構造的類似性をオートエンコーダーに適用した無監督欠陥セグメンテーションの改善

Paul Bergmann[1], Sindy Löwe[1,2], Michael Fauser[1], David Sattlegger[1], and Carsten Steger[1]

[1]*MVTec Software GmbH*
www.mvtec.com
{*bergmannp,fauser,sattlegger,steger*}*@mvtec.com*

[2]*University of Amsterdam*
sindy.lowe@student.uva.nl

概要—畳み込みオートエンコーダーは、画像データにおける非監督型欠陥セグメンテーションの人気の手法として台頭しています。最も一般的なアプローチは、`^{p} -距離に基づくピクセル単位の再構築誤差の閾値化です。しかし、この手順は、再構築にエッジ周辺のわずかな局所化誤差が含まれる場合、大きな残差を生じさせます。また、強度値が概ね一致している場合、視覚的に変更された欠陥領域を検出できません。私たちは、これらの問題がこれらのアプローチを複雑な現実世界のシナリオに適用することを妨げ、変分オートエンコーダーや特徴マッチングオートエンコーダーのようなより複雑なアーキテクチャを採用しても容易に回避できないことを示します。私たちは、輝度、コントラスト、構造情報を考慮し、単一のピクセル値を比較するのではなく、局所的な画像領域間の相互依存性を評価する構造的類似性に基づく知覚損失関数を使用することを提案します。この手法は、ナノファイバー材料の挑戦的な現実世界のデータセットと、2種類の織物からなる新規データセットにおいて、ピクセル単位の再構築誤差メトリクスを使用する最先端の無監督欠陥セグメンテーション手法と比較して、著しい性能向上を実現します。

1. 導入

産業製造において、高品質な生産とコスト効率の向上を迅速に実現するため、視覚検査は不可欠です。人間による手動検査は遅く、高コストで誤りやすいことから、完全自動化されたコンピュータビジョンシステムの活用がますます普及しています。監督学習手法では、システムが欠陥のあるサンプルとないサンプルの両方で訓練することで、欠陥領域のセグメンテーションを学習します。しかし、これらの手法はデータのannotationに大きな労力が必要であり、すべての欠陥タイプを事前に把握する必要があります。さらに、一部の製造プロセスでは、特にデータ依存度の高い深層学習モデルの場合、トレーニング用の十分な欠陥サンプルを収集するための廃棄率が低すぎる場合があります。

本研究では、視覚検査における非監督型欠陥セグメンテーションに焦点を当てています。目標は、非欠陥サンプルのみで訓練した後、画像内の欠陥領域をセグメント化することです。畳み込み神経ネットワーク（CNN）に基づくアーキテクチャ（例：オートエンコーダー（Goodfellow et al., 2016）や生成対抗ネットワーク（GANs；Goodfellow et al., 2014））がこのタスクに適用可能であることが示されています。第2章では、このような手法の概略を説明します。これらのモデルは、ボトルネックなどの制約下で入力を再構築しようとするため、高次元データ（例：画像）の本質を低次元空間で捉えることができます。テストデータにおける異常は、トレーニングデータのマニフォールドから逸脱すると仮定され、モデルはそれらを再現できません。その結果、大きな再構築誤差が欠陥を示します。通常、使用される誤差測定はピクセルごとの`^{p} -距離であり、これは単純さと速度の観点から任意に選択されたものです。しかし、これらの測定は、再構築がわずかに不正確な場所（例えば、エッジの局所的な不正確さによるもの）で高い残差を生じます。また、入力画像と再構築画像のピクセルの色値が概ね一致する場合、構造的な違いを検出できません。私たちは、このような方法が複雑な現実世界のシナリオで用いられる際に、その有用性が制限されることを示します。

上記の問題を緩和するため、私たちは構造的類似性（SSIM）メトリクス（Wang et al., 2004）を用いて再構築精度を測定することを提案します。SSIMは、エッジの配置に敏感ではなく、入力と再構築の間の目立つ違いに重点を置くように設計された距離測定指標です。これは、現在の最先端の無監督欠陥セグメンテーション手法（オートエンコーダーに基づくピクセル単位の損失関数を使用する手法）が無視する、局所的なピクセル領域間の相互依存関係を捕捉します。SSIMを損失関数として採用することで得られる性能向上を、2つの現実世界の産業検査データセットで評価し、ピクセル単位のアプローチに比べて顕著な性能向上を示します。図1は、ナノファイバー材料のNanoTWICEデータセット（Carrera et al., 2017）において、知覚損失関数がピクセル単位の`$^2$`損失よりも優れていることを示しています。両方のオートエンコーダーは欠陥領域の再構築を変化させますが、SSIMオートエンコーダーの残差マップのみがこれらの領域のセグメンテーションを可能にします。損失関数を変更しつつ、オートエンコーディングアーキテクチャを同じままに保つことで、他の最先端手法と肩を並べる性能を達成しました。

