콜라주 기법에 의한 비디오 생성을 위한 탐색적 실험 분석

*조형래 **박구만
*서울과학기술대학교 일반대학원 미디어IT공학과
***서울과학기술대학교 전자미디어IT공학과
*artjow@naver.com

Exploratory Experiment Analysis for Video Generation by Collage Technique

*Hyeongrae Cho **Gooman Park
*Dept. of Media IT Engineering
**Dept. of Electronics and IT Media Engineering
Seoul National University of Science and Technology

요약

딥러닝이 정답을 찾아가는 연구과정이라면 미술은 정답이나 오답의 단정적 결과보다는 미추(아름다움과 추함)를 포함하는 과정적, 창조적 행위에 가깝다고 할 수 있다. 다시 말하면 미술은 0과 1로만 환원할 수 없는 세계를 기술하여 감동을 주는 유기적 규칙이 내재되어 있고 때로는 과학이 만들어낸 결론을 뒤집는 반상식적 추론을 하기도 한다. 그러므로 딥러닝은 예술적 방식을 통하여 과학의 상식적 추론과의 좋은 거리(Fine distance)를 유지할 필요성이 있는데, 이를 위해서 기존 딥러닝의 이미지 생성과 관련하여 Distance, Classification, Optimization 등의 문제를 미술 표현 기법과 목적이 담겨있는 창작자의 Statement 키워드와의 유사성과 차이점을 비교 분석할 필요가 있다고 생각한다. 시각적 표현과 관련된 딥러닝의 성능은 아직 사람의 표현능력에 못 미치고 있어 본 논문에서는 콜라주 기법에 의한 비디오 생성을 위한 탐색적 실험 분석을 목적으로 GAN을 활용한 콜라주 비디오를 제작하고 그 문제점과 개선점을 제안하고자 한다.

1. 서론

콜라주(collage)란 실밥 ·깡통 등 캔버스와는 전혀 이질적인 재료나, 잡지의 삽화 ·기사를 오려 붙여 보는 사람에게 이미지의 연쇄반응을 일으키게 하는 기법을 뜻한다. 상관관계가 없는 재료를 영상으로 표현하여 색다른 의미와 유머, 복고를 전달하기도 하고 사회풍자, 부조리와 냉소적인 충동, 초현실주의적 특징을 가지고 있다. 대표적인 작가로는 막스 에른스트 (Max Ernst), 쿠르트 슈비터스(Kurt Schwitters), 로버트 라우션버그(Robert Rauschenberg), 에두아르도 파올로치(Eduardo Luigi Paolozzi), Eugenia Loli (유지니아 롤리), Joseba Elorza(조주바엘루자)등이 있다. 인공지능 분야 중 GAN(Generative adversarial network)은 생성자와 판별자간의 적대적 학습으로 진짜 같은 가짜 이미지나 비디오를 생성하고 이를 판별자가 가짜를 진짜처럼 판별하게 되어 결국 생성자 입장에서 진짜 같은 가짜를 만들게 된다. 이러한 행위나 결과는 콜라주 기법과 유사성이 있다고 생각하는데 예를 들면 학습을 거쳐 존재하지 않는 사람의 얼굴을 생성하는 Style GAN2, 스케치와 실제이미지의 Pair를 판별하여 실제 같은 이미지를 만들어내는 pix2pix, 비디오를 한 장의 사진 이미지에 매핑 시켜 동작하는 Photo-2-Video, 비디오의 과거와 현재의 학습을 통해 미래를 예측하는 video-to-video등이 그러하다. 본 논문의 실험방법은 예측하는 생성이미지를 직접 그리거나 이미지 작업 한 것과 알고리즘을 활용하여 제작한 것을 비교분석하였고, GAN을 활용한 비디오 제작을 통해 문제점을 파악하고, 그 개선점도 함께 제안한다. pix2pix의 경우 학습데이터에 있을 법한 오브젝트를 스케치한 것과 약간 변형 왜곡시킨 스케치 결과를 테스트 하고, 이것을 콜라주 기법으로 그린 실제 그림과 비교함으로서 그 문제점과 개선방향을 제시한다. Photo-2-Video의 경우 합성할 비디오의 인체의 움직임을 학습데이터에 있는 것보다 동작을 크게 주고, 사진이미지가 아닌 그림이미지를 실험해 봄으로써 비디오 변화의 문제점을 발견하고 개선점도 제안한다. Style GAN2의 경우 사진이미지와 직접 그린 그림이미지와의 성능 변화를 살피고, LipGAN의 경우는 사운드와 사진이미지의 입모양 Shink가 잘 동작하는지를 파악하는 실험을 하였다. 이처럼 본 논문은 GAN을 활용하여 비디오를 만들 때 발생하는 문제점과 이를 보완하기 위한 방법을 제안하고 있는데 성능의 탐색 및 분석은 콜라주로 대표되는 창작자의 Statement 키워드와 효과적인 기법의 실현정도를 기준으로 하였다. 본 논문의 다수의 표와 그림은 YouTube[1], 솔트파파의 트렌드 인사이드[2], 블로그 sfckth 이미지[3], 블로그 신윤복 미인도 부분도 액자[4], Arinze Stanley의 웹사이트[5]에서 가져왔다. 사람이 만든 비디오와 GAN을 활용한 비디오의 유사성과 차이점을 비교 분석하기 위해 알고리즘은 photo-to-video, StyleGAN2, pix2pix, LipGAN을 GitHub 소스코드와 Google Colaboratory을 사용하여 프로그래밍 하였다. 영상편집은 VEGAS를 사용했으며, 시각적 표현은 직접 그린 이미지와 Photoshop을 활용하여 이미지 제작 및 가공을 하고 GIF애니메이션을 함께 구현하였다.

