Journal of the Korea Industrial Information Systems Research Vol.25 No.2, Apr. 2020 : 11-18
http://dx.doi.org./10.9723/jksiis.2020.25.2.011 ISSN:1229-3741

이미지를 사용한 가상의상착용을 위한 개선된 알고리즘+

(An Improved VTON (Virtual-Try-On) Algorithm using a Pair of Cloth and Human Image)

미나르 마드울 라흐만[1], 따이 트안 투안[2], 안 희 준[3]*
(Matiur Rahman Minar, Thai Thanh Tuan, and Heejune Ahn)

요 약 최근 이미지를 사용한 가상착용기술 (Virtual try-on: VTON)에 대한 일련의 연구들이 발표되었다. 이에 의상과 사용자 이미지를 사용한 대표적 방식 (SCMM 기반의 비-딥러닝 방식, 딥러닝 기반 VITON 과 CP-VITON)에 대해 인물의 자세 및 체형, 의상의 가려짐 정도, 의상의 특성 등에 따라 분석한 연구가 보고되었다. 본 논문에서는 이중 가장 좋은 성능을 보이는 CP-VTON의 문제점을 살펴보고 이에 따른 해결책을 제시한다. 구체적으로 대상인물의 분할 표현 문제, 교체 대상이 아닌 영역이 유지되지 못하는 문제, 합성 마스크 생성네트워크의 학습에 사용되는 비용함수 문제, 합성 네트워크의 마스크 문제를 지적하고 이를 개선하는 알고리즘을 제안하였다. 그 결과 SSIM 등에서 5%내외의 주관적으로는 상당한 개선을 보였다.

핵심주제어 : 가상착용, 딥러닝, 인간 표현, 성능 개선, 비용 함수

Abstract Recently, a series of studies on virtual try-on (VTON) using images have been published. A comparison study analyzed representative methods, SCMM-based non-deep learning method, deep learning based VITON and CP-VITON, using costumes and user images according to the posture and body type of the person, the degree of occlusion of the clothes, and the characteristics of the clothes. In this paper, we tackle the problems observed in the best performing CP-VTON. The issues tackled are the problem of segmentation of the subject, pixel generation of un-intended area, missing warped cloth mask and the cost function used in the learning, and limited the algorithm to improve it. The results show some improvement in SSIM, and significantly in subjective evaluation.

Keywords: Virtual-try-on, Deep-learning, Human representation, Quality improvement, Loss function

* Corresponding Author: heejune@seoultech.ac.kr
+ 이 논문은 본 연구는 서울과학기술대학교 교내연구비의 지원으로 수행되었습니다(2019-0396)
Manuscript received January 23, 2020 / revised February 22, 2020, accepted February 22, 2020
1) 서울과학기술대학교 전기정보공학과, 제1저자
2) 서울과학기술대학교 전기정보공학과, 제2저자
3) 서울과학기술대학교 전기정보공학과, 제3저자, 교신저자

1. 서 론

최근 2차원 이미지를 사용한 가상착용기술에 관심이 증가하고 있다(Ahn, 2018a, 2018b; Han et al., 2018; Wang et al. 2018, Tuan, 2019). 이미지 기반 가상 착용기술로 현재까지 발표된 연

구는 (Ahn, 2018, 2018b)를 포함하여 딥러닝을 중심으로 한 VITON (Virtual Try-ON) (Han et al., 2018), CP-VTON (Content Preserving Virtual Try-ON) (Wang et al. 2018), SwapNet (Raj et al. 2018) 등이 있다. 이들은 이차원 이미지를 입력으로 사용하고 2차원 영상처리 알고리즘을 사용한다. 알고리즘 세부적으로는 이미지에서의 사람의 2차원 자세 예측, 이미지 분할 기술, 2차원 기하변환, 2차원 이미지 블렌딩 기술을 사용하며, 전통적인 규칙기반 방식과 딥러닝에 의한 학습기법이 같이 사용되고 있다.

