(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2011-253528
(P2011-253528A)
(43)公開日　平成23年12月15日(2011.12.15)

(51)Int.Cl.		F I			テーマコード（参考）
G06T 7/00	(2006.01)	G06T	7/00	350B	5L096

審査請求 未請求　請求項の数 18　ＯＬ　外国語出願　（全 33 頁）

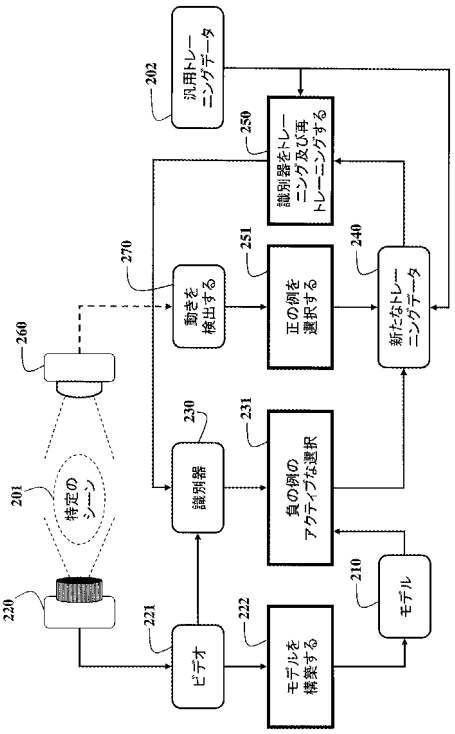
(21)出願番号　特願2011-108179 (P2011-108179)
(22)出願日　平成23年5月13日 (2011.5.13)
(31)優先権主張番号　12/791,786
(32)優先日　平成22年6月1日 (2010.6.1)
(33)優先権主張国　米国 (US)

(71)出願人　597067574
ミツビシ・エレクトリック・リサーチ・ラボラトリーズ・インコーポレイテッド
アメリカ合衆国、マサチューセッツ州、ケンブリッジ、ブロードウエイ　２０１
２０１　ＢＲＯＡＤＷＡＹ，　ＣＡＭＢＲＩＤＧＥ，　ＭＡＳＳＡＣＨＵＳＥＴＴＳ　０２１３９，　Ｕ．Ｓ．Ａ．
(74)代理人　100110423
弁理士　曾我　道治
(74)代理人　100094695
弁理士　鈴木　憲七
(74)代理人　100111648
弁理士　梶並　順

最終頁に続く

(54)【発明の名称】識別器を特定のシーン内のオブジェクトを検出するように適応させるためのシステム及び方法

(57)【要約】　（修正有）
【課題】汎用識別器を、特定のシーン、この識別器がトレーニングされたときには未知であるか又は利用可能でなかった特定のシーンに適応させることができるトレーニング方法を提供する。
【解決手段】汎用識別器が、特定のシーン内のオブジェクトを検出するように適応される。特定のシーンは識別器が汎用トレーニングデータを用いてトレーニングされたときに未知であった。カメラが、特定のシーンのフレームのビデオを取得する。ビデオ内のフレームを用いて特定のシーンモデルのモデルが構築される。識別器はモデルに適用され、負の例が選択される。そして、新たな負の例がトレーニングデータに付加される一方で、不確実性基準に基づいて、トレーニングデータから既存の負の例の別のセットが除去される。選択された正の例もトレーニングデータに付加され、識別器はシーン固有の識別器を得るための所望の精度レベルに達するまで再トレーニングされる。
【選択図】図２



【特許請求の範囲】
【請求項１】
識別器を特定のシーン内のオブジェクトを検出するように適応させるための方法であって、前記特定のシーンは、前記識別器がトレーニングデータを用いてトレーニングされたときに未知であり、前記方法は、
前記特定のシーンのフレームのビデオを、カメラを用いて取得するステップと、
前記ビデオ内の前記フレームを用いて、前記特定のシーンモデルのモデルを構築するステップと、
前記識別器を前記モデルに適用するステップであって、負の例を選択する、適用するステップと、
新たな前記負の例のサブセットを前記トレーニングデータに付加する一方で、不確実性基準に基づいて、前記トレーニングデータから既存の負の例の別のセットを除去するステップと、
選択された正の例を前記トレーニングデータに付加するステップと、
前記識別器を再トレーニングするステップと、
シーン固有の識別器を得るための所望の精度レベルに達するまで、前記付加するステップ及び前記再トレーニングするステップを反復するステップと、
を含む識別器を特定のシーン内のオブジェクトを検出するように適応させるための方法。
【請求項２】
前記構築するステップは、
混合モデルのセットを前記フレーム内の各ピクセルに適合させることによって、ベイズ背景更新メカニズムを用いて前記特定のシーンの背景を推定するステップであって、ピクセルモデルを生成する、推定するステップと、
最も可能性の高いピクセルモデルを選択するステップと、
をさらに含む請求項１に記載の識別器を特定のシーン内のオブジェクトを検出するように適応させるための方法。
【請求項３】
前記モデルは、フレーム差分を求めると共に、小さい差分値を有するピクセルをグループ化すること、及びオブジェクトサイズ窓を前記グループ化されたピクセルに適合させることによって構築され、ここで、前記窓は新たな負の例である請求項１に記載の識別器を特定のシーン内のオブジェクトを検出するように適応させるための方法。
【請求項４】
動きセンサーを用いて前記特定のシーン内の動きを検出するステップと、
前記フレーム差分を、前記動きの検出前、検出中、及び検出後に適用するステップであって、差分値を求める、適用するステップと、
最も大きな差分値を有する前記フレーム内の領域を求めるステップと、
前記オブジェクトサイズ窓を前記グループ化されたピクセルに適合させるステップであって、ここで、前記窓は新たな正の例である、適合させるステップと、
をさらに含む請求項３に記載の識別器を特定のシーン内のオブジェクトを検出するように適応させるための方法。
【請求項５】
メモリ要件及びリアルタイム処理要件に従って前記トレーニングデータを固定サイズに設定及び維持するステップ
をさらに含む請求項１に記載の識別器を特定のシーン内のオブジェクトを検出するように適応させるための方法。
【請求項６】
現在のモデルと現在のフレームとの間の差分が大きい場合、前記再トレーニングを反復することによって、前記特定のシーン内の変化に適応させるステップ
をさらに含む請求項１に記載の識別器を特定のシーン内のオブジェクトを検出するよう

