Hierarchical Video Generation from Orthogonal Information: Optical flow and Texture

Katsunori Ohnishi*§¹, Shohei Yamamoto*¹, Yoshitaka Ushiku¹, and Tatsuya Harada¹²

¹ The University of Tokyo  ² RIKEN

* indicates equal contribution.
§ currently belongs to DeNA Co., Ltd.

Paper&Slides: http://katsunoriohnishi.github.io/
Goal

- Video generation
  - Applications)
    Human AI collaboration, dataset extension

Generate videos → OK
Generative Adversarial Network (GAN)
[I. Goodfellow+, NIPS14]
Generative Adversarial Networks

- **Application**

  - **Image Translation**
    - [JY Zhu, et al., ICCV17]
  
  - **Text to Image**
    - [H. Zhang, et al., ICCV17]
  
  - **Dataset Extension**
    - [A. Shrivastava, et al., CVPR17]
  
  - **Domain Adaptation**
    - Y. Ganin, et al., ICML15
GANs for Video

- Previous works
  - Video GAN (VGAN) [C. Vondrick, et al., NIPS16]
  - Temporal GAN (TGAN) [M. Saito, et al., ICCV17]*

* We refer their first arXiv version in the paper because ICCV17 papers were not published yet at our submission time.
Challenges in video generation

Important factors for realistic video generation:
1. Realistic frame
2. Scene consistency
3. Reasonable motion
GANs for Video

- Video GAN (VGAN) [C. Vondrick, et al., NIPS16]

1. Realistic frame
2. Scene consistency
3. Reasonable motion
GANs for Video

- Temporal GAN (TGAN) [M. Saito, et al., ICCV17]*

1. Realistic frame
2. Scene consistency
3. Reasonable motion

* We refer their first arXiv version in the paper because ICCV17 papers were not published yet at our submission time.
Challenges in video generation

- Important factors for realistic video generation:
  1. Realistic frame
  2. Scene consistency
  3. Reasonable motion

It is important to consider structure of video and to make a video generation pipeline that can express the structure.
Hierarchical Video Generation

- Generating video via optical flow
  1. Generate optical flow as motion information
  2. Give texture to generated optical flow

Optical Flow Generator → Generated Optical Flow → Texture Generator → Generated Video

\[ z_{\text{flow}} \sim \mathcal{N}(0, 1) \]
\[ z_{\text{tex}} \sim \mathcal{N}(0, 1) \]
Features of Optical flow

- Extractable unsupervisedly
- Holding the contour of a moving object
- Continuity in the time direction
- No texture information

Generating optical flow first makes it …
- possible to generate a video with reasonable motion
- easier to generate a realistic video than without optical flow
Proposed Method

- Overview of generator
Proposed Method

- Optical flow generator

\[ z_{\text{flow}} \sim \mathcal{N}(0, 1) \]

\[ z_{\text{tex}} \sim \mathcal{N}(0, 1) \]
Optical flow generator

- Optical flow generator is constructed based on the pipeline of VGAN [C. Vondrick+, NIPS16].
Optical flow generator

- Background optical flow should be zero
  - If the camera is fixed.

Eliminate background stream!
Proposed Method

- Texture generator

\[ z_{\text{flow}} \sim \mathcal{N}(0, 1) \]

\[ z_{\text{tex}} \sim \mathcal{N}(0, 1) \]
Texture generator

- Generate RGB video from random noise and optical flow.
Texture Generator

- Auto-encoder that converts optical flow to RGB video
Texture Generator

- Add background stream as VGAN
- to obtain scene consistency
Texture Generator

- In order to keep the contour of input, we add U-net architecture [O. Ronneberger+, MICCAI15].
  cf.) Pix2Pix [P. Isola+, CVPR17]
Texture Generator

- The whole pipeline of our texture generator
Overview of Proposed Method

- Hierarchical video generation via optical flow

\[ z_{\text{flow}} \sim \mathcal{N}(0, 1) \]

[Diagram of optical flow generation process]

\[ z_{\text{tex}} \sim \mathcal{N}(0, 1) \]
Experiments

- Experiment 1:
  - Examples of generated results
  - Qualitative comparison with baseline
  - Human evaluation
- Experiment 2:
  - Walk in dual z
- Experiment 3:
  - Unsupervised action classification
Experiments

- Experiment 1:
  - Examples of generated results
  - Qualitative comparison with baseline
  - Human evaluation
- Experiment 2:
  - Walk in dual z
- Experiment 3:
  - Unsupervised action classification
Experiments

- **Dataset**
  - Resolution: 64x64
  - Time: 32frames (≈1~2 seconds)

