動物撮影支援のためのベストショット選定システム

岡部研究室 524C5009 高橋 卓

1　はじめに

撮影した大量の画像や映像の中から、カメラ目線かつ瞬きをしていないフレームを効率的に選定することは、写真撮影や映像制作の質を向上させるための重要な課題である。しかし、既存の視線推定や瞬き検出技術は主に人間を対象として設計・学習されており、動物への適用には多くの課題が残されている。例えば、動物の顔の形状や目の位置が人間と異なるため、視線方向の推定や瞬きの検出精度が大きく低下することが指摘されている。本研究では、視線推定と瞬き検出の技術を統合し、動物のビデオデータからベストショットを自動選定するシステムを構築することを目指す。このシステムでは、視線推定モデルを用いてよりカメラ目線であるフレームを評価し、瞬き検出モデルを活用して瞬きをしていないフレームを選別する。さらに、動物特有の特徴に対応するために、Unity を活用して動物に特化した視線ベクトル付きデータセットを作成し、それを用いて既存モデルを再学習させる。この新たなデータセットとモデルにより、動物撮影におけるベストショット選定の効率と精度を大幅に向上させることが期待される。

2　既存研究

2.1　視線推定

視線推定の一般的な手法として、動画を画像フレームに分割し、各フレームを深層学習モデルに入力する方法がある。この手法では、まず顔や目の特徴ベクトルを抽出し、それを基にカメラ基準の 3 次元視線ベクトルを推定する。顔全体や頭部の向き、目の状態など、複数の要素を組み合わせて視線方向を計算する点が特徴である。Y. Guan らの研究では、顔・頭部・目の動きを統合的に捉え、視線の時空間的な変化を考慮することで、視線推定を実現している【1】。

2.2　瞬き検出

瞬き検出は、顔画像や動画フレームから目の開閉状態を推定する技術であり、深層学習モデルを活用することで高精度な解析が可能となる。一般的な手法では、動画をフレームごとに分割し、各フレームを事前学習済みのモデルに入力して顔の特徴ベクトルを取得する。この特徴ベクトルを基に、各フレームにおける瞬きの発生確率を推定する。W. Zeng らは、未編集の野外動画において多人数の瞬きをリアルタイムで検出する手法を提案した【2】。この研究では、顔領域の目元の特徴に着目し、深層学習モデルを用いて目の開閉動作を解析している。

3　提案手法

本研究では、視線推定と瞬き検出を統合し、ビデオからカメラ目線かつ瞬きをしていないフレームを選定するアルゴリズムを提案する。このアルゴリズムでは、まず動画をフレーム単位に分割し、各フレームを解析可能な画像データとして準備する。次に、視線推定モデルを用いて各フレームの視線方向を推定する。出力された視線ベクトルは大きさが 1 に正規化された 3 次元ベクトルとして表現され、そのうちカメラ目線方向を示す Z 座標の値が大きいほど、よりカメラ目線であるフレームと定義する。このカメラ目線方向を示す Z 座標の値を基準に、よりカメラ目線のフレームを候補として抽出する。しかし、この候補フレームの中には瞬きが発生している可能性があるため、瞬き検出モデルを適用し、各フレームにおける瞬き発生確率を評価する。瞬きの確率があらかじめ設定した閾値未満であるフレームを「瞬きをしていないフレーム」と判定し、その中からよりカメラ目線であるフレームを最終的なベストショットとして選定する。このアルゴリズムにより、よりカメラ目線で、瞬きをしていない高品質なフレームを自動的に選定することが可能となる。

4　実験結果

本研究では、提案した視線推定と瞬き検出を統合したアルゴリズムを用いて、人間の動画データを入力し、その有効性を確認した。実験の結果、提案手法はカメラ目線で瞬きをしていないフレームを正確に選定することができ、ベストショットの自動選択が可能であることが示された。これにより、人間を対象とした場合、提案システムが高い精度で動作することを確認した。

図 1: 提案手法によるベストショットの選定結果

4.1　現状の課題

一方で、このモデルに動物の動画を入力すると、いくつかの課題が生じる。現在の視線推定と瞬き検出のモデルは、主に人間の画像データセットを用いて学習されており、動物の顔は人間とは形状が大きく異なるため、モデルが顔を正確に認識できない場合がある (図 2)。また、視線ベクトルの推定においても、動物の目の特徴や配置が人間と異なるため、不正確な視線ベクトルが出力されることが多く、精度が大幅に低下する (図 3)。このような課題により、動物に対して既存のモデルをそのまま適用することは困難であり、動物特化型のデータセットやモデルの再学習が必要不可欠である。

図 2: 顔の認識が難しい

図 3: 不正確な視線ベクトル

5　データセット作成

視線推定モデルの構築には、画像データとそれに対応する視線ベクトルのペアを含むデータセットが必要不可欠である。しかし、人間のデータセットとは異なり、動物の場合は視線ベクトルのデータセット作成が非常に困難である。人間の場合、被写体に特定の方向を向くよう指示することで、画像と正確な視線ベクトルを対応付けることが可能である。一方で、動物に同様の指示を与えることは現実的に不可能であり、視線方向のラベル付けが大きな課題となる。この課題を解決するために、本研究では Unity を活用した仮想環境内で動物の画像と視線ベクトルのデータセットを作成する方法を提案する。Unity は、3D モデルとカスタマイズ可能な視線ベクトルを容易に生成できるツールとして非常に有用である。