に適応させるための方法。
【請求項７】
前記新たな負の例を用いてマルチクラス識別器を適応させるステップ
をさらに含む請求項１に記載の識別器を特定のシーン内のオブジェクトを検出するように適応させるための方法。
【請求項８】
前記新たな負の例及び前記正の例を用いて前記識別器を再トレーニングするステップと、
前記識別器を前記識別器内のカスケード層として付加するステップと、
をさらに含む請求項１に記載の識別器を特定のシーン内のオブジェクトを検出するように適応させるための方法。
【請求項９】
前記トレーニングデータは、最初汎用である請求項１に記載の識別器を特定のシーン内のオブジェクトを検出するように適応させるための方法。
【請求項１０】
前記窓のサイズは、７５×５０ピクセルであり、５０×３０ピクセルの水平方向及び垂直方向の重複を有する請求項３に記載の識別器を特定のシーン内のオブジェクトを検出するように適応させるための方法。
【請求項１１】
各前記フレームから特徴を抽出するステップであって、特徴ベクトルにする、抽出するステップと、
前記特徴ベクトルを分類するステップと、
をさらに含む請求項１に記載の識別器を特定のシーン内のオブジェクトを検出するように適応させるための方法。
【請求項１２】
前記特徴は、勾配ヒストグラムである請求項１１に記載の識別器を特定のシーン内のオブジェクトを検出するように適応させるための方法。
【請求項１３】
前記識別器は、サポートベクターマシンである請求項１に記載の識別器を特定のシーン内のオブジェクトを検出するように適応させるための方法。
【請求項１４】
前記識別器は、マルチクラス識別器である請求項１に記載の識別器を特定のシーン内のオブジェクトを検出するように適応させるための方法。
【請求項１５】
前記識別器は、最初汎用である請求項１に記載の識別器を特定のシーン内のオブジェクトを検出するように適応させるための方法。
【請求項１６】
前記オブジェクトは、人である請求項１に記載の識別器を特定のシーン内のオブジェクトを検出するように適応させるための方法。
【請求項１７】
ユーザーが、アクティブトレーニング中に選択されたラベル付けされていない例を選択する請求項１に記載の識別器を特定のシーン内のオブジェクトを検出するように適応させるための方法。
【請求項１８】
識別器を特定のシーン内のオブジェクトを検出するように適応させるためのシステムであって、前記特定のシーンは、前記識別器がトレーニングデータを用いてトレーニングされたときに未知であり、前記システムは、
前記特定のシーンのフレームのビデオを、取得するように構成されるカメラと、
前記ビデオ内の前記フレームを用いて、前記特定のシーンモデルのモデルを構築する手段と、

前記識別器を前記モデルに適用して負の例を選択する手段と、
新たな前記負の例のサブセットを前記トレーニングデータに付加する一方で、不確実性基準に基づいて、前記トレーニングデータから既存の負の例の別のセットを除去する手段と、
選択された正の例を前記トレーニングデータに付加する手段と、
前記識別器を再トレーニングする手段と、
を備える識別器を特定のシーン内のオブジェクトを検出するように適応させるためのシステム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、包括的にはコンピュータービジョンに関し、より詳細には、移動しているオブジェクト、特に人を検出するように一般的なオブジェクト識別器を特定のシーンに適応させることに関する。
【背景技術】
【０００２】
ビデオフレーム内のオブジェクトを検出又は分類するためのほとんどのトレーニング方法は、ビデオのラベル付けされたフレーム例を提供することによってトレーニングされる。識別器がトレーニングされた後、既知のテストフレームを処理して識別器の性能精度を求めることができる。
【発明の概要】
【発明が解決しようとする課題】
【０００３】
そのような方法は、トレーニング及びテストが、同じシーン等の類似した条件において行われるときに良好に実行される。しかしながら、トレーニング及び配備は、幅広く変動する照明、カメラ位置、オブジェクトの見かけの大きさ、及びオブジェクトの姿勢を有する様々なシーン内であり得るので、条件は多くの場合に変化する。すなわち、多くの場合に、識別器が適用されるシーンのタイプを事前に求めることができない。
【０００４】
本発明の目的は、汎用識別器を、特定のシーン、この識別器がトレーニングされたときには未知であるか又は利用可能でなかった特定のシーンに適応させることである。
【課題を解決するための手段】
【０００５】
多くのコンピュータービジョンタスクにおいて、シーン変化によって、汎用的にトレーニングされた識別器の能力が妨げられる。たとえば、１つのフレームセットを用いて人検出器用にトレーニングされた識別器は、異なるシーン条件において良好に機能する可能性が低い。
【０００６】
したがって、本発明の実施の形態は、汎用トレーニングデータを取り、識別器を異なる特定のシーンに適応させることができる（人）オブジェクト検出のためのインクリメンタルトレーニング方法及びシステムを提供する。
【０００７】
自律的モードにおいて、特定のシーン内に人が存在しない間の、ビデオの最初のいくつかのフレームが適応のために用いられる。すなわち、特定のシーンは概ね静止している。ほとんどの場合、背景シーンの単一のフレームがあれば十分である。ユーザーの助けにより、最初のいくつかのフレームが静止していないときにアクティブトレーニングモードを用いることができる。
【発明の効果】
【０００８】
本方法は、汎用トレーニング例を適応させてシーン固有のオブジェクト検出器を提供す

るのに用いることができる。これによって、シーンにおいてデータ収集のコストのかかる動作を伴うことなく、特定のシーンにおける迅速な配備が可能になる。インクリメンタルトレーニングを用いて、識別器は、利用可能な汎用トレーニング例の利点を、シーン固有の例と同様に結合することができる。

【図面の簡単な説明】

【０００９】

【図１Ａ】本発明の実施の形態による、オブジェクトを検出するための識別器が適応される特定のシーンのビデオフレームである。

【図１Ｂ】適応されていない識別器が人オブジェクトを含むと識別した窓を含むビデオフレームである。

【図２】本発明の実施の形態による、識別器を特定のシーンに適応させるための方法の流れ図である。

【図３】本発明の実施の形態による、適応された識別器が人オブジェクトを含むと識別した窓を含むビデオフレームである。

【発明を実施するための形態】

【００１０】

図１Ａは、本発明の実施の形態に従って人が検出されることになる特定のシーンを示している。本発明の目的は、汎用的にトレーニングされた識別子を、汎用識別器が汎用トレーニングデータを用いてトレーニングされたときには未知であった特定のシーン内のオブジェクトを検出するように適応させることである。

【００１１】

図１Ｂに示すように、テストビデオのフレーム１０２毎に、それぞれ水平方向及び垂直方向の５０×３０ピクセルの重複を有する７０×５０ピクセルのスライド窓１０１を用いる。窓は、ラスター走査順でフレームを横切る。勾配ヒストグラム（ＨＯＧ：Ｈｉｓｔｏｇｒａｍ　Ｏｆ　Ｇｒａｄｉｅｎｔ）特徴が窓毎に抽出され、特徴ベクトルが構築される。特徴ベクトルは、トレーニングされた識別器、たとえば汎用サポートベクターマシン（ＳＶＭ：Ｓｕｐｐｏｒｔ　Ｖｅｃｔｏｒ　Ｍａｃｈｉｎｅ）に渡され、特定のシーン内の人が検出される。マルチクラス識別器等の他の識別器も用いることができることに留意されたい。

【００１２】

図１Ｂに示される各窓が、正の識別器出力を示す。図１Ｂは、主にフレームの左上部分における紛らわしいテクスチャに起因する過度に多数の誤検出を示している。このため、汎用的にトレーニングされたオブジェクト識別器は、良好に一般化されず、トレーニング例の詳細に大きく依存する。

【００１３】

通常、特定のシーン内の背景は、トレーニング中未知であるか又は利用可能でない。したがって、背景の部分は多くの場合に、特定のシーンから取得されたフレームにおいて人であると誤って分類される。

【００１４】

他方で、人１００を含む窓が、図１Ｂのフレーム内で正しく検出されていることも見て取ることができる。したがって、識別器は、検出問題のいくつかの局面、特に人の外観を正しく捉えている。

【００１５】

トレーニングされた識別器の部分的な正確性に動機付けされ、本発明者らの目的は、識別器を特定のシーンに効率的かつ迅速に、すなわちユーザー入力をほとんど又は全く伴うことなく適応させることである。

