

1 枚の写真をインタラクティブ動画に変換する一手法

爲谷秀太^{†1} 佐藤隆^{†1}

概要: 本発表では、物体認識と動画生成 AI を用いて、1 枚の写真を直接操作可能なインタラクティブ動画に変換する方法を提案する。これにより実際では起こりえない超現実的なユーザ体験をもたらすことができる。本手法では、物体認識された領域と動きをユーザが指定すると、プロンプト文を与えて動画生成 AI により動画を生成し、指定した動きの折れ線型スライダを埋め込んだインタラクティブ動画コンテンツを自動編集する。今回はプロトタイプ動作検証を発表する。プロンプトによる生成動画 AI の不確実性の克服が課題である。

1. はじめに

著者の 1 名はかつて、動画再生の制御に用いられる時間スライダの形状を折れ線型に拡張し、動画フレーム上に重畳して、あたかも被写体を直接つかんで動かしているかのように動画再生する技術を提案した[1-4]。この技術の事業化において、次のような提供価値があることを確認した。

- 直接触るので被写体を身近に感じる。商品を直接手にとらせて購買意欲をかきたてるのと似た効果がある
- ユーザはスライダを往復して繰返し再生する傾向があり、動画の尺が短くてもユーザの滞留時間が長くなる
- 危険な動物など、直接的に触りにくい被写体でも仮想的に触ることができる
- コントロールできない対象をコントロールできる支配欲を満たす

このうち最後の支配欲を満たす効果は、実際にはありえないことを仮想体験できる楽しさを提供する。例えば、時間を遡って動かせることは、時間の支配欲を満たし、何度も再生と巻き戻しを繰り返す行動を引き起こす。また、偉い人や配偶者など、ふだんは思い通りに動いてくれない対象が、このインタフェースを使って思い通りに動いてくれるのは、極上の楽しさを提供するであろう。

デモを多数こなしたが、事前に撮影した動画を再生しているのではなく、写真が動画として合成されていると誤解を受けることがしばしばあった。当時の技術では、不可能であったが、最近、1 枚の静止画から数秒の動画を生成する生成 AI 技術が使えるようになったので、この組み合わせを試してみたのが本発表の主旨である。

2. 提案手法

本発表では、以下の手順で 1 枚の静止画を直接操作可能なインタラクティブ動画にする方法を提案する。図 1 に提案手法の処理の流れを示す。

1. ユーザが写真を 1 枚アップロードする
2. システムが物体認識し被写体の領域と種類を求めてユーザに提示する

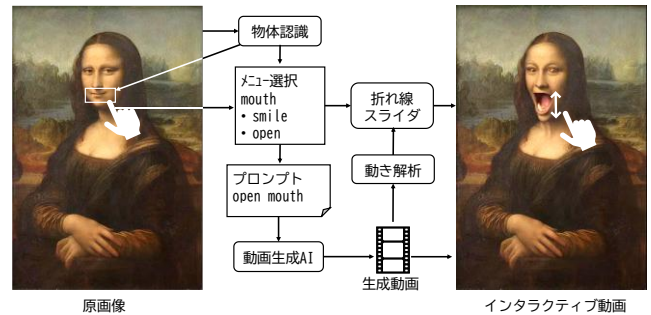


図 1 提案手法の処理の流れ

3. ユーザが領域を指定し、物体の種類に対応した動きを選択する
4. システムが被写体と動きに対応したプロンプトを用いて動画生成する
5. システムが 2 で指定した領域近辺の動きベクトルを求めて折れ線スライダにする

物体認識では、写真に含まれる、顔、目や口などの顔のランドマーク、車両、建物などの物体の種類と領域を認識する。物体認識には OpenCV や YOLO を用いた。

事前に物体の種類に応じて動きの項目を用意しておき、ユーザが認識された領域を触ると、動きメニューを表示して選択できるようにする。

物体の動きには事前にプロンプトを割り当てておき、ユーザが選択した物体と動きに対応するプロンプト文を動画生成 AI に与える。物体の動きに対応するプロンプトの例を表 1 に示す。

表 1 物体と動きとプロンプトの対応例

物体	動き	プロンプト
顔 (頭)	左右に動かす	move head left and right, shake head
	うなずく	move down, take a bow
口	開ける	open mouth
	微笑む	smile mouth
建物	動かす	move building
	回転する	rotate building
	伸ばす	grow building

^{†1} 東京電機大学

この他、画角や背景、物体の意図しない箇所に変更がおきないように奇妙な画像生成を抑制するプロンプトを補っている。補ったプロンプトは次の通りである。

ポジティブプロンプト： best quality, fixed camera

ネガティブプロンプト： low quality, deformed, deformed, distorted, fused fingers, bad anatomy, weird hand, ugly, zoom, tilt, pan

折れ線スライダ形状は、領域周辺の生成された動画の動きベクトルを集約して求めている。単純な動きなら、上記の動きに対応した単純なスライダ形状をあてはめても、操作性に影響は少ない。

生成動画と折れ線データは、それぞれ H.264 と JSON にエンコードして、JavaScript によるプレーヤで再生する。ブラウザが VideoDecoder に対応していれば再生可能であり、多種のスマートホンやタブレットで動作可能である。

図2に生成したインタラクティブ動画の例を示す。モナリザは、図1のように口を開ける加工の他、手を上げたり、顔を左右にふらせたりすることができた。ピサの斜塔は、傾きを戻すような大きな加工はできず、回転したり上下に伸縮したりする程度であった。