Improving Unsupervised Defect Segmentation by Applying Structural Similarity To Autoencoders

Paul Bergmann[1], Sindy Löwe[1,2], Michael Fauser[1], David Sattlegger[1], and Carsten Steger[1]

[1]*MVTec Software GmbH*
www.mvtec.com
{*bergmannp,fauser,sattlegger,steger*}*@mvtec.com*

[2]*University of Amsterdam*
sindy.lowe@student.uva.nl

***Abstract*—Convolutional autoencoders have emerged as popular methods for unsupervised defect segmentation on image data. Most commonly, this task is performed by thresholding a per-pixel reconstruction error based on an ℓ^p-distance. This procedure, however, leads to large residuals whenever the reconstruction includes slight localization inaccuracies around edges. It also fails to reveal defective regions that have been visually altered when intensity values stay roughly consistent. We show that these problems prevent these approaches from being applied to complex real-world scenarios and that they cannot be easily avoided by employing more elaborate architectures such as variational or feature matching autoencoders. We propose to use a perceptual loss function based on structural similarity that examines inter-dependencies between local image regions, taking into account luminance, contrast, and structural information, instead of simply comparing single pixel values. It achieves significant performance gains on a challenging real-world dataset of nanofibrous materials and a novel dataset of two woven fabrics over state-of-the-art approaches for unsupervised defect segmentation that use per-pixel reconstruction error metrics.**

1. INTRODUCTION

Visual inspection is essential in industrial manufacturing to ensure high production quality and high cost efficiency by quickly discarding defective parts. Since manual inspection by humans is slow, expensive, and error-prone, the use of fully automated computer vision systems is becoming increasingly popular. Supervised methods, where the system learns how to segment defective regions by training on both defective and non-defective samples, are commonly used. However, they involve a large effort to annotate data and all possible defect types need to be known beforehand. Furthermore, in some production processes, the scrap rate might be too small to produce a sufficient number of defective samples for training, especially for data-hungry deep learning models.

In this work, we focus on unsupervised defect segmentation for visual inspection. The goal is to segment defective regions in images after having trained exclusively on non-defective samples. It has been shown that architectures based on convolutional neural networks (CNNs) such as autoencoders (Goodfellow et al., 2016) or generative adversarial networks (GANs; Goodfellow et al., 2014) can be used for this task. We provide a brief overview of such methods in Section 2. These models try to reconstruct their inputs in the presence of certain constraints such as a bottleneck and thereby manage to capture the essence of high-dimensional data (e.g., images) in a lower-dimensional space. It is assumed that anomalies in the test data deviate from the training data manifold and the model is unable to reproduce them. As a result, large reconstruction errors indicate defects. Typically, the error measure that is employed is a per-pixel ℓ^p-distance, which is an ad-hoc choice made for the sake of simplicity and speed. However, these measures yield high residuals in locations where the reconstruction is only slightly inaccurate, e.g., due to small localization imprecisions of edges. They also fail to detect structural differences between the input and reconstructed images when the respective pixels' color values are roughly consistent. We show that this limits the usefulness of such methods when employed in complex real-world scenarios.

To alleviate the aforementioned problems, we propose to measure reconstruction accuracy using the structural similarity (SSIM) metric (Wang et al., 2004). SSIM is a distance measure designed to capture perceptual similarity that is less sensitive to edge alignment and gives importance to salient differences between input and reconstruction. It captures inter-dependencies between local pixel regions that are disregarded by the current state-of-the-art unsupervised defect segmentation methods based on autoencoders with per-pixel losses. We evaluate the performance gains obtained by employing SSIM as a loss function on two real-world industrial inspection datasets and demonstrate significant performance gains over per-pixel approaches. Figure 1 demonstrates the advantage of perceptual loss functions over a per-pixel ℓ^2-loss on the NanoTWICE dataset of nanofibrous materials (Carrera et al., 2017). While both autoencoders alter the reconstruction in defective regions, only the residual map of the SSIM autoencoder allows a segmentation of these areas. By changing the loss function and otherwise keeping the same autoencoding architecture, we reach a performance that is on par with other state-of-the-art