2. 콜라주 기법으로 실험 분석한 GAN

실험에 사용된 GAN 성능의 탐색 분석 기준은 콜라주(collage)로 대표되는 창작자의 Statement 키워드로 표1과 같으며 유사한 딥러닝 키워드도 함께 비교하여 제시한다. 단, 딥러닝 키워드를 실험에 적용하지는 못했다.

표 1. Statement와 Deep learning 키워드 비교

스테이트먼트 키워드	딥러닝 키워드
생성력	Pdata(X)학습데이터에 ½이 되는 지점의 Pmode(X) 데이터의 생성, 내쉬 평형(Nash Equilibrium)
한 객체의 다양한 모습	Inception Score (IS)에서 diversity(다양한 이미지의 정도)의 높은 수치
데페이즈망(dépaysement), 초현실주의, 우연, 꿈, 잠재의식	World model(VAE와 MDN-RNN구조)
생성의 다양한 모습	VAE의 latent vector의 encoder에서의 추상화의 정도, CGAN의 latent variable Y의 추가 및 그 다양성 $\min_G \max_D V(D,G) = \mathbb{E}_{x \sim p_{data}(x)}[\log D(x\|y)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z\|y)))]$
실체의 이면(또 다른 실체의 모습), 추상적	latent variable(내재변수,잠재변수), 사전에 정의될 수 없는noise

첫 번째) Face Image Motion Model (Photo-2-Video)을 활용한 이미지 애니메이션은 소스 이미지의 객체가 구동 비디오의 움직임에 따라 비디오 시퀀스를 생성하게 된다. 프레임 워크는 애니메이션할 특정 개체에 대한 주석이나 사전 정보를 사용하지 않으며 같은 카테고리의 사물을 묘사하고(예 :얼굴, 인체) 이를 달성하기 위해 외모와 동작 정보를 분리하게 된다.

그림 1. 얼굴 바꾸기

그림1과 같이 사전 훈련된 Face-Image-Motion-Model을 활용하여 구동이미지와 한 장의 소스 이미지를 매핑시켜 시퀀시애니메이션을 만들어 내고 아래 그림2와 같이 원본영상과 함께 듀엣으로 노래 부르는 비디오로 제작하였다.

그림 2. 원본영상(좌)과 생성된 비디오(우)

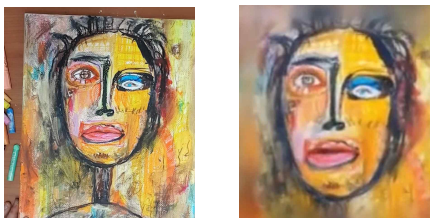

그림 3. 직접 그린 소스 이미지(좌)와 생성된 비디오(우)

그림3은 직접 그린 소스 이미지와 생성된 비디오의 또 다른 실험의 예로 소스 이미지가 현실적인 얼굴의 모습에서 멀어질수록 생성비디오는 더욱 Blur해 지는 현상이 관측되었다. 실험을 통해 Photo-2-Video의 문제점과 보완점을 콜라주 기법 측면에서 보면 표2와 같다.