Tuan (2019)의 연구는 이러한 최근 발표된 의상과 대상 인물의 2차원 이미지를 사용한 가상착용 알고리즘들을 인물의 자세 및 체형, 의상의 가려짐 정도, 의상의 특성 등을 기준으로 자세히 비교하였다. 분석결과 기존 발표된 알고리즘 중에 CP-VTON 방식이 가장 좋은 성능을 보이긴 하지만, 여러 가지 문제점과 한계점을 갖고 있는 것을 보였다. 다음 절에서 상세히 지적하겠지만, 저자들이 판단해 보았을 때 그 문제점 중에는 기존의 기본 2차원 처리 구조에서 해결이 가능한 사항도 있고, 2차원 방식으로는 근본적으로 해결이 어려운 방법도 있는 것으로 판단된다. 본 연구에서는 2차원 방식으로 해결이 가능한 요소들에 대한 개선책을 제시하고 이를 실험을 통하여 검증하였다.

본 논문은 다음과 같이 구성하였다. 우선 제 2절에서 2차원 방식에서 가장 좋은 성능을 보이고 있는 CP-VTON 알고리즘의 문제점에 대하여 설명한다. 제 3절에서는 이 문제점들을 해결하기 위한 방법을 설명하고, 제 4절에서 의상 데이터셋을 사용한 실험한 결과를 보이고 이를 평가한다. 마지막으로 제 5절에서 실험 결과를 바탕으로 의미와 향후 개선 방향에 대하여 제시한다.

2. CP-VTON (Wang et al., 2018)

CP-VTON 네트워크를 사용한 가상착용 알고리즘들은 사용자 사진과 의상사진을 입력으로 하지만, 우선 사전에 사람의 2차원 (골격) 자세와 의상분할이 되어 있다고 가정한다. 본 연구에 사용한 데이터는 VITON 논문에서 저자들이 처음 사용하였고 이후 CP-VTON을 비롯한 많은 논문들에서 사용하고 있는 VITON 데이터셋을 사용하였다. 이 데이터 셋은 자세예측에는 OpenPose (Cao et al., 2017) 방식과 사용자 분할에는 LIP (Liang et al., 2018) 방식을 사용하여 얻어 졌다.

Fig. 1은 CP-VTON의 기본 구조는 의상의 기하변환을 행하는 GMM (Geometric manipulation module) 단계와 의상을 합성하는 TON (Try-on network)으로 구분된다. GMM 단계에서 VITON에서 처럼 마스크를 만드는 대신, STN (Spatial transformation network) (Jaderberg, 2015)을 사용하여 직접 변형된 결과를 생성한다. TON 단계에서는 초벌 가상착용 이미지를 생성함과 함께 알파-블렌딩을 위한 합성 마스크를 생성한다. 이렇게 구해진 초벌착용영상에 합성마스크를 가중치로 사용하여 GMM결과 변형의상과 합성을 수행하여 최종 가상착용이미지를 생성한다. 논문에서 저자들은 CP-VITON이 의상의 텍스쳐를 VITON에 비하여 잘 유지한다고 주장하는데, 그 근본적인 이유는 의상이 Encoder-Decoder 네트워크를 통과하지 않고, 원본 이미지를 GMM에서 기하변환 결과를 블렌딩으로 합치기 때문이다.

인공신경망의 블랙박스 모델이라는 특성상 네트워크 내부적으로 어떤 로직을 따지는지 상세한 설명은 어렵다. 하지만, 직관적으로 GMM 네트워크와 TON 네트워크의 동작을 설명하면 다음과 같다. 우선 GMM 네트워크는 입력의상과 사용자의 조인트와 실루엣 정보의 상관도를 이용하여 TPS 변수를 Regression 방식으로 찾아낸다. TON 네트워크는 사용자 입력실루엣을 바탕으로 배경영역과 전경 영역을 구분하고 전경 내역인 경우는 의상의 화소정보를 가져올 수 있도록, 전경영역이 아닌 곳은 의상의 화소정보로부터 유추하며, 그 외의 영역은 피부색으로 예측하도록 네트워크가 구성되는 것으로 생각할 수 있다.

CP-VTON을 의상과 사용자의 자세 등을 기준으로 분류하고 실험을 분석한 결과 크게 다음과 같이 기존의 알고리즘에 개선요소가 확인되었다. Fig. 2에 대표적인 CP-VTON 결과를 제안하는 방식 (CP-VTON+)와 비교하여 제시하였다.

