第 30 回日本バーチャルリアリティ学会大会論文集（2025 年 9 月）

顔への動的投影のためのずれ補償に向けた高精度な特徴点位置の動き予測

松下浩樹 [1)]，渡辺義浩 [1)]

1) 東京科学大学 工学院 情報通信系（〒 226-8501 神奈川県横浜市緑区長津田町 4259-G2-31, matsushita.h.6824@m.isct.ac.jp）

概要: 顔への Dynamic Projection Mapping では，実体と投影像のずれを知覚される問題がある．これは，撮像から投影までの遅延や特徴点検出の精度不足による小さな揺れ (ジッター) が原因である．そこで本稿では，遅延とジッターによる投影ずれを高速かつ高精度に補償するために，位置・加速度誤差を最小化する軽量なニューラルネットワークモデルを用いた特徴点位置の動き予測を提案する．

キーワード： 拡張・複合現実，計測・認識，プロジェクタ，動き予測

1. はじめに

Dynamic Projection Mapping (DPM) は，動的物体の形状や位置に合わせて投影を行うことで，対象の外観を変容させることができる技術である．DPM では，対象のセンシングから投影までの間に遅延が生じる．このような遅延により実体と投影像のずれが起こり，没入感が低下する．

DPM の中でも，顔を対象としたものを Dynamic Facial Projection Mapping (DFPM) と呼び，演劇やバーチャルメイクなどへの応用が可能である．DFPM においても，遅延による投影ずれの知覚で没入感が低下する問題がある．これに対して，Peng らは DFPM において知覚可能な遅延時間に関する調査を行った [1]．同研究では，最小で 3.87 ms の遅延が知覚されることが報告された．このように，DFPM においても遅延時間が短いことが重要である．

上記のような遅延時間の要求に対し，Bermano らは赤外線カメラを用いて顔の特徴点 (ランドマーク) を検出することで，物理的なマーカーを必要としない DFPM を提案した [2]．同手法では，システムの遅延時間を最小化するため，CPU/GPU を組み合わせたマルチスレッド計算が用いられており，遅延時間は 16.4 ms である．また同手法では，カルマンフィルタを用いたランドマーク位置予測が行われており，遅延を補償している．しかし，このような汎用的なフィルタリング手法による予測は線形であり，予測の精度は不十分である．

さらに Peng らは，高速なランドマーク検出と高精度である補助的なランドマーク検出を並列に実行し，時間的ずれを補正しながら 2 つの結果を組み合わせることで，高速かつ高精度な顔追跡処理を達成し，遅延による投影ずれを低減した [3]．同手法の遅延時間は，4.87 ms である．しかし，既存のシステムは遅延時間の要件である 3.87 ms を達成しておらず，観察者にずれを知覚される可能性がある．

また，投影対象の位置を把握するための顔のランドマーク検出において，検出手法の精度が不十分であり，検出位置のフレーム間での小さな揺れ (ジッター) により投影ずれが生じる．

このような遅延とジッターによる投影ずれを補償するための，DFPM に特化した予測に関する研究は行われていないが，DPM では関連する研究が行われている [4, 5]．しかし，これらの手法は非剛体運動の予測ができないことや，予測値の生成方法に関して，入力された値が正確であることを前提としている特性上，実際のジッターを含む状況において予測精度が低下する問題があると考えられる．

そこで本稿では，遅延とジッターによる投影ずれを補償するために，位置誤差と加速度誤差を最小化する軽量なニューラルネットワークモデルを用いた顔ランドマーク位置の動き予測を提案する．ここで，加速度誤差を最小化するのは，ジッターが加速度誤差として現れることが一般的に知られているためである．実験では，汎用的なフィルタリング手法による予測性能を上回ることを確認した．

2. 関連研究

2.1 投影ずれ補償のためのフィルタリング

DFPM では，遅延とジッターによる投影ずれを補償するために，汎用的なフィルタリング手法を用いている．具体的には，まずフィルタリングによりジッター抑制を行い，その後現在フレームの速度に応じて線形予測をしている．

ここで，動的なデータの位置や速度を推定するフィルタとして，カルマンフィルタがある．同手法は，状態空間モデルを事前に定義する必要があり，通常この状態空間モデルは経験的に決定することになる．そのため，適切なパラメータを設定することは難しい．