【００１６】

目標は、以前のトレーニング例の情報性のある局面を保持する一方で、特定のシーンのための分類タスクに関するより多くの情報も集め、それによって、汎用識別器からシーン固有の識別器を構築することである。

【００１７】
人検出の用途に焦点を置く。これは、ほとんどの監視用途において重要である。しかしながら、本発明者らの方法は、他の検出及びオブジェクト追跡タスクにも適用することができる。概して、本発明者らの方法は、トレーニングのための新たな例を選択し、古い情報性のない例を除去することにより、インクリメンタル更新を実行することによって機能する。情報性のない例を除去することによって、固定サイズのトレーニングデータセットを維持することが可能になるので、トレーニングが効率的であり、固定メモリ及びリアルタイム処理要件と共に機能することができる。
【００１８】
方法
図２は、本方法のステップをより詳細に示している。本方法のステップは、当該技術分野において既知のメモリ及び入力／出力インターフェースを備えるプロセッサにおいて実行することができる。
【００１９】
特定のシーン２０１のビデオ２２１が、カメラ２２０によって取得される。シーンモデルのモデル２１０が、ビデオを用いて構築される（２２２）。識別器２３０がモデルに適用され、負の例が選択される（２３１）。
【００２０】
最初に、識別器２３０は、汎用の、たとえばサポートベクターマシン（ＳＶＭ）、カスケード識別器、又はマルチクラス識別器である。識別器は、シーンに固有の識別器となるように適応され、この識別器を用いて、特定のシーン内の人等のオブジェクトを検出することができる。このオブジェクトは、国立情報学自動制御研究所（ＩＮＲＩＡ：Ｉｎｓｔｉｔｕｔ Ｎａｔｉｏｎａｌ ｄｅ Ｒｅｃｈｅｒｃｈｅ ｅｎ Ｉｎｆｏｒｍａｔｉｑｕｅ ｅｔ ｅｎ Ａｕｔｏｍａｔｉｑｕｅ）人物データ、又はマサチューセッツ工科大学（ＭＩＴ：Ｍａｓｓａｃｈｕｓｅｔｔｓ Ｉｎｓｔｉｔｕｔｅ ｏｆ Ｔｅｃｈｎｏｌｏｇｙ）の生物学及びコンピューター学習センター（ＣＢＣＬ：Ｃｅｎｔｅｒ ｆｏｒ Ｂｉｏｌｏｇｉｃａｌ ＆ Ｃｏｍｐｕｔａｔｉｏｎａｌ Ｌｅａｒｎｉｎｇ）の歩行者データセット等の既知の汎用データセットからの汎用トレーニングフレーム例の大きなセットを用いて最初にトレーニングされたときは未知であった。双方のデータセットが、人オブジェクト検出及び同様の用途のために識別器をトレーニングするのにコンピュータービジョンコミュニティにおいて広く用いられる、ラベル付けされたトレーニングフレーム及びラベル付けされていないテストフレームの大きなセットを含む。しかしながら、オブジェクトが存在する特定のシーンは、汎用識別器がトレーニングされたときに未知である。
【００２１】
新たな負の例のサブセットがトレーニングデータに加えられる一方、不確実性基準に基づいて、トレーニングデータから既存の負の例の別のセットが除去され、新たなトレーニングデータ２４０が生成される。同様に、正の例をトレーニングデータに付加する。これによってデータセットが固定サイズに維持される。
【００２２】
次に、識別器は新たなトレーニングデータ２４０を用いて再トレーニングされる（２５０）。選択するステップ、付加するステップ、及びトレーニングするステップは、所望の精度レベルに達するまで反復される。
【００２３】
代替的な実施の形態では、動きセンサー２６０を用いて特定のシーン内の動きを検出する（２７０）ことができ、この検出は正の例の選択２５１をトリガーする。動きが検出されると、フレームは、動きフレームとしてマーキングされる。フレーム差分は、動きの検出前、検出中、及び検出後にフレームに適用される。最も大きな差分値を有するフレーム内の領域が求められ、オブジェクトサイズ窓がグループ化されたピクセルに適合される。ここで、窓は新たな正の例である。

【００２４】
半教師付きモードにおいて、ユーザーがトレーニング中に参加し、オプションのユーザー入力データを提供する。次に、本方法は、フレーム窓を示してこの窓がオブジェクトを含むか否かをクエリする等の、ユーザーに対して行われるいくつかのクエリに基づいて特定のシーンに適応する。このモードは、人の外観が大幅に異なる場合があるか、又は空の（動きのない）フレームが自律的適応に利用可能でない、より困難な環境に用いることができる。
【００２５】
自律モードは、汎用データセット内の汎用データ、及び動きを一切含まない特定のシーン（単なる背景）のビデオからの最初のいくつかのフレームを用いて、本発明者らのシーン固有の識別器２３０をトレーニングする。このモードでは、最初のいくつかの空のフレーム、たとえば１つ又は２つを、自動背景除去に用いることができる。
【００２６】
ループ内のユーザーを用いた適応
アクティブトレーニング
アクティブトレーニング、その後に続く本発明者らのアクティブ選択方法の短い概観を与える。アクティブトレーニングにおける基本的な着想は、ユーザーに「情報性のある例」をクエリし、それによって受動方法、すなわちより少ないトレーニング例を用いるよりも高速にトレーニングを行うことである。アクティブトレーニングは、複数のコンピュータービジョンアプリケーションにおいて利用されてきた。たとえば、米国特許第７，５９３，９３４号及び同第７，５８７，０６４号を参照されたい。
【００２７】
アクティブ選択プロセスは通例反復的であり、プロシージャはユーザーに、選択されたラベル付けされていない例に対するラベルをクエリし、ユーザーフィードバックを取得し、ここでラベル付けされた例をトレーニングセットに付加する。識別器は各反復中に再トレーニングされ（２５０）、所望の精度レベルに達するか、トレーニングデータがこれ以上利用可能でなくなるまでプロセスが反復される。
【００２８】
インテリジェントなクエリ選択を通じて、アクティブトレーニングは、汎用識別器を非常に少ないトレーニング例を用いてトレーニングすることができる。アクティブトレーニングの最も重大な局面は、クエリ選択メカニズムである。未来の分類率に関してラベル付けされていない例の潜在的な情報性を基準することは、クエリ選択の場合と同様に困難である。
【００２９】
ほとんどの方法は、不確実性サンプリング、すなわち現在の識別器が最も不確実である例又は換言すれば最も不確実な例を選択すること等の代用物を用いる。たとえば、ＳＶＭ識別器２３０について、分類境界に最も近い例は不確実であり、ラベル付けされている場合、潜在的に情報性のあるものとなり得る。不確実性サンプリングに焦点を置く。
【００３０】
インクリメンタルトレーニング及び忘却
このセクションでは、インクリメンタルトレーニングのためのアクティブトレーニング及び忘却を利用する。主な着想は、汎用ラベル付けされたトレーニングフレームを所与とすると、トレーニングセットに付加するために、配備中のシーンから新たな情報性のあるフレームをユーザーにクエリすることができる一方、古い情報性のないフレームを除去することができるということである。選択（付加）及び削除（忘却）プロセスは、共にアクティブ選択を通じて機能する。削除の場合、アクティブ選択基準は逆にされる。すなわち、最も情報性のない例が選択される。
【００３１】
本発明者らの知る限り、これは、アクティブ忘却を用いると共にアクティブ忘却をインクリメンタル識別器トレーニングのためのアクティブトレーニングと組み合わせる最初の