표 2. Photo-2-Video의 문제점과 보완점

문제점	보완점
이미지와 비디오의 얼굴의 위치나 방향이 서로 비슷해야 성능이 좋음, 거리나 방향이 멀어질수록 Blurring현상 심함, 소스이미지가 그림이미지인 경우 생성비디오에서 Blurring현상	이미지와 비디오의 방향과 위치가 달라도 거리를 추론하여 Blur현상을 방지할 필요성이 있음, Poisson Image Inpainting 제안
배경과 얼굴이미지간 자연스러운 분리가 안됨	Image Inpainting with GAN에서 backprop을 이용해 입력으로 들어온 손상된 이미지에 가까운 Encoding vector z를 찾도록 backprop의 loss function을 새롭게 디자인
해상도 사이즈가 256*256만 가능	Resolution을 확장하기 위해 fine discriminator를 추가, Pix2pix HD에 사용한 Coarse-to-fine Generator 와 Muliti-scale Discriminators 방법
구동 비디오에서 얼굴이미지가 사라지면 얼굴인식을 못하는 오류	Face Dectection기능의 보강, 높은 Mean averaged precision (mAP)이 필요(0.5~1.0), MTCNN은 3개의 neural network(P-Net, R-Net, O-Net)을 활용한 joint learning
얼굴이미지 Crop기능의 획일성	얼굴이미지 Crop기능의 사용자 조작이 가능한 편집툴의 필요

두 번째) LipGAN은 음성 신호로부터 입 모양을 생성하는 연구인데 사운드에 맞도록 입모양을 만들어내는 기술이다. 사진이미지의 얼굴과 그림이미지의 얼굴에 각각 LipGAN을 측정해 보았는데 아래와 같은 문제점을 확인하였다.

(i) 사진이미지 보다 그려진 얼굴의 입술동작이 더 명확하지 않았다.

(ii) 소스 이미지 얼굴의 입이 벌어진 경우에는 입 모양이 어색하고 다물어지지가 않는다.

(iii) 모든 사운드에 입모양이 반응하는 문제점(사람의 목소리와 같은 특정 사운드에만 반응하지 않고 모든 사운드에 반응함으로 입술의 작은 떨림이 지속적으로 발생한다)

(iv) 생성 비디오 데이터 입부분이 소스 이미지보다 블러링 되는 현상이 있다. 특히 실제 사람이 대화를 할 때는 입모양만 움직이는 것이 아니고, 얼굴표정과 방향, 손동작, 상체의 움직임 등을 같이 활용하는데 이러한 동작의 추가가 필요하다고 생각된다. 특히 아래 그림4와 같이 입술의 모양이 분명하지 않은 벽화그림의 경우 입술이외의 영역(코 일부)까지 블러링 처리 되며 미세한 움직임이 발생하나 그 움직임 또한 부정확하다.

그림 4. (좌)벽화그림, (우)입과 코의 블러링 현상

세 번째) Style GAN2는 StyleGAN을 더욱 개선한 것으로 이미지에 발생하는 물방울 같은 노이즈를 제거하기 위해 AdaIN 대신 추정 통계(estimated statistics)로 정규화(normalization)하여 물방울을 없애고, 잠재 공간에 지속성(continuity)을 제공하여 이미지 품질을 보다 향상시켰다. 또한 Progressive Growing 대신에 skip connection을 갖고 있는 계층(hierarchical) 생성자(Generator)를 사용하여 눈과 이의 침체(stagnation)를 줄여서 얼굴 방향의 변화와 함께 자연스럽게 변하도록 하였다. 본 논문에서는 Style GAN2의 얼굴변화(얼굴 일부분의 움직임, 회전등)와 생성된 얼굴이미지가 정답 얼굴을 닮아가는 과정을 탐색하며 특히 직접 그린 반추상적 얼굴이미지로 점차적으로 변화되는 과정을 측정하고 그 결과를 확인했다.

그림 5. 원본이미지(좌)와 얼굴회전과 입과 눈의 변화(우)

그림 5는 신윤복의 미인도로 (우)그림은 GAN에 의해 생성된 얼굴이미지가 원본이미지를 닮아가는 과정과 이후 눈과 입의 변화와 얼굴회전을 시킨 결과인데 얼굴 중심에서 멀어질수록 이미지가 더욱 Blur해지고 눈의 움직임도 명확하지 않다는 것을 알 수 있다.

그림6은 (좌)직접 그린 원본그림에서 (우)원본 그림을 Style GAN2로 생성한 이미지인데 원본에 비해 실재감이 떨어지고, 이미지의 뭉개짐 현상, 옅은 색감(낮은 채도), 화면에 생기는 층(픽셀층)등이 실험을 통해 확인 되었다.