これに対して Casiez らは，実装が容易なフィルタとして 1€ フィルタを提案した [6]．同手法は，簡単にチューニングできる 2 つのパラメータの設定を行い，速度に応じたカットオフ周波数を適用することにより効果的にジッターを抑制している．このようにパラメータの設定が容易である 1€ フィルタは，カルマンフィルタよりも高い精度を得ることができる．しかし，顔ランドマークの動きは非線形であることが多いことや，顔ランドマーク特有の動きを反映でき

ないため，予測精度が不十分であると考えられる．

2.2 ニューラルネットワークによる動き予測

現在，顔ランドマーク位置の動き予測に特化した研究はないが，DPM において，ニューラルネットワークを用いた予測手法がある．例えば Maeda らは，LSTM と全結合層を用いた 4 層のニューラルネットワークを用いて，投影対象の位置と姿勢を予測した [4]．しかし，同手法は学習データと同じ形状の剛体に対する予測を行っており，学習データに含まれていない物体や非剛体運動に対する予測は試されていない．

これに対して，Yoshimura らは物体にマーカーを設置し，各時点での前のフレームからのマーカーの変化量を入力とし，出力も現在フレームの座標からの変化量とすることで未来の予測を行った [5]．しかし，変位量を入出力とする場合，検出されるマーカーの現在位置にジッターが含まれていると，誤った位置からの変位量を出力するため，予測精度が低下すると考えられる．

また DPM とは関連していないが，人体の姿勢に関する未来の予測を行っている研究が数多くある．しかし，未来の予測を行うモデルでは，入力にジッターが含まれている状況が考えられていない．そのため実際のアプリケーションに適用する際に，ジッターによる誤った情報が含まれた入力が与えられると，精度が低下する可能性がある．

またジッターを抑制することによる，人体の滑らかな姿勢推定に着目した研究も数多くある．中でも Zeng らは SmoothNet と呼ばれる軽量かつ高精度な姿勢推定が可能なニューラルネットワークを提案した [7]．同手法では，空間方向に対してモデルの重みを設けないことにより，ジッターを含む正確でない空間情報は一切モデル化させず，時間方向のみに重みを用いることで精度の向上を図っている．さらに，ジッター誤差が加速度誤差として現れることから，位置誤差のみでなく，加速度誤差を最小化することで効果的にジッターを抑制している．

本稿では，DFPM に予測モデルを適用する際，高速性が必要であることと，入力にジッターが含まれている状況に対応する必要があることから，SmoothNet を応用することで投影ずれ補償を行うことを目指す．ただし，SmoothNet は入力フレームのジッターを抑制するための手法であり，動作予測を行うための手法ではないため，そのままでは DFPM に適用し予測を行うことはできないことに注意が必要である．

3. 高速かつ高精度な顔ランドマーク位置の動き予測モデル

3.1 モデルアーキテクチャ

2.2 節で述べた通り，SmoothNet [7] を応用した投影ずれの補償を行う．ここで図 1 のように，全結合層と残差ブロックを用い，隣接するフレームに対して明確に重みをもたせることで時系列的関係を学習する．また，予測モデルに入力されたランドマーク座標から速度と加速度を離散微分により計算し，位置，速度，加速度を予測モデルが学習することにより精度を向上させる．

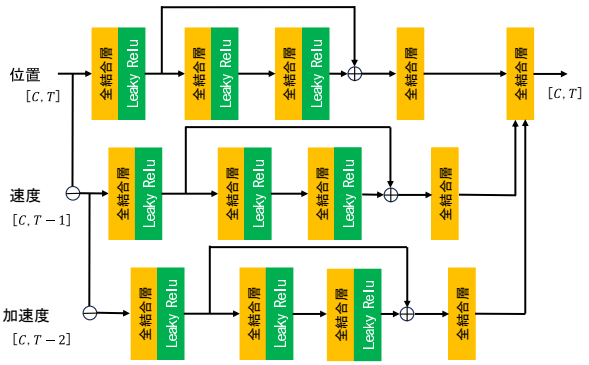

図 1: 予測モデルのアーキテクチャ．全結合層と残差ブロックからなり，入力されたランドマーク座標から離散微分によって速度と加速度を求め，図の上段，中段，下段それぞれで位置と速度と加速度を学習している．ここで，$C = L \times 2$ である．