研究である。

【００３２】

図２に示されるように、配備用の特定のシーン２０１が汎用ラベル付けされたトレーニングデータと共に与えられると、本方法は、ユーザーにクエリし、新たなフレームからいくつかのトレーニング例フレームを選択及び付加する。トレーニングデータを用いて識別器を特定のシーンに適応させる。

【００３３】

同時に、古い情報性のないデータがトレーニングセットから除去され、このため固定サイズであることが要求されるメモリが維持され、リアルタイム処理が可能になる。除去される例がアクティブに選択されるので、それらは比較的情報性がなく、除去によって精度が大幅に減少することはない。

【００３４】

このプロセスは、反復して実行され、その結果、汎用トレーニングデータを少量のユーザー入力を用いて適応させることによって達成された、シーン固有のトレーニングされた識別器となる。通常、特定のシーンにおいて、ビデオの最初のいくつかのフレーム、たとえば１つ又は２つは、更新を実行するのに用いることができ、そして結果としての識別器を特定のシーンに配備することができる。

【００３５】

不確実性ベースの選択基準

本発明者らが利用する選択基準は、ＳＶＭ識別器の超平面への距離に基づく。特に、ＳＶＭがトレーニングされた後、ＳＶＭを用いて、ラベル付けされていないフレームのクラスメンバーシップ確率値を推定する。以下で確率推定技法の短い概観を与える。

【００３６】

マージンに基づく確率推定

マージンからクラスメンバーシップ確率の推定値を得るために、プラットの逐次最小最適化（ＳＭＯ：Ｓｅｑｕｅｎｔｉａｌ　Ｍｉｎｉｍａｌ　Ｏｐｔｉｍｉｚａｔｉｏｎ）手順の変更版を用いて（米国特許第７，１１７，１８５号を参照されたい）、ＳＶＭから確率出力を抽出する。基本的な着想は、シグモイド関数を用いてクラス確率を概算することである。

【００３７】

本発明者らの特徴ベクトルはx_iであり、$y_i \in 2\{-1, 1\}$はベクトルの対応するラベルであり、ｆ（ｘ）がＳＶＭの決定関数である。クラスメンバーシップの条件付き確率Ｐ（ｙ＝１｜ｘ）は、次式（１）を用いて概算することができる。

【００３８】

【数１】

$$p(y=1|x)=\frac{1}{1+\exp(Af(x)+B)} \qquad (1)$$

【００３９】

ここで、Ａ及びＢは、最大尤度技法を用いて推定されたパラメーターである。

【００４０】

ラベル付けされたトレーニングデータのセットは、任意の時点においてＬである。ｘを、そのアクティブ選択基準（不確実性スコア）が対象とするラベル付けされていない例の特徴ベクトルとする。ｙを、選択中未知である、ｘの真のラベルとする。

【００４１】

選択基準を、２つのクラスに関して推定された確率間の差｜Ｐ（ｙ＝１｜Ｌ）－Ｐ（ｙ＝０｜Ｌ）｜として定義する。このため、大きなプールＡからのアクティブな例選択は、次式（２）のように定式化することができる。

【００４２】

【数２】

$$x^* = \underset{x_i \in A}{\operatorname{argmin}} \quad |P(y_i = 1|\mathcal{L}) - P(y_i = 0|\mathcal{L})| \qquad (2)$$

【００４３】

上記のスコアは、ラベル付けされていない例の場合の識別器の不確実性を表している。スコアが低いほど不確実性が高く（マージンがより小さい）、例は現在の識別器を更新する可能性がより高い。上記と同じ不確実性スコアを用いて、識別器境界から最も離れていることを示す最も高いスコアを有する例を除去することができる。

【００４４】

ＳＶＭ識別器の場合、これらの例は、ベクトルをサポートしない。このため、例を除去しても識別器の精度が変化しない。新たな例を付加することによって、除去される例が潜在的なサポートベクトルとなる場合があることに留意されたい。しかしながら、実際は、これは極度に稀にしか発生しないことを観測している。したがって、この基準を用いた例の除去は識別器の精度を減少させない。

【００４５】

二値分類の場合、マージンへの距離で十分である。しかしながら、推定確率値を用いて、上記の方法をマルチクラス識別器にも拡張することができる。ｋクラス問題の場合の選択基準は、次式（３）のとおりである。

【００４６】

【数３】

$$x^* = \underset{x_i \in A}{\operatorname{argmin}} \quad |P(y_{k_1}|\mathcal{L}) - P(y_{k_2}|\mathcal{L})|, \text{ where} \qquad (3)$$

$$k_1 = \underset{i=1:k}{\operatorname{argmax}} P(y_i), \; k_2 = \underset{i=1:k, i \neq k_1}{\operatorname{argmax}} P(y_i).$$

【００４７】

本発明者らの方法は、他の検出技法に取って代わることを意図しているのではなく、インクリメンタルアクティブトレーニングを追加することによって、他の検出技法を補うことを意図している。したがって、本発明者らの方法は、人検出アプリケーションにおいて良好な性能を与えることで知られている、識別器カスケード等の、特定のドメインにおいて良好に機能する他の既知の技法と共に用いることができる。

【００４８】

上記の半教師付き適応方法は、トレーニング条件とテスト条件が概ね異なり、他の情報が利用可能でない場合であっても、多くのインクリメンタルトレーニングタスクに適用することができる。

【００４９】

多くの人検出アプリケーションにおいて、より多くの情報が利用可能である。たとえば、特定のシーンにおいて、特定のシーン内に人が一切いないビデオのいくつかのフレーム（すなわち、この特定のシーンは、本質的に静止背景である）にアクセスすることができる場合がある。

【００５０】

代替的に、動きセンサーは、監視環境において多くの場合に利用可能である。動きセンサーは、人のいないフレーム（すなわち特定のシーンが概ね静止している）の存在を示すプライマリセンサーとして用いることができる。動きセンサーが動きを検出すると、正のサンプルを選択することができる。この実施の形態では、汎用識別器を、以下のように完全に自律的に特定のシーンに適応させることができる。

【００５１】

自律的適応

図１Ｂの例において、多数の誤検出が存在する。誤ったサンプルを根絶する一方、正しい検出をそのままにしておくことを目的とする。特定のシーン内に人が存在しないビデオフレームにアクセスすることができる場合、そのフレームからのフレーム窓を用いて、より多くの負のトレーニング例を集めることができる。

【００５２】

負の例の選択

フレームあたりのスライディング窓の数は、小さな窓サイズ及び大幅な重複に起因して非常に大きくなり得る。したがって、トレーニングセットのサイズ及び再トレーニング時間の双方の視点から、全ての窓を負のトレーニング例として用いることは実際的でない。

【００５３】

このセクションでは、例の選択、付加、及び除去の本発明者らの方法を説明する。汎用識別器２３０は、空のフレーム、すなわち人のいないフレームに適用され、識別器が正の応答を与える全ての窓がトレーニング用に選択される。

【００５４】

フレームは空であることが分かっているので、正の検出は、本質的に識別器による誤分類である。したがって、正の検出をトレーニングデータに付加することによって汎用識別器がシーン固有の識別器に変化すると共に、誤検出の数を低減する可能性が高い。

