現化できる。コンピュータ可読媒体は、プロセッサによって使用され又はプロセッサに関連して使用されるプログラムを格納、保存、通信、伝播、又は輸送できる任意の有形デバイスである。更に、コンピュータ可読媒体は、電子、磁気、光学、電磁、赤外線、半導体デバイス、又は伝搬媒体であってもよい。コンピュータ読取可能媒体の例としては、半導体メモリ、ランダムアクセスメモリ（random access memory：ＲＡＭ）、読取専用メモリ（read-only memory：ＲＯＭ）、フラッシュメモリ、ハードディスクドライブ、ソリッドステートドライブ、光ディスク、メモリカード等がある。現在の光ディスクの例としては、ＣＤ、ＤＶＤ、ブルーレイ等がある。現在のメモリカードの例としては、ＵＳＢフラッシュドライブ、ＳＤカード、ｍｉｃｒｏＳＤカード、ＭＭＣカード、ｘＤカード、メモリスティック等がある。

【００８３】

ハードウェアで実施する場合、本発明は、本明細書で説明する特定のハードウェアに限定されない。本発明は、図２及び図３に関して説明したものとは異なるハードウェアで実装しても、上述のように機能させることができることは、当業者にとって明らかである。

【００８４】

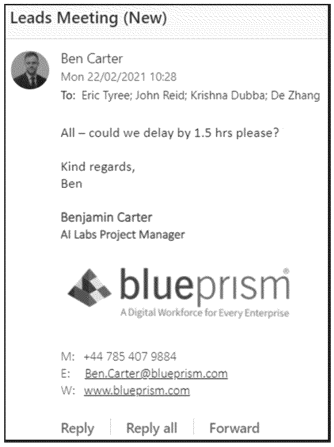
図示したフローチャートは、本発明の方法の可能な実施側面のアーキテクチャ、機能、及び動作を示している。幾つかの代替的な具体例では、図に示すステップは、図に示す順序とは異なる順序で実行してもよい。例えば、連続して示されている２つのステップは、実際には、実質的に同時に実行してもよく、ステップを表すブロックは、関係する機能によっては、逆の順序で実行してよいこともある。

【００８５】

上の説明は、例示のみを目的とし、当業者によって様々に変更できることは明らかである。上では、ある程度の特殊性をもって又は１つ又は複数の個々の実施形態を参照して様々な実施形態を説明したが、当業者は、本発明の範囲から逸脱することなく、ここに開示した実施形態に多数の変更を加えることができる。