未来の予測を行うため，学習時に予測フレーム数分の未来の座標を含む時系列データを真値とする．ここで，$\mathbf{p}_{i,j} = (x_{i,j}, y_{i,j})^T$ をフレーム j における L 点の顔ランドマーク座標のうち，i 番目の座標として，$\mathbf{P}_j = (\mathbf{p}_{0,j}^T, \mathbf{p}_{1,j}^T, ..., \mathbf{p}_{L,j}^T)^T$ をフレーム j におけるすべてのランドマーク座標とする．また，T を予測モデルへの入力ウィンドウサイズ，t を現在フレーム，M を予測フレーム数とする．このとき，予測モデルへの入力が

$$(\mathbf{P}_{t-(T-1)}, \mathbf{P}_{t-(T-2)}, ..., \mathbf{P}_t)^T \tag{1}$$

であれば出力は，

$$(\mathbf{P}_{t-(T-1)+M}, \mathbf{P}_{t-(T-2)+M}, ..., \mathbf{P}_{t+M})^T \tag{2}$$

となる．

3.2 損失関数

予測モデルの学習における損失関数は位置誤差と加速度誤差を最小化するために 2 つの要素からなり，以下の式で表される．

$$\mathcal{L} = \mathcal{L}_{pos} + \mathcal{L}_{acc} \tag{3}$$

ここで，$\mathcal{L}_{pos}$ は位置に関する損失であり，$\mathcal{L}_{acc}$ は加速度に関する損失である．

位置に関する損失 $\mathcal{L}_{pos}$ と加速度に関する損失 $\mathcal{L}_{acc}$ について，SmoothNet [7] では，L1 損失を用いていた．しかし本手法では，投影ずれが知覚されないことを目指している．大きな誤差はずれの知覚につながるため，最大誤差を小さくする L2 損失を用いる．以下に，位置損失 $\mathcal{L}_{pos}$ の式を示す．

$$\mathcal{L}_{pos} = \frac{1}{T \times L} \sum_{j=t-(T-1)+M}^{t+M} \sum_{i=1}^{L} \|\mathbf{G}_{i,j} - \mathbf{p}_{i,j}\|_2^2 \tag{4}$$

また，加速度損失 $\mathcal{L}_{acc}$ は以下の式で表される．

$$\mathcal{L}_{acc} = \frac{1}{(T-2) \times L} \sum_{j=t-(T-1)+M}^{t+M} \sum_{i=1}^{L} \|\mathbf{G}''_{i,j} - \mathbf{p}''_{i,j}\|_2^2 \tag{5}$$

表 1: 本手法と 1€ フィルタを用いた手法の位置および加速度の平均誤差の比較

手法	位置誤差 [pix]	加速度誤差 [pix/frame^2]
本手法	0.3392	0.2787
1€ フィルタ	0.9735	0.7137

ここで，$\mathbf{G}_{i,j}$ は j フレーム，i 番目の顔ランドマークにおける真値の座標であり，$\mathbf{p}_{i,j}$ は同様に，j フレーム，i 番目の顔ランドマークにおける予測値の座標である．$\mathbf{G}^{''}_{i,j}$ は真値における加速度を表し，$\mathbf{p}^{''}_{i,j}$ は予測値における加速度を表す．予測値の加速度の計算は以下の式によって行われる．

$$\mathbf{p}^{''}_{i,t} = \mathbf{p}_{i,t+1} - 2\mathbf{p}_{i,t} + \mathbf{p}_{i,t-1} \tag{6}$$

また，$\mathbf{G}''_{i,t}$ に関しても同様である．

4. 予測モデルの学習と評価

4.1 実験条件

本稿では，Miao らの顔画像データセットを用いた [8]．同データセットは 108 人分の顔画像データを 400 fps，720 × 540 pix で撮像している．1 人につき 10 種類のモーションが含まれており，1 つのモーションにつき 1000 枚の画像がある．顔ランドマークデータセットは，同手法の顔画像に対し LDEQ によるランドマーク検出を行うことで作成した [9]．ここで，顔ランドマーク点数は $L = 98$ である．真値データにはアノテーションしたデータにガウシアンフィルタを適用し，可能な限りジッターをなくしたものを使用し，入力データには同フィルタを適用する前のジッターが含まれたデータを用いた．