【００５５】

本発明の実施の形態は、ベイズ背景更新メカニズムを用いて特定のシーンの背景を推定すること、及び混合モデルのセットを各ピクセルに適合させて、最も有望なピクセルモデルを選択することによって、特定のシーンのモデルを構築する。この背景から、オブジェクトサイズにされた窓が選択される。

【００５６】

代替的に、ビデオからのフレームのセットに関して、差分が小さいピクセルをグループ化することによって（すなわち、グループ化されたピクセルが概ね静止した特定のシーンの部分を表す）、フレーム内のピクセル間の差分が求められる。次に、オブジェクトサイズ窓がグループ化されたピクセルに適合される。双方の場合に、窓は動きを表現しないので、そのような窓は新たな負の例に対応し、この窓は、動いているオブジェクトを一切含まない可能性が非常に高い。

【００５７】

新たな正の例を得るために、動きセンサーを用いて動きを有する動きフレームを検出することができる。このとき、フレーム差分は、動きの検出前、検出中、及び検出後のフレームにしか適用されない。そのようなフレーム差分マップにおいて、最大の差分値を有する領域が動いているオブジェクトを示し、このため新たな正の例を示す。

【００５８】

トレーニングセットサイズの維持

他方で、新たなトレーニング例を付加することによって、トレーニングデータセットのサイズが増加する。これは、メモリが制約された用途、及び処理レートがたとえばリアルタイムの人検出のために重要である場合において望ましくない。したがって、等しい数の古い負の例の、汎用トレーニング例からの除去も行う。これは、前のセクションの方法を用いることによって、すなわち境界から最も遠い例を除去することによって達成される。

【００５９】

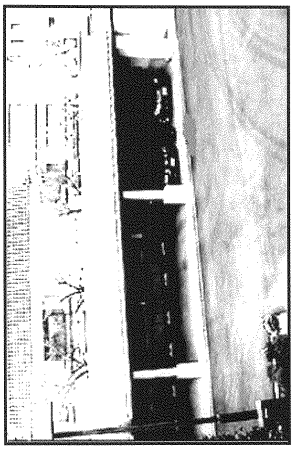
（人）オブジェクト検出器のための汎用識別器を特定のシーンに適応させるための完全に自律的なモードを提供する。また、ユーザーが識別器を再トレーニングするための正の例及び負の例をクエリされる半自律的モードを提供する。図３は、本発明者らの識別器が、歩行者３０１を含む窓を正確に識別するように適用されるビデオフレームを示している。

【００６０】

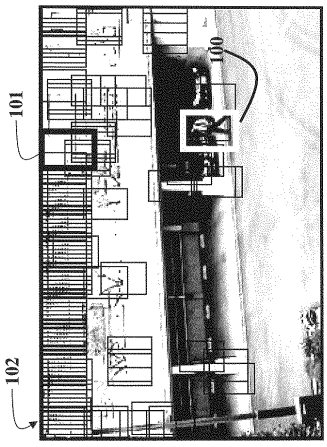
本方法は、汎用トレーニング例を適応させてシーン固有のオブジェクト検出器を提供するのに用いることができる。これによって、シーンにおいてデータ収集のコストのかかる

動作を伴うことなく、特定のシーンにおける迅速な配備が可能になる。インクリメンタルトレーニングを用いて、識別器は、利用可能な汎用トレーニング例の利点を、シーン固有の例と同様に結合することができる。

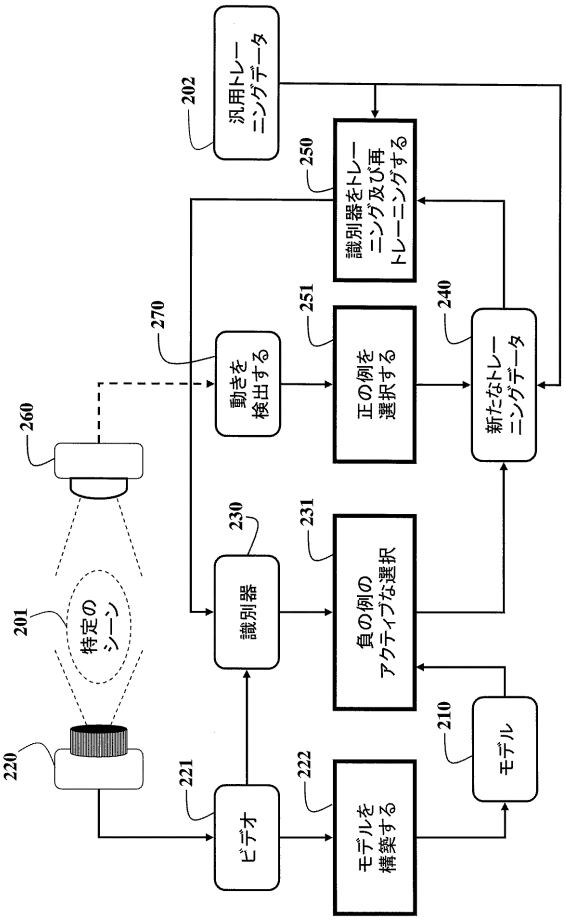
【図１Ａ】



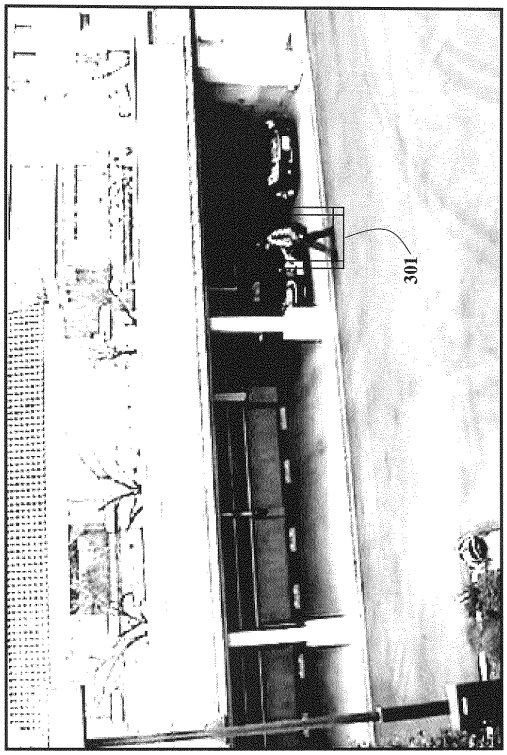
【図１Ｂ】



【図２】


汎用トレーニングデータ
202
識別器をトレーニング及び再トレーニングする
250
動きを検出する
270
正の例を選択する
251
新たなトレーニングデータ
240
260
識別器
230
負の例のアクティブな選択
231
特定のシーン
201
モデル
210
ビデオ
221
モデルを構築する
222
220

【図３】


301

フロントページの続き

(74)代理人 100122437
弁理士 大宅 一宏
(74)代理人 100147566
弁理士 上田 俊一
(74)代理人 100161171
弁理士 吉田 潤一郎
(74)代理人 100161115
弁理士 飯野 智史
(72)発明者 ファティー・エム・ポリクリ
アメリカ合衆国、マサチューセッツ州、ウォータータウン、マウント・オーバーン・ストリート７１２

Ｆターム(参考) 5L096 AA06 BA08 BA18 CA04 FA35 GA02 GA08 GA17 GA51 HA03
JA03 JA09 JA11 KA04

【外国語明細書】

Title of Invention

System and Method for Adapting Classifier to Detect Object in Particular Scene

Field of the Invention

This invention relates generally to computer vision, and more particularly to adapting a general object classifier to a particular scene to detect moving objects, specifically humans.