그림 6. (좌)원본 그림, (우)Style GAN2가 생성한 이미지

생성된 이미지가 얼굴의 형상을 띄고 있지만 실제 사진 같은 이미지와는 다소 다른 모습(그려진 이미지)일 때는 회전 및 얼굴 중 일부를 움직일 경우에 형태를 알아보기 힘들 정도로 이미지가 좋지 못한 성능을 보임을 그림 7과 같은 실험을 통해 확인 하였다.

그림 7. (좌)생성이미지, (우)이미지 변환과정

결론적으로 사실적인 이미지의 경우 그 생성과 변화에 원본이미지와 유사하게 동작하지만 스케치와 같은 그림을 원본으로 하고 이와 유사하게 생성하는 경우나 얼굴의 일부분(눈, 코, 입 등)에 변화를 주게 되면 Style GAN2는 정확한 성능을 나타내지 못하는 결과를 보였다. 하지만 이러한 현상은 콜라주 기법 관점으로 보았을 때 추상 또는 반추상화 이미지 생성화 알고리즘을 추가로 보강한다면 (반)추상화 비디오 생성이라는 좋은 결과를 보여줄 것이라고 추론해 본다.

마지막으로) Pix2pix는 흑백과 칼라 영상이 한 쌍으로 있는 Pair Data를 모아서 CNN을 기반으로 학습시켜 문제에 적용하는 것인데 이미지의 Style을 변형시키는 Supervised Learning으로, Pair로 되어있는 Dataset을 이용해서 Image to Image Translation하는 것이다.

그림 8. 스케치를 실제 고양이처럼 바꿔주는 Pix2pix

그림 8은 pix2pix를 활용하여 스케치한 고양이 그림을 실제와 유사한 고양이 이미지로 Translation한 것이다. 실험방법은 아래 그림9와 같이 약간 물체의 형태를 외곡하거나 기괴하게 표현한 경우에는 표현이 불분명, 색감의 뭉개짐, 낮은 채도, 단조로운 색 등의 결과가 나타났는데 이것은 논문에서 말하는 예술가나 디자이너들의 상상력을 자극하는 도구로 활용하기에는 문제가 있어 보인다.

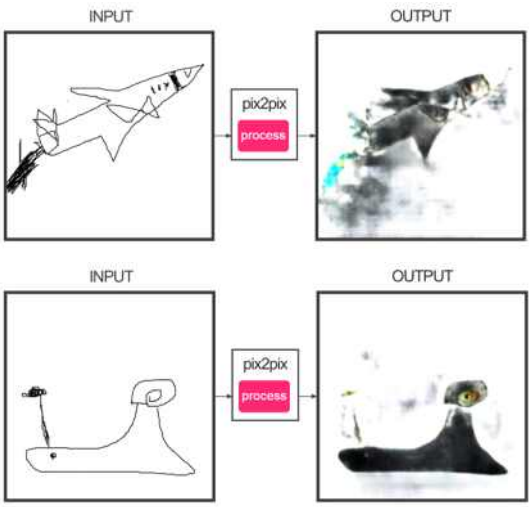

그림 9. Pix2pix를 활용한 테스트

Pix2pix는 특히 콜라주 기법과 유사성이 많아 보인다고 생각하는데 스케치를 하면 이를 변환시켜 실제 이미지처럼 나타내려는 특징을 가지고 있다. 그림 10은 수작업으로 스케치하고 채색한 그림인데 Pix2pix의 결과보다 선명한 색감, 기계적이지 않코 유연한 특징, 이미지의 뭉개짐 현상이 없다.

그림 10. 사람이 그린 스케치와 생성이미지

특히 제작기법에 따른 이미지 생성의 부족함은 보완할 부분 중에 하나인데, 만약 초현실적인 이미지의 생성이 목표라고 했을 때 Pix2pix를 통해 초현실적인 순간과 강력한 감정을 포착하기에는 많이 부족하다. 그림11은 Arinze Stanley의 초현실적 드로잉인데 Pix2pix에게는 이러한 초현실적 기법과 미적인 표현의 명령은 불가능한 일이다.