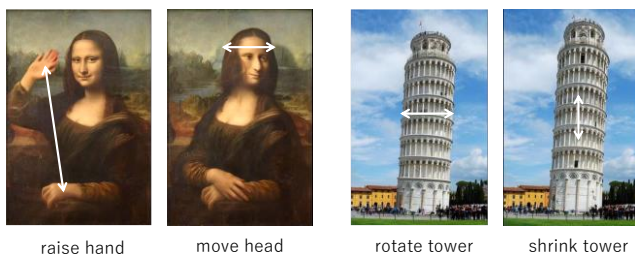


図2 生成したインタラクティブ動画の例

本システムのプロトタイプでは、動画生成 AI として、CogVideo[12,13]を用いた。生成に20分程度の時間がかかるのが欠点であるが、ローカル環境で動作するオープンソースの実装であるから採用した。クラウド上には高速・高品質なサービスがあり、ローカルで動作するものでも数十秒で生成可能な LTXV が発表された。最適な動画生成 AI の選定は、今後の課題である。

また、プロンプト文どおりに動画が生成されないことが多く、成功率は1割程度である。画角や背景の動きを抑制するプロンプトと、特定の部分だけを動かすプロンプトが拮抗関係にあることが影響していると考えられる。

3. 関連研究

動画の直接操作再生をもたらす手法は、様々なものが提案されている。

前述の[1-4]では編集ソフトを用いて折れ線スライダを埋め込む作業が必要であった。Kimber[5]や Dragicevic[6], Karrer[7]は、被写体の自動トラッキングにより、折れ線ス

ライダを自動作成する方法を提案している。物体の軌跡は折り返しが生じる場所では、ポインタの移動とスライダ上のツマミの移動が不連続になる問題がある。[1]ではポインタの位置とツマミの位置を幾何学的に滑らかに対応づける二等分線法を提案している。[6]では時空間上の距離を評価してポインタの位置を決めている。Nguyen[8]は、再生時間に応じてスライダを縦方向に展開することにより、この問題を解決している。

NTT アイティ「サイバー巧知」[9][10]では、ユーザが被写体の動きそうなところを触ると、その周辺の前後フレームから局所的に動きベクトルを計算して、動的に折れ線スライダを作成する。当時の Android タブレット端末の計算力が非力だったため、全フレームを対象にした自動トラッキングの代わりに、人間の「ここが動きそうだ」という知識を活用して計算量を削減した。

本研究では、1枚の静止画をユーザがどのように動かしたいかが先行し、それに合わせて動画を生成する。折れ線スライダの形状は簡素なもので済むことが多い。

なお、暦本のエッセイ[11]によれば、CHI 研究において、ダイレクトマニピュレーション（直接操作）の時代が一区切りつき、次は AI と何らかの「会話」をしながら目的を達成する方向になることが示唆されている。たしかに、何らかのタスクを達成する手段としての直接操作は一区切りついているのかもしれない。一方、被写体に仮想的に触るという直接操作が目的になる場面では、まだ開拓の余地があると考えられる。

4. まとめ

以上、本発表では、物体認識と動画生成 AI を用いて、1枚の写真を直接操作可能なインタラクティブ動画に変換する方法を提案した。今後の課題は、以下の通りである。

- より高速な動画生成 AI の検討
- 作成したコンテンツについての心理的効果測定
- テキストによるプロンプトを介してではなく、動画生成 AI の生成過程の中に特定の領域の動きを直接指示する方法
- 動画生成 AI の生成過程で、動きベクトルを出力する方法

参考文献

- [1] 佐藤隆, 阿久津明人, 外村佳伸. Coaster: 折れ線スライダによる時空間メディアインタフェース, 情処研報 HI-79, 1998, p.37-41.
- [2] 佐藤隆, 阿久津明人, 外村佳伸. Coaster: 時空間メディアのダイレクトマニピュレーション, インタラクシオン 99, 1A-21 1999
- [3] 佐藤隆, 阿久津明人, 南憲一, 外村佳伸. Coaster: 映像の時空間直観的操作による可変速再生方法とその応用. 情報処理学会論文誌. 1999, vol. 40, no. 2, p. 529-536.

- [4] 佐藤隆. 映像を簡単にインタラクティブ化する DRAGRI, 映像学会冬季大会, 2003.
- [5] Kimber D., et.al.. Trailblazing: Video playback control by direct object manipulation. ICME, 2007, p. 1015-1018
- [6] Dragicevic, P., et al.. Video Browsing by Direct Manipulation. CHI '08, 2008, p. 237-246.
- [7] Karrer, T., et.al.. DRAGON: A Direct Manipulation Interface for Frame-Accurate in-Scene Video Navigation. CHI '08, 2008, p. 247-250.
- [8] Nguyen, C. et.al.. Direct manipulation video navigation in 3D. CHI'13, 2013, p. 1169-1172.
- [9] 特許 5844697. 映像再生装置, 映像再生方法, および映像再生用プログラム, 2014
- [10] ICT 授業でニーズの高い動画再生・比較・手書き書込みが直観的にタッチ操作できる「サイバー巧知(コーチ)」をリリース <https://www.atpress.ne.jp/news/66861>
- [11] 暦本純一, CHI2024 を振り返って. <https://medium.com/blog-two/1a3bc3b60b1f>
- [12] Yang, Z. et.al.. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer, 2024, arXiv:2408.06072
- [13] Hong, W., CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers, 2022, arXiv:2205.15868