Background of the Invention

Most training methods for detecting or classifying objects in video frames are trained by providing labeled example frames of a video. After the classifier is trained, known test frames can be processed to determine a performance accuracy of the classifier.

Problems to be solved by the Invention

Such methods perform well when training and testing is done in similar conditions, such as on the same scene. However, conditions often change because training and deployment can be in different scenes with widely varying illumination, camera position, apparent object sizes, and pose of the object. That is, often it can not be determined beforehand to what types of scene the classifier will be applied.

It is object of the invention to adapt a general classifier to a particular scene, which is a particular scene that was unknown or not available when the classifier was trained.

Summary of the Invention

In many computer vision tasks, scene changes hinder the ability of generically trained classifiers. For example, a classifier trained for human detector with one set of frames is unlikely to perform well in different scene conditions.

Therefore, the embodiments of the invention provide an incremental training method and system for (human) object detection that can take generic training data and adapt a classifier to a different particular scene.

In an autonomous mode, the first few frames of a video, while there is no human present in the particular scene are used for the adaptation, i.e., the particular scene is substantially stationary. In most cases, a single frame of the background scene is sufficient. With the assistance of a user, an active training mode can be used when the first few frames are not stationary.

Effect of the Invention

The method can be used to adapt generic training examples to provide scene-specific object detectors. This enables a quick deployment in particular scenes, without involving expensive operations of data collection at the scene.

Using incremental training, the classifiers can combine the advantages of available generic training example as well as scene-specific examples.

Brief Description of the Drawings

Fig. 1A is a video frame of a particular scene to which a classifier for detecting object is to be adapted according to embodiments of the invention;

Fig. 1B is a video frame including windows that an unadapted classifier identified as containing human objects;

Fig. 2 is a flow diagram of a method for adapting the classifier to the particular scene according to embodiments of the invention;

Fig. 3 is a video frame including windows that an adapted classifier identified as containing human objects according to embodiments of the invention.

Detailed Description of the Embodiments

Figure 1A shows a particular scene in which a human is to be detected according to embodiments of our invention. It is an object of our invention to adapt a generically trained classifier to detect objects in the particular scene, which was unknown when the generic classifier was trained with generic training data.

As shown in Fig. 1B, we use a sliding window 101 of 75 x 50 pixels, with a horizontal and vertical overlap of 50 x 30 pixels, respectively, for each frame 102 of a test video. The window is passed over the frame in a raster scan order. Histogram of gradient (HOG) features are extracted for each window to construct a feature vector. The feature vector is passed to a trained classifier, e.g., a generic support vector machine (SVM), to detect humans in the particular scene. It should be noted that other classifiers, such as multi-class classifiers, can also be used.

Each window that is shown in the Fig. 1B indicates a positive classifier output. Fig. 1B shows an extremely large number of false positive detections, primarily due to misleading texture in the upper left part of the frame. Thus, generically trained object classifiers do not generalize well, and heavily rely on the specifics of the training examples.

Typically, the background in the particular scene is not known or unavailable during training. Consequently, parts of the background are often wrongly classified as being human in a frame acquired of the particular scene.

On the other hand, we can also see that the window including the human 100 is detected correctly in the frame in Fig. 1B. Therefore the classifier correctly captures some aspects of the detection problem, specifically, the appearance of the human.

Motivated by the partial correctness of the trained classifier, our objective is to adapt the classifier to the particular scene efficiently and quickly, i.e., with little or no user input.

The goal is to retain informative aspects of previous training example, while also gathering more information about the classification task for the particular scene, thereby constructing a scene-specific classifier from a generic classifier.

We focus on the application of human detection, which is important in most surveillance applications. However, our method can also be applied to other detection and object tracking tasks. Broadly, our method works by performing incremental updates by selecting new examples for training and removing old uninformative examples. The removal of the uninformative examples enables us to maintain a training dataset of a fixed size, so training is efficient, and can work with fixed memory and real-time processing requirements.

Method

Figure 2 shows the steps of the method in greater detail. The steps of the method can be performed in a processor including memory and input/output interfaces as known in the art.

A video 221 of a particular scene 201 is acquired by a camera 220. A model 210 of the scene model is constructed 222 using the video. The classifier 230 is applied to the model to select 231 negative examples.

Initially, the classifier 230 is generic, e.g., a support vector machine (SVM), a cascaded classifier, or a multi-class classifier. The classifier is

adapted to be a scene specific classifier, which can be used to detect object, such as humans, in a particular scene, which was unknown when the classifier was initially trained using a large set of generic example training frames from well known generic datasets, such as the Institut National de Recherche en Informatique et en Automatique (INRIA) person data, or the Center for Biological & Computational Learning (CBCL) at Massachusetts Institute of Technology (MIT) pedestrian dataset. Both data sets include a large set of labeled training frames, and unlabeled test frames, that are extensively used in the computer vision community to train classifiers for human object detection, and similar applications. However, the particular scenes in which the objects reside is unknown when the generic classifier was trained.

A subset of the new negative examples are added to the training data while removing another set of existing negative examples from the training data based on an uncertainty measure to produce new training data 240. Similarly, positive examples are added to the training data. This maintains the data set at a fixed size.

Then, the classifier is retrained 250 with the new training data 240. The selecting, adding, and training steps are repeated until a desired accuracy level is reached.

In an alternative embodiment, a motion sensor 260 can be used to detect motion 270 in the particular scene, which triggers the selection 251 of the positive examples. When motion is detected, the frames are marked as motion frames. Frame differencing is applied to the frames before, while and

after the motion is detected. Regions in the frames that have largest difference values are determined and object size windows are fitted to the grouped pixels where the windows are the new positive examples.

In a semi-supervised mode, a user participates during the training to provide optional user input data. Then, the method adapts to the particular scene based on a few queries made to the user, such as showing a frame window and querying whether the window includes an object, or not. This mode can be used for more challenging environments where human appearance may differ significantly, or where empty (motion free) frames are not available for autonomous adaptation.

An autonomous mode uses the generic data in the generic data set and the first few frames from the video of the particular scene, which does not contain any motion – just background, to train our scene-specific classifier 230. In this mode, the first few empty frames, e.g., one or two, can be used for automatic background subtraction.

Adaptation with User in the Loop

Active Training

We give a short overview of active training, followed by our active selection method. The basic idea in active training is to query the user for "informative examples," so as to train faster than passive methods, i.e., with fewer training examples. Active training has been employed in a number of computer vision applications, see e.g., U.S. Patents 7,593,934, and 7,587,064.

The active selection process is usually iterative, wherein the procedure queries the user for a label on selected unlabeled examples, obtains user feedback, and appends the now labeled example to the training set. The classifiers are retrained 250 during each repetition, and the process is repeated until a desired accuracy level is reached, or until no more training data are available.

Through intelligent query selection, active training can train a generic classifier with very few training examples. The most crucial aspect in active training is the query selection mechanism. Measuring the potential informativeness, in terms of future classification rate, of unlabeled examples is difficult, as is the case for query selection.

Most methods use proxies such as uncertainty sampling, i.e., selecting examples for which the current classifier is most uncertain, or in other words the most uncertain examples. For example, for the SVM classifier 230, examples closest to the classification boundary are uncertain and can be potentially informative if labeled. We focus on uncertainty sampling.