本稿では，学習には 48 人分のデータを, 評価には 10 人分のデータを用いた．また，CPU は Intel® Xeon® Gold 6136，GPU は NVIDIA GeForce RTX 2080 Ti を用いて実験を行った．

図 1 に示すモデルアーキテクチャのうち，1 層目の全結合層の出力次元数を 196，残差ブロックの全結合層の出力次元数を 98 とした．また，評価は位置誤差と加速度誤差によって行った．

さらに，1€ フィルタ [6] を用いて顔ランドマーク座標をフィルタリングし，その後各ランドマーク座標において速度による線形予測をした手法と比較をした．今回使用したフィルタのパラメータは，最小カットオフ周波数が $f_{c_{min}} = 0.35$，カットオフスロープが $\beta = 1.006$ である．

ここで，誤差の要求精度を，Peng らの遅延知覚に関する調査 [1] を参考に決定する．同手法では最小で 1.41 mm の投影像と実体のずれが知覚できることが報告されている．本稿で用いたデータセットでは，目の動きが小さく，正面を向いているモーションにおける画像上の瞳孔間距離が約 93 pix であり，実寸の瞳孔間距離は平均 60-70 mm であることから，1 pix あたりの距離は約 0.65-0.75 mm と計算される．したがって，投影ずれ知覚の最小距離と合わせて，データセット画像上における要求精度は約 2 pix となる．

4.2 4 フレーム先の予測

本節では予測フレーム数を $M = 4$，すなわち 10 ms 先の予測に関する結果を示す．ここで，入力ウィンドウサイズを $T = 15$ として，モデルの学習を行う．表 1 に，学習させたモデルによる予測結果を示す．1€ フィルタを用いた従来手法と比較して，本手法は位置誤差，加速度誤差ともに優れていることがわかる．また，推論時間は $L = 98$ で 1.2 ms と高速であることを確認した．

さらに，予測されたランドマーク座標と真値を顔画像に重ね合わせた結果を図 2 に示す．この図より，本手法の方がずれが小さいことがわかる．

ここで図 3 に，あるモーションでの各顔の部位における最大誤差を示す．表 1 のように精度の向上が見られた一方，図のように，目標精度の 2 pix よりも誤差が大きい場合も確認できた．

4.3 予測フレーム数に関する検証

本節では，予測フレーム数を長くしたときの最大誤差に関する検証を行う．予測フレーム数は小さいほうから順に，$M = 4, 8, 12, 20, 30, 50$ である．図 4 に，学習させたモデルによる予測結果を示す．この図より，どの予測フレーム数においても従来手法に対して，本手法が優位になることがわかる．

5. まとめ

本稿では，DFPM における投影ずれを補償するための，位置誤差と加速度誤差を最小化する軽量なニューラルネットワークモデルによる顔ランドマーク位置の動き予測を提案した．実験では，4 フレーム先の予測に関して従来手法との比較を行い，位置誤差，加速度誤差ともに従来手法よりも優れた性能であることを確認した．また，長い予測フレーム数に対しても，本手法は従来手法よりも優れていることを確認した．これらの結果より，本手法は従来手法と比較してより投影ずれを抑えた DFPM を行うことが可能であると考えられる．

一方，目標精度に達していない部分が見られた．現状のモデルでは，すべての顔ランドマーク点に対して，同じモデルパラメータを用いている．しかし実際には顔の部位ごとに動きの特性が異なるため，それぞれに異なるモデルパラメータを用いて，部位ごとに最適な重みとなるようにすることで精度を向上させられる可能性がある．

参考文献

[1] Hao-Lun Peng, Shin'ya Nishida, and Yoshihiro Watanabe. Studying user perceptible misalignment in simulated dynamic facial projection mapping. In *2023 IEEE International Symposium on Mixed and Augmented Reality*, pp. 493–502, 2023.