具体的には、以下の手順でデータセットを構築する。まず、動物の 3D モデルを仮想環境内に配置し、頭部の姿勢や視線方向を自由に制御する。この際、視線ベクトルの正確なラベル付けが可能であり、動物が特定の方向を向く条件を再現できる。また、照明条件、背景、撮影角度を多様に設定することで、現実環境に近いデータを生成する。さらに、データを大量に収集することで、さまざまな視線方向をカバーするデータセットを構築する。このデータセットは、視線推定モデルの再学習に使用され、動物への適用を可能にする。Unity を用いることで、動物の視線ベクトルデータセット作成における課題を克服し、高精度な視線推定を実現する基盤を提供する。

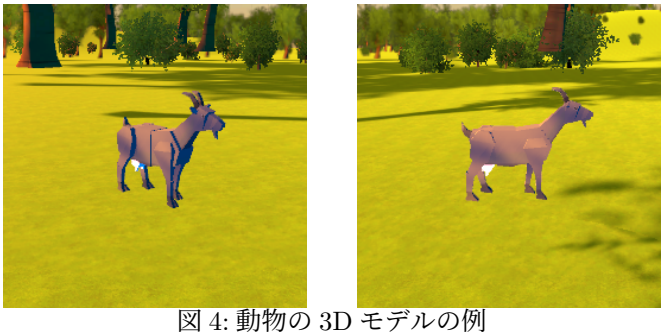

図 4: 動物の 3D モデルの例

6　今後の方針

6.1　仮想データの現実環境への適応

仮想空間で作成した画像データは、動物の視線推定モデルの学習において非常に重要な役割を果たす。このデータを活用することで、動物がさまざまな状況でどのように視線を動かすのかを再現でき、現実では収集が困難なデータセットを生成することが可能である。しかし、仮想空間で生成される画像データは、3D モデルをベースとしているため、被写体の質感や背景、光の反射といった要素が現実の写真とは異なる場合が多く存在する。このため、仮想データで学習したモデルが現実環境でも同様の精度を発揮するとは限らないという課題がある。この課題を解決するためには、仮想データと現実データの間に存在するギャップを埋める手法を導入する必要がある。その一つのアプローチとして、仮想データを現実に近づけるためのスタイル変換技術や、仮想データと現実データの特徴空間を一致させるための技術が挙げられる。具体的には、生成的敵対ネットワーク（GAN）を活用して、仮想空間で生成されたデータに現実的な質感や光の表現を付加することが可能である。この技術を用いることで、仮想データの見た目を現実の写真に近づけ、モデルの学習におけるギャップを効果的に軽減できる。さらに、スタイル変換技術を適用することで、仮想データと現実データが視覚的および特徴的に一致するように調整が可能である。これにより、仮想データと現実データの違いによって生じるモデルの精度低下を最小限に抑え、実際の撮影環境での高い適用性を実現できる。

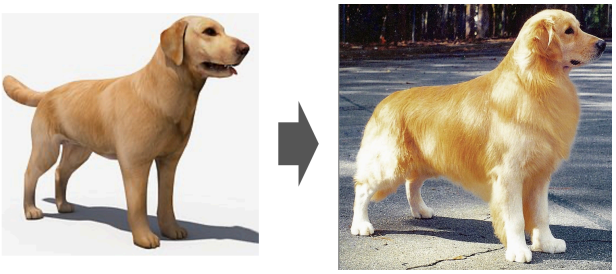

図 5:3D モデル画像から現実に近い画像への変換の例

6.2　リアルタイム処理の実現

現在の提案手法は、あらかじめ撮影された動画データを入力し、その中からカメラ目線かつ瞬きをしていないフレームを選定するバッチ処理を基本としている。この方法では、動画全体をフレームごとに分割し、それぞれのフレームについてカメラ目線や瞬きの有無を解析し、すべてのフレームを対象とした評価を行うことで、最適なフレームを選定する。しかし、実際の撮影現場においては即時的な判断とフィードバックが求められる場合が多く、リアルタイム処理を実現する必要がある。リアルタイム処理を実現するためには、従来のように動画全体を一括して解析し、最適なフレームを選定するバッチ処理の方法から、リアルタイム性を考慮したデータが読み込んだ時点で一つずつ順番に処理する方式へと移行する必要がある。そのためのアプローチとして、カメラ目線のスコアに閾値を設定し、その閾値を超えるフレームを「カメラ目線フレーム」として定義する。この方法では、フレーム全体のベストを選択するのではなく、フレームを読み込むたびにスコアが計算され、閾値を満たした段階で即座に判定を行うことが可能になる。また、瞬き検出についても、同様にリアルタイムで瞬きの発生確率を推定し、一定の閾値未満のフレームのみを「瞬きをしていないフレーム」としてフィルタリングする。このような逐次的な判定を組み合わせることで、撮影中にリアルタイムでカメラ目線かつ瞬きをしていないフレームを特定することができる。リアルタイム処理の実現により、撮影者はその場でベストショットを確認することが可能となり、撮影後の選定作業を大幅に削減することが期待される。

参考文献

[1] Y. Guan, Z. Chen, W. Zeng, Z. Cao, and Y. Xiao, ”End-to-end Video Gaze Estimation via Capturing Head-face-eye Spatial-temporal Interaction Context” Huazhong University of Science and Technology, 2023.

[2] W. Zeng, Y. Xiao, S. Wei, J. Gan, X. Zhang, Z. Cao, Z. Fang, and J. T. Zhou, ”Real-time Multi-person Eyeblink Detection in the Wild for Untrimmed Video”, 2023.