Incremental Training and Forgetting

In this section, we employ active *training* and *forgetting* for incremental training. The main idea is that given a set of generic labeled training frames, new informative frames from the scene of deployment can be queried to the user for adding to the training set, while old uninformative frames can be removed. The selection (adding) and deletion (forgetting)

processes both work through active selection. For deletion, the active selection measure is inverted, i.e., examples which are least informative are selected.

To our knowledge, this is the first work that employs active forgetting, and combines active forgetting with active training for incremental classifier training.

As shown in Figure 2, given the particular scene 201 for deployment, along with the generic labeled training data, the method queries, the user selects and adds a few training examples frames from the new frame. The training data are used to adapt the classifier to the particular scene.

At the same time, old uninformative data are removed from the training set, thus maintaining the memory required at a fixed size, and enabling real-time processing. As the examples to be removed are selected actively, they are relatively uninformative and the removal does not significantly decrease accuracy.

This process is performed iteratively, and results in a trained classifier that is scene-specific, achieved by adapting the generic training data with a small amount of user input. In general, in particular scenes, the first few frames of video, e.g., one or two, can be used for performing the update, and the resulting classifier can then be deployed at the particular scene.

Uncertainty-Based Selection Measure

The selection measure we employ is based on distance to a hyperplane of the SVM classifier. In particular, after the SVM is trained, the SVM is used to estimate class membership probability values for the unlabeled frames. We give a brief overview of the probability estimation technique below.

Probability Estimation Based on Margins

In order to obtain estimates of the class membership probability from margins, we use a modified version of Platt's Sequential Minimal Optimization (SMO) procedure, see U.S. Patent 7,117,185, to extract probabilistic outputs from the SVM. The basic idea is to approximate the class probability using a sigmoid function.

Our feature vectors are x_i, $y_i \in 2$ {-1, 1} are corresponding labels for the vectors, and $f(x)$ is a decision function of the SVM. The conditional probability of class membership $P(y = 1|x)$ can be approximated using

$$p(y = 1|x) = \frac{1}{1 + \exp(Af(x) + B)}, \qquad (1)$$

where A and B are parameters estimated using a maximum likelihood technique.

The set of labeled training data is $\mathcal{L}$ at any instant. Let x be the feature vector of the unlabeled example for which the active selection measure

(uncertainty score) is to be. Let y be the true label of x, which is unknown during selection.

We define the selection measure as a difference between the estimated probabilities for the two classes $|P(y = 1|\mathcal{L}) - P(y = 0|\mathcal{L})|$. Thus, active example selection from a large pool A can be formulated as

$$x^* = \underset{x_i \in A}{\operatorname{argmin}} \quad |P(y_i = 1|\mathcal{L}) - P(y_i = 0|\mathcal{L})| \qquad (2)$$

The above score represents the classifier uncertainty for an unlabeled example. The lower the score, the higher is the uncertainty (smaller margin), and the example is more likely to update the current classifier. We can use the same uncertainty score above, and remove examples having the highest score, indicating that they are farthest away from the classifier boundary.

For the SVM classifier, these examples are not support vectors. Hence, removing the example does not change the accuracy of the classifier. Note that adding new examples might make the removed examples potential support vectors. However, in practice, we observed that this happens extremely rarely. Consequently, example removal using this measure does not decrease the accuracy of the classifier.

For binary classification, the distance to the margin suffices. However, using estimated probability values, we can extend the above method to multi-class classifiers as well. The selection measure for a k-class problem is

$$x^* = \underset{x_i \in \mathcal{A}}{\operatorname{argmin}} \quad |P(y_{k_1}|\mathcal{L}) - P(y_{k_2}|\mathcal{L})|, \text{ where} \qquad (3)$$

$$k_1 = \underset{i=1:k}{\operatorname{argmax}} P(y_i), \; k_2 = \underset{i=1:k, i \neq k_1}{\operatorname{argmax}} P(y_i).$$

Our method is not intended to replace other detection techniques, but rather to complement them by adding *incremental active training*. As such, our method can be used with other known techniques that perform well in particular domains, such as classifier cascades, which are known to give good performance in human detection applications.

The above method of semi-supervised adaptation can be applied to many incremental training tasks, even when training and test conditions differ substantially, and no other information is available.

In many human detection applications, more information is available. For example, at the particular scene, we might have access to a few frames of video without any human in the particular scene, i.e., the particular scene is essentially stationary background.

Alternatively, motion sensors are often available in surveillance environments. The motion sensors can be used as a primary sensor to indicate the presence of a frame without a human, i.e., the particular scene is substantially stationary. When the motion sensor detects motion, positive samples can be selected. In this embodiment, we can adapt the generic classifier to the particular scene completely autonomously as follows.

Autonomous Adaptation

In the example of Figure 1B, there are a large number of false positives. We aim to eradicate false samples, while keeping the correct detection as is. If we have access to the video frames when there is no human in the particular scene, we can use the frame windows from that frame to gather more negative training examples.

Selecting Negative Examples

The number of sliding windows per frame can be very large, because of the small window size and substantial overlap. As such, it is impractical to use all of the windows as negative training examples, from both perspectives of training set size, and retraining time.

In this section, we describe our method of example selection and addition, and removal. The generic classifier 230 is applied to the empty frame, i.e., no human, and all the windows on which the classifier gives a positive response are selected for training.

As the frame is known to be empty, the positive detections are essentially misclassifications by the classifier. Therefore, adding the positive detections to the training data is likely to change the generic classifier to a scene specific classifier, and reduce the number of false positive detections.

The embodiments of the invention construct a model of a particular scene by estimating a background of the particular scene using a Bayesian

background update mechanism, and by fitting a set of mixture models to each pixel and selecting a most likely pixel model. From this background, object-sized windows are selected.

Alternatively, for a set of frames from video, differences between pixels in the frames are determined, by grouping the pixels that have small differences, i.e., the grouped pixels represent portions of the particular scene that are substantially stationary. Then, object size windows are fitted to the grouped pixels. In both cases the windows correspond to new negative examples as such windows depict no motion, and the windows are highly likely to not contain any moving objects.

To obtain new positive examples, the motion sensor can be used to detect motion frames with motion. The frame differencing can then only be applied frames before, while and after the motion is detected. In such frame difference maps, regions that have largest difference values indicate moving objects, and thus, new positive examples.

Maintaining Training Set Sizes

On the other hand, adding new training examples increases the size of the training data set. This is undesirable in memory-constrained applications, and where a processing rate is critical, e.g., for real-time human detection. Therefore, we also remove an equal number of old negative examples from the generic training examples. This is accomplished by using the method of the previous section, i.e., removing examples that are farthest away from the boundary.

We provide a completely autonomous mode for adapting a generic classifier for (human) object detector to a particular scene. We also provide a semi autonomous mode where a user is queried for positive and negative example to retrain the classifier. Figure 3 shows a video frame on which our classifier is applied to correctly identify the window that includes a pedestrian 301.

The method can be used to adapt generic training examples to provide scene-specific object detectors. This enables a quick deployment in particular scenes, without involving expensive operations of data collection at the scene. Using incremental training, the classifiers can combine the advantages of available generic training example as well as scene-specific examples.