[2] Amit H Bermano, Markus Billeter, Daisuke Iwai, and Anselm Grundhöfer. Makeup lamps: Live augmen-

Frame 1　Frame 21　Frame 41　Frame 61　Frame 81

図 2: 顔ランドマーク座標の予測値 (赤) と真値 (緑) を顔画像に重ね合わせた結果を 20 フレームごとに並べたもの．上段は 1€ フィルタの結果，下段は本手法の結果である．本手法の方は従来手法と比較して，ずれが小さいことがわかる．

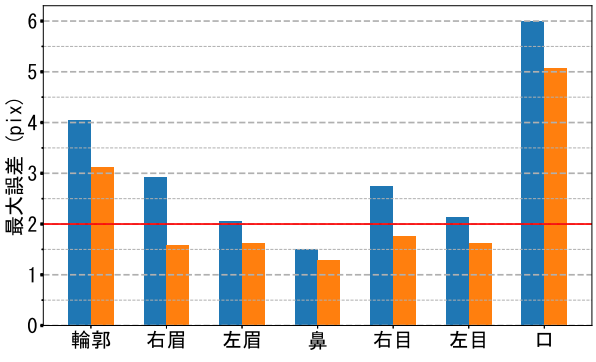

図 3: 口を大きく開けるモーションの最大誤差．青色の棒グラフが 1€ フィルタを用いた手法，橙色のグラフが本手法による予測である．赤色の線が目標精度の 2 pix であり，口の最大誤差が目標精度よりも顕著に大きくなっている．

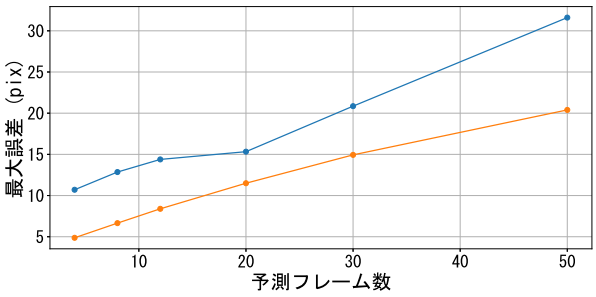

図 4: 予測フレーム数による最大誤差の遷移．青色のグラフが 1€ フィルタを用いた手法，橙色のグラフが本手法による予測である．どの予測フレーム数においても本手法の方が優れていることがわかる．

tation of human faces via projection. In *Computer Graphics Forum*, Vol. 36, pp. 311–323, 2017.

[3] Hao-Lun Peng, Kengo Sato, Soran Nakagawa, and Yoshihiro Watanabe. Perceptually-Aligned Dynamic Facial Projection Mapping by High-Speed Face-Tracking Method and Lens-Shift Co-Axial Setup. *IEEE Transactions on Visualization and Computer Graphics*, 2025.

[4] Kosuke Maeda and Hideki Koike. MirAIProjection: Real-time projection onto high-speed objects by predicting their 3D position and pose using DNNs. In *Proceedings of the 2020 International Conference on Advanced Visual Interfaces*, pp. 1–5, 2020.

[5] 吉村一真, 橋本直己. プロジェクションマッピングにおける動き予測を用いた変形する非剛体への実時間光学的補正. 映像情報メディア学会誌, Vol. 78, No. 3, pp. 348–354, 2024.

[6] Géry Casiez, Nicolas Roussel, and Daniel Vogel. 1€ filter: a simple speed-based low-pass filter for noisy input in interactive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2527–2530, 2012.

[7] Ailing Zeng, Lei Yang, Xuan Ju, Jiefeng Li, Jianyi Wang, and Qiang Xu. Smoothnet: A plug-and-play network for refining human poses in videos. In *European Conference on Computer Vision*, pp. 625–642, 2022.

[8] Langning Miao, Ryo Kakimoto, Kaoru Ohishi, and Yoshihiro Watanabe. Improving Real-Time Near-Infrared Face Alignment With a Paired VIS-NIR Dataset and Data Augmentation Through Image-to-Image Translation. In *2024 IEEE International Conference on Image Processing*, pp. 2368–2374, 2024.

[9] Andrés Prados-Torreblanca, José M Buenaposada, and Luis Baumela. Shape Preserving Facial Landmarks with Graph Attention Networks. In *33rd British Machine Vision Conference*, 2022.