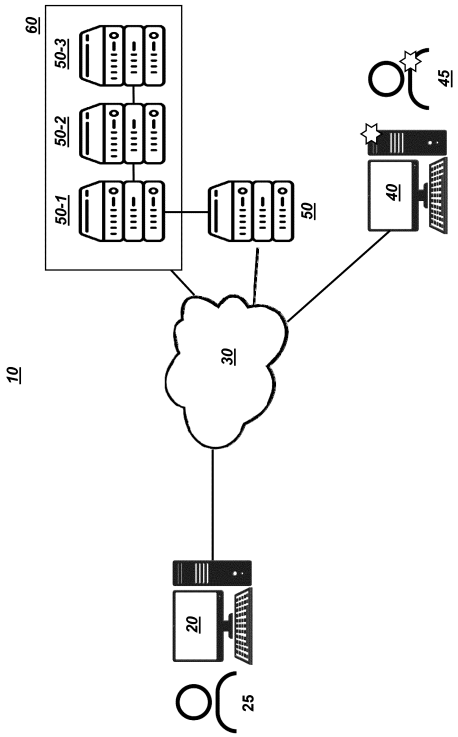
【図面】

【図１Ａ】

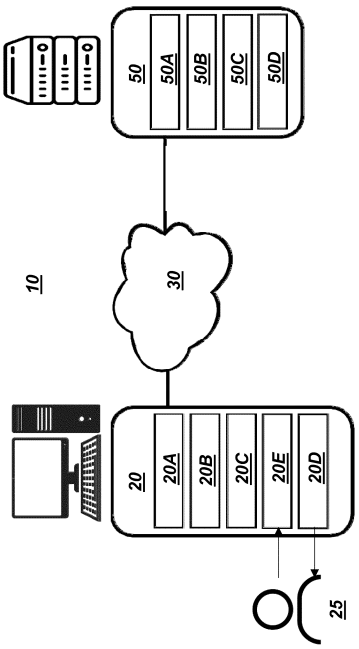


【図１Ｂ】

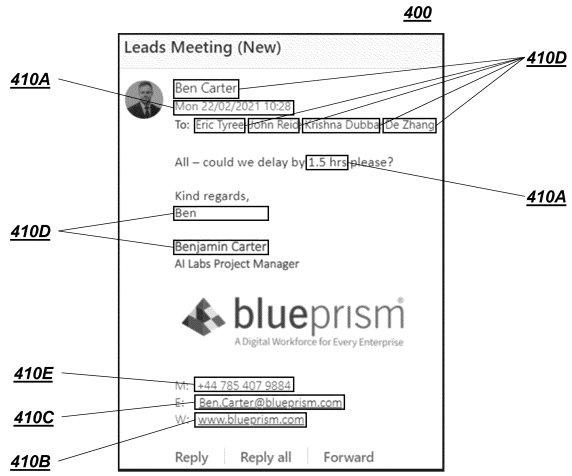
【図２】



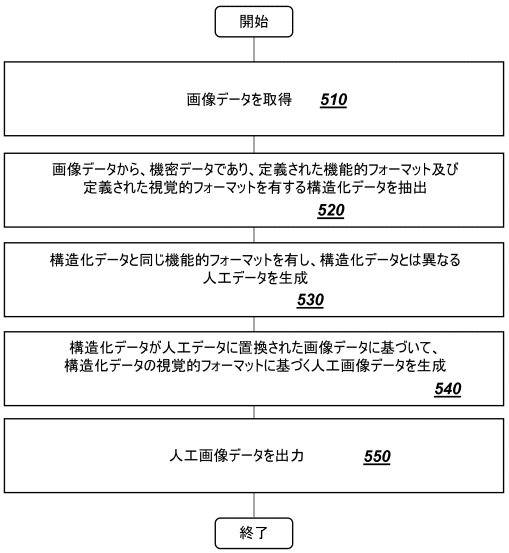
【図３】



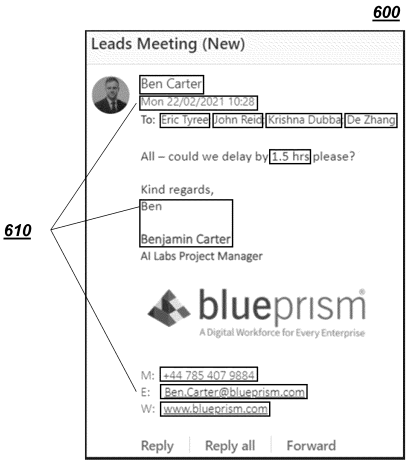
【図４】



【図５】



【図６Ａ】

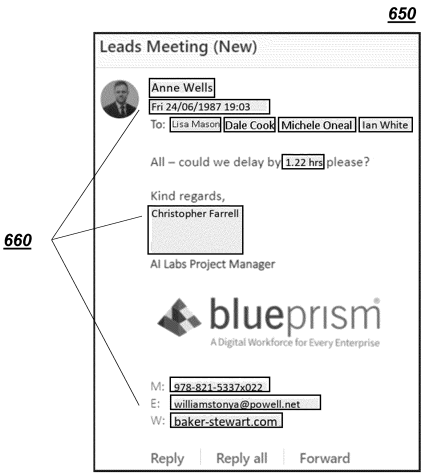


【図６Ｂ】

620

エンティティタイプ 625	構造化データ 615	人工データ 660
DATE_TIME	1.5 hrs	1.22 hrs
DATE_TIME	Mon 22/02/2021 10:28	Fri 24/06/1987 19:03
DOMAIN_NAME	www.blueprism.com	Baker-stewart.com
EMAIL_ADDRESS	Ben.Carter@blueprism.com	williamstonya@powell.net
PERSON	Ben Benjamin Carter	Christopher Farrell
PERSON	Ben Carter	Anne Wells
PERSON	De Zhang	Ian White
PERSON	Eric Tyree	Lisa Mason
PERSON	John Reid	Dale Cook
PERSON	Krishna Dubba	Michele Oneal
PHONE_NUMBER	+44 785 407 9884	978-821-5337x022

【図６Ｃ】


650
Leads Meeting (New)
Anne Wells
Fri 24/06/1987 19:03
To: Lisa Mason Dale Cook Michele Oneal Ian White
All – could we delay by 1.22 hrs please?
Kind regards,
Christopher Farrell
660
AI Labs Project Manager
blueprism
A Digital Workforce for Every Enterprise
M: 978-821-5337x022
E: williamstonya@powell.net
W: baker-stewart.com
Reply
Reply all
Forward

フロントページの続き

審査官　石坂 知樹
(56)参考文献　国際公開第２０２０／０８２１８７（WO，Ａ１）
米国特許出願公開第２０１８／０２８５５９１（ＵＳ，Ａ１）
英国特許出願公開第０２５３１７１３（ＧＢ，Ａ）
(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　４０／１５１
Ｇ０６Ｆ　２１／６２