그림 11. Arinze Stanley의 초현실적 드로잉

Pix2pix와 사람의 수작업과의 비교실험을 위해 낙서를 실사와 같은 디지털 이미지로 생성하는 콜라주비디오를 제작해 보았는데 이처럼 어린아이 같은 감성의 사실적인 표현을 Kiddie Arts(키디아트)[6]라고 하며 이 역시 Pix2pix알고리즘으로 표현하기에는 개선할 성능(소스 이미지가 상식적으로 생각하는 모습에서 약간만 달라져도 색감의 뭉개짐, 낮은 채도, Blur해지는 현상, 표현의 Reality 감소 등)이 많아 보인다.

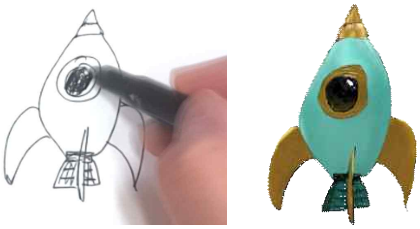

그림 12. (좌)직접 그린 이미지, (우)디지털이미지

3. 결론

본 논문에서는 GAN을 콜라주라는 관점에서 해석해 보기 위한 실험 및 관측을 하였는데, 이를 위한 비디오 제작에는 사람이 직접 그린 이미지와 GAN이 생성하는 이미지를 비교분석함으로써 그 성능의 문제점과 개선점을 제안하였다. 테스트로 사용된 GAN의 종류는 모두 4가지로 photo-to-video, StyleGAN2, pix2pix, LipGAN등이다. 전반적으로 사람의 수작업과 영상 및 그래픽 편집프로그램(VEGAS, Photoshop, GIF애니메이션)을 활용하여 만들기 어렵거나 제작에 시간이 많이 소요되는 것에는 GAN의 장점을 볼 수 있는데 예를 들면 photo-to-video의 소스 이미지에 구동비디오를 매핑시켜 생성하는 비디오, StyleGAN2의 1024X1024의 높은 해상도로 생성하는 존재하지 않는 사람의 생성과 그 변화(얼굴의 부분별 변화, 방향의 변화 등), pix2pix의 스케치를 실사이미지로 빠르게 바꿔주는 기능, LipGAN의 사운드에 맞게 움직이는 입술의 움직임 등이 이에 해당한다. 하지만 실험을 통해 관측한 것과 같이 각각의 GAN은 창작자의 다양한 목적에 맞도록 좋은 성능을 내기에는 개선할 부분이 있었다. 전체적인 문제점으로 지적되는 것으로 생성이미지가 Blur해지는 현상, 학습한 데이터와 약간만 형태가 달라도 생성을 잘 못하는 경우, 고해상도 비디오 생성의 어려움(시간이 오래 걸리거나 1920X1080이상의 해상도 출력은 불가능, 본 논문에서 실험하지는 않았지만 Video-to-Video Synthesis와 이를 더욱 개선한 World-Consistent Video-to-Video Synthesis에서는 2048X1024해상도의 비디오 출력이 가능함),구동비디오에서 얼굴이 사라지면 Face Dectection이 떨어져 Blur해지는 현상, 사진이미지에 비해 직접 그린 그림이미지에 낮은 인식률을 보였고, 사용자 조작을 통해 유저의 목적에 따른 이미지 Crop기능의 부족, 비디오 생성시 사운드와 영상의 Shink가 잘 안 맞는 문제점, 얼굴방향 변화의 경우 중심에서 멀어질수록 외곡되거나 Blur되는 현상, 원본과 다른 이미지의 생성(낮은 채도, 형태의 왜곡, 단색, 화면에 보이는 픽셀의 딱딱함) 등의 문제점을 보였다. 이를 개선하기 위해 본 논문에서는 각각의 GAN의 기본적인 성능 개선점을 말하였고 특히 콜라주 기법으로 대표되는 창작자의 Statement를 제시하였는데 이러한 딥러닝 키워드와의 비교분석을 통한 연구는 앞으로 진행할 예정이다.

참고문헌

[1] YouTube: https://dreamgonfly.github.io/blog/gan-explained/
[2] 솔트파파의 트렌드 인사이드: https://saltpapa.tistory.com/269
[3] 블로그 sfckth 이미지: https://m.blog.naver.com/PostView.nhn?blogId=sfckth&logNo=220959908696&proxyReferer=https:%2F%2Fwww.google.com%2F
[4] 블로그 신윤복 미인도 부분도 액자: https://ilee3.tistory.com/m/1013
[5] Arinze Stanley의 웹사이트: https://www.arinzestanley.com/drawings
[6] Kiddie Arts(키디아트): http://www.telmopieper.com/kiddie-arts