1. A method for adapting a classifier to detect an object in a particular scene, wherein the particular scene was unknown when the classifier was trained with training data, comprising the steps of:

acquiring a video of frames of the particular scene with a camera;

constructing a model of the particular scene model using the frames in the video;

applying the classifier to the model to select negative examples;

adding a subset of the new negative examples to the training data while removing another set of existing negative examples from the training data based on an uncertainty measure;

adding selected positive examples to the training data;

retraining the classifier; and

repeating the adding and retraining steps until a desired accuracy level is reached to obtain a scene specific classifier.

2. The method of claim 1, wherein the constructing further comprises:

estimating a background of the particular scene using a Bayesian background update mechanism by fitting a set of mixture models to each pixel in the frames to produce a pixel model; and

selecting a most likely pixel model.]]

3. The method of claim 1, wherein the model is constructed from the frames by determining frame differences and grouping the pixels that have small difference values, and fitting object size windows to the grouped pixels where the windows are new negative examples.

4. The method of claim 3, further comprising:

detecting motion in the particular scene with a motion sensor;

applying the frame difference to the frames before, while, and after the motion is detected to determine difference values;

determining regions in the frames that have largest difference values; and

fitting object size windows to the grouped pixels where the windows are the new positive examples.

5. The method of claim 1, further comprising:

setting and maintaining the training data to a fixed size according to memory and real-time processing requirements.

6. The method of claim 1, further comprising:

adapting to changes in the particular scene by repeating the retraining if a difference between a current model and a current frame is large.

7. The method of claim 1, further comprising:

adapting a multi-class classifier using the new negatives example.

8. The method of claim 1, further comprising:

retraining the classifier with the new negative examples and the positive examples; and

adding the classifier as a cascade layer in the classifier.

9. The method of claim 1, wherein the training data are initially generic.

10. The method of claim3, wherein a size of the window is 75 x 50 pixels, with a horizontal and vertical overlap of 50 x 30 pixels.

11. The method of claim 1, further comprising:

extracting features from each frame into a feature vector; and

classifying the feature vector.

12. The method of claim 11, wherein the features are histogram of gradients.

13. The method of claim 1, wherein the classifier is a support vector machine.

14. The method of claim 1, wherein the classifier is a multi-class classifier.

15. The method of claim 1, wherein the classifier is initially generic.

16. The method of claim 1, wherein the object is human.

17. The method of claim 1, wherein a user selects selected unlabeled examples during active training.

18. A system for adapting a classifier to detect an object in a particular scene, wherein the particular scene was unknown when the classifier was trained with training data, comprising:

a camera configured to acquire a video of frames of the particular scene;

means for constructing a model of the particular scene model using the frames in the video;

means for applying the classifier to the model to select negative examples;

means for adding a subset of the new negative examples to the training data while removing another set of existing negative examples from the training data based on an uncertainty measure;

means for adding selected positive examples to the training data; and

means for retraining the classifier.

Abstract

A generic classifier is adapted to detect an object in a particular scene, wherein the particular scene was unknown when the classifier was trained with generic training data. A camera acquires a video of frames of the particular scene. A model of the particular scene model is constructed using the frames in the video. The classifier is applied to the model to select negative examples, and new negative examples are added to the training data while removing another set of existing negative examples from the training data based on an uncertainty measure;. Selected positive examples are also added to the training data and the classifier is retrained until a desired accuracy level is reached to obtain a scene specific classifier.

Representative Drawings

Figure 2

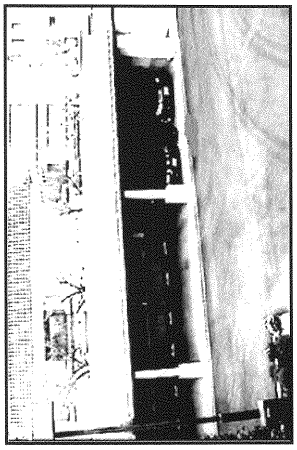


Fig. 1A

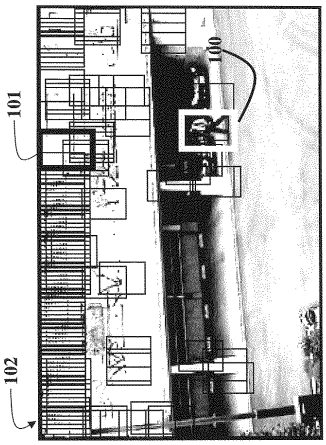


Fig. 1B

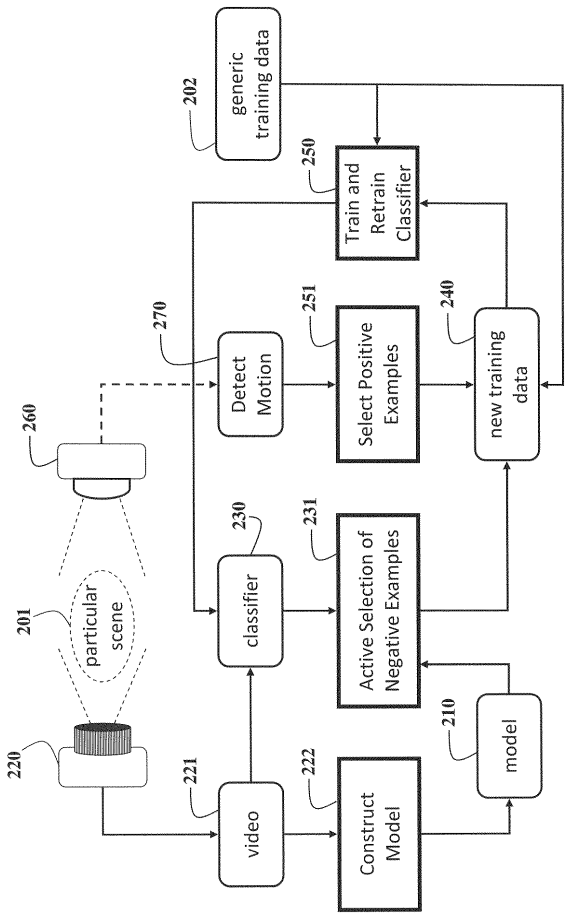


Fig. 2

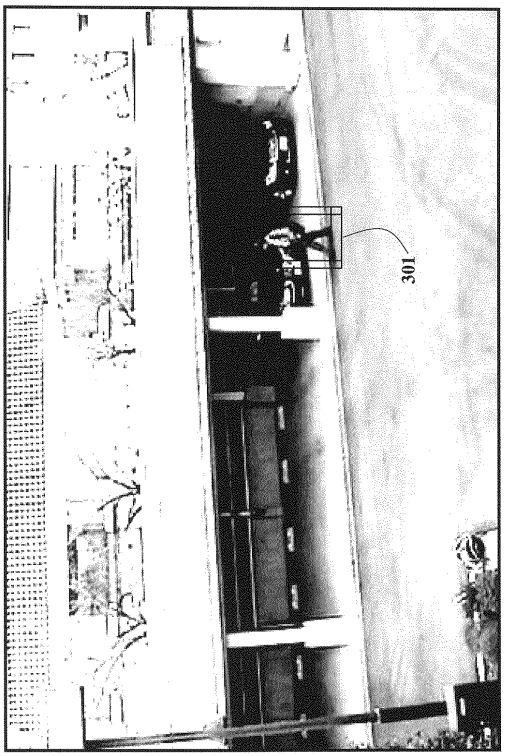


Fig. 3