Penn Action
[W. Zhang+, ICCV13]

Penn Action Cropped

SURREAL
[G. Varol+, CVPR17]
Experiments

- Examples of generated results
- Various videos

Penn Action

Penn Action Cropped

SURREAL
Experiments

- Qualitative comparison with VGAN
  - FTGAN generates a video with reasonable motion

[Images: VGAN vs FTGAN]

- A person is walking without moving his legs.
- A person is walking by moving their left and right feet in turn.
Experiments

- Qualitative comparison with VGAN
  - FTGAN generates a video with reasonable motion

  **VGAN**
  - The outline and the axis of rotation are unclear.

  **FTGAN** (ours)
  - The outline and the axis of rotation are clear.

Result on PennAction cropped

Pull-up
Experiments

- AB test on Amazon Mechanical Turk
  - Q: In which video is it easier to figure out what action is being performed?
  - 200 videos
  - 9 votes on each video

Number of videos with better evaluation

<table>
<thead>
<tr>
<th></th>
<th>Penn Action</th>
<th>Penn Action Cropped</th>
<th>SURREAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGAN</td>
<td>76</td>
<td>91</td>
<td>95</td>
</tr>
<tr>
<td>FTGAN (ours)</td>
<td><strong>124</strong></td>
<td><strong>109</strong></td>
<td><strong>105</strong></td>
</tr>
</tbody>
</table>

As the complexity of the dataset increases, the proposed method becomes effective.
Experiments

- Experiment 1:
  - Examples of generated results
  - Qualitative comparison with baseline
  - Human evaluation

- Experiment 2:
  - Walk in dual z

- Experiment 3:
  - Unsupervised action classification
Experiments

- Overview of generator

$z_{flow} \sim \mathcal{N}(0, 1)$ generates motion

$z_{tex} \sim \mathcal{N}(0, 1)$ generates appearance
Experiments

- **Walk in z**
  - Conforming mode-collapse avoiding
  - Mixing two images at any ratio

[A. Radford+, ICLR16]
Experiments

- Dataset
  - SURREAL [G. Varol+, CVPR17] cropped

Crop videos with bounding box
Experiments

- Walk in $z_{\text{flow}}$
  - The same motion
  - The appearance changes gradually
- Walk in $z_{\text{tex}}$
  - The same appearance
  - The motion changes gradually
Experiments

- Walk in $z_{\text{flow}}$
  - The same motion
  - The appearance changes gradually

- Walk in $z_{\text{tex}}$
  - The same appearance
  - The motion changes gradually
Experiments

- Walk in $z_{\text{flow}}$
  - The same motion
  - The appearance changes gradually

- Walk in $z_{\text{tex}}$
  - The same appearance
  - The motion changes gradually
Experiments

- Walk in z_flow
  - The same motion
  - The appearance changes gradually

- Walk in z_tex
  - The same appearance
  - The motion changes gradually

Our method can generate videos by independently controlling motion and appearance.
Experiments

- Experiment 1:
  - Examples of generated results
  - Qualitative comparison with baseline
  - Human evaluation
- Experiment 2:
  - Walk in dual z
- Experiment 3:
  - Unsupervised action classification
Experiments

Purpose
- Investigate the unsupervised feature expression learning capability as the same way with previous works.

Method
- Extract the last layer in discriminator as feature.

Setting
  - 101 classes
  - 13320 videos
- Classifier: SVM

All of these settings are following previous works (VGAN and TGAN).
Experiments

(Repost) Overview of our networks.

Experiments
Experiments

- Extract the last layer of discriminator as feature vector

- Optical Flow Discriminator
- Texture Discriminator

- SVM

Real Optical Flow
Real Video
Experiments

- Late fusion of flow-discriminator and texture-discriminator improves recognition accuracy.
- Features learned by each network is complementary, which means…
  - Flow-discriminator learns motion information
  - Texture-discriminator learns appearance information

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chance</td>
<td>0.9%</td>
</tr>
<tr>
<td>(a) Flow-discriminator + Linear SVM (ours)</td>
<td>48.0%</td>
</tr>
<tr>
<td>(b) Texture-discriminator + Linear SVM (ours)</td>
<td>50.3%</td>
</tr>
<tr>
<td>(a) + (b) FTGAN (fusion by Linear SVM) (ours)</td>
<td>59.7%</td>
</tr>
</tbody>
</table>
Experiments

- FTGAN outperforms VGAN and TGAN
- Separating information ensures the capture of much richer video characteristics

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chance</td>
<td>0.9%</td>
</tr>
<tr>
<td>VGAN + Random Init [C. Vondrick+, NIPS16]</td>
<td>36.7%</td>
</tr>
<tr>
<td>TGAN: Image-discriminator + Linear SVM [M. Saito et al, arXiv]</td>
<td>38.6%</td>
</tr>
<tr>
<td>TGAN: Temporal-discriminator + Linear SVM [M. Saito et al, arXiv]</td>
<td>23.3%</td>
</tr>
<tr>
<td>(a) Flow-discriminator + Linear SVM (ours)</td>
<td>48.0%</td>
</tr>
<tr>
<td>(b) Texture-discriminator + Linear SVM (ours)</td>
<td>50.3%</td>
</tr>
</tbody>
</table>
| (a) + (b) FTGAN (fusion by Linear SVM) (ours)        | 59.7%    | outperform!
Summary

- We propose a hierarchical video generative model via optical flow: FTGAN.

Experiments:

Table:

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGAN</td>
<td>36.7%</td>
</tr>
<tr>
<td>TGAN</td>
<td>38.6%</td>
</tr>
<tr>
<td>FTGAN</td>
<td>59.7%</td>
</tr>
</tbody>
</table>

UCF101

It is important to consider structure of video and to make a video generation pipeline that can express the structure.

Thank you!
Fin

終
Experiments

- Does $z_{\text{flow}}$ generates motion and $z_{\text{tex}}$ generates appearance independently?
  - vertical: generated from the same $z_{\text{flow}}$
  - Horizontal: generated from the same $z_{\text{tex}}$
Experiments

- Does $z_{flow}$ generates motion and $z_{tex}$ generates appearance independently?
  - vertical: generated from the same $z_{flow}$
  - Horizontal: generated from the same $z_{tex}$

The same movements and different appearance
Experiments

- Does $z_{\text{flow}}$ generates motion and $z_{\text{tex}}$ generates appearance independently?
  - vertical: generated from the same $z_{\text{flow}}$
  - Horizontal: generated from the same $z_{\text{tex}}$

Penn Action

- Cropped

SURREAL

Different movements and the same appearance
動画生成の難しさ

- 生成された動画が本物らしくあるためには、以下の3つの条件を満たしている必要がある
  a. 各フレームがきれいな画像になっている
  b. 動画内でのシーンの一貫性が保たれている
  c. 動きが妥当なものになっている

GANを動画生成に拡張した手法

Video GAN (VGAN)
[C. Vondrick, et al., NIPS16]

Temporal GAN (TGAN)
[M. Saito et al, arxiv]
関連研究 -動画生成-

生成された動画が本物らしくあるためには、以下の3つの条件を満たしている必要がある

a. 各フレームがきれいな画像になっている
b. 動画内でのシーンの一貫性が保たれている
c. 動きが妥当なものになっている

手法: VGAN [C. Vondrick, et al., NIPS16]

・ 動くものを前景、動かないものを背景として生成
・ 動画内で同じシーンが現れる
・ 3D convolutionを使用
・ 見た目と動きを同時に学習
関連研究 -動画生成-

生成された動画が本物らしくあるためには、以下の3つの条件を満たしている必要がある

a. 各フレームがきれいな画像になっている
b. 動画内でのシーンの一貫性が保たれている
c. 動きが妥当なものになっていない生成結果

手法: TGAN [M. Saito et al, arxiv]

・2D convolutionをXY方向にかけた後、1D convolutionをT方向に
  動き情報が見た目情報を抽象化された状態でしか取れていない
実験

生成結果: TextureGAN (Optical flowを与えた場合)

<table>
<thead>
<tr>
<th>Penn Action</th>
<th>SURREAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>GT optical flow</td>
<td>GT optical flow</td>
</tr>
<tr>
<td>Generated video</td>
<td>Generated video</td>
</tr>
<tr>
<td>GT video</td>
<td>GT video</td>
</tr>
</tbody>
</table>
実験

生成結果: FTGAN (Optical flowも生成)

Penn Action
- Generated optical flow
- Generated video

SURREAL
- Generated optical flow
- Generated video
実験

- VGANとの比較
  - Penn Action

<table>
<thead>
<tr>
<th>VGAN</th>
<th>TextureGAN</th>
<th>FTGAN</th>
</tr>
</thead>
</table>

![Image of Penn Action results for VGAN, TextureGAN, and FTGAN]
実験

- VGANとの比較
  - SURREAL

<table>
<thead>
<tr>
<th>VGAN</th>
<th>TextureGAN</th>
<th>FTGAN</th>
</tr>
</thead>
</table>

未完成