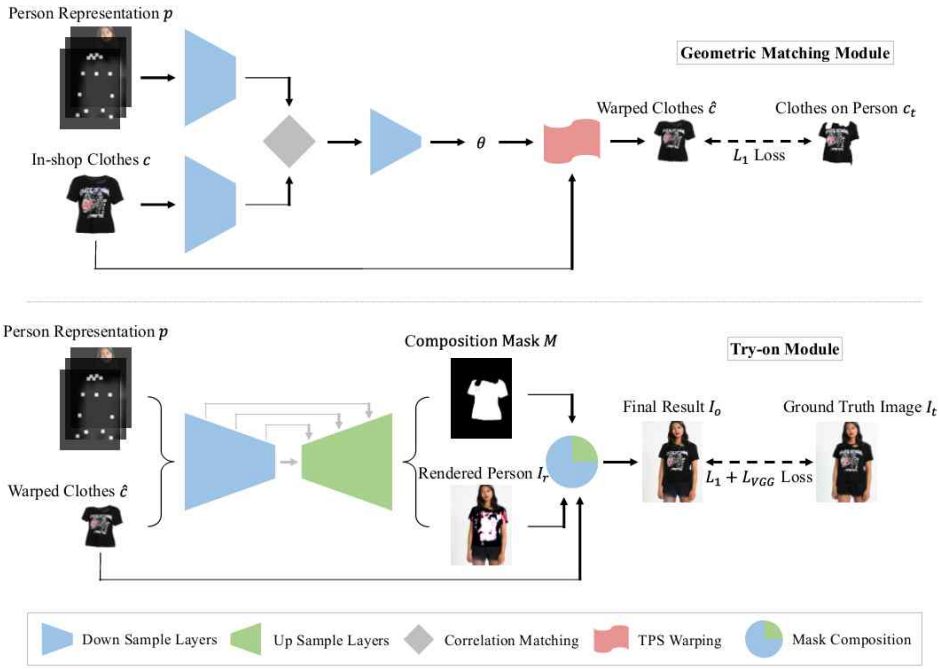


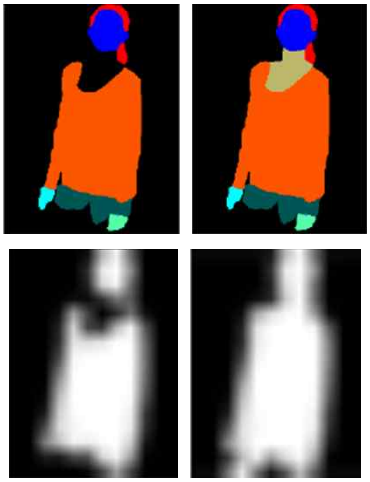
Fig. 1 CP-VTON NN Pipeline (Wang et al., 2018)

Fig. 2 Clear Improvement Cases

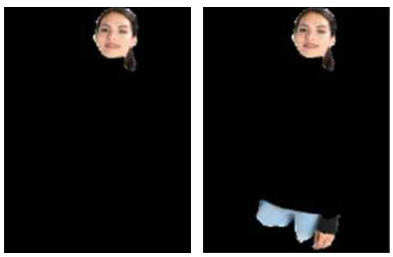
- 문제점 1: 타겟 사용자 실루엣 표현에서 Neck/check 부분이 백그라운드로 구분됨으로써, 의상의 목 부분이 타겟 사용자의 현재 의상에 영향을 받게 된다. 또한, 사용자의 머리카락이 옷을 가리는 경우 이 또한 사용자 실루엣에 변화를 주고 결국 의상의 변형에 영향을 주게 된다 (Fig. 3 (a) 좌측).
- 문제점 2: 대체 대상인 의상을 제외한 다른 의상이나 영역은 원본의 색상을 유지해야 하는데 (현재 테스트에서는 바지와 하체), TON 네트워크에 의하여 생성된다. GAN 네트워크의 특성상 그럴듯한 모습을 보이지만, 이는 일반적으로 VTON 응용에서 원하는 바가 아니다 (Fig. 3 (b) 좌측).
- 문제점 3: 식 (1)은 학습 손실인데, 복원 손실과 VGG 손실 Ma나에 대한 손실로 구성되어 있다. 그러나 Composition 마스크에 사용하는 비용함수가 단순히 정규화 목적함수이기 때문에 마스크의 영역이 선명하지 않아, 이로 인하여 의상의 색상과 텍스쳐가 선명하지 못하다.

$$L = c_1|L_0 - L_{GT}| + c_2 L_{VGG} + c_3|1 - M_0| \quad (1)$$

- 문제점 4: 의상의 색상과 배경이 같은 경우 TON 네트워크가 이를 혼동하는 경우가 발생한다. (Fig. 3 (c))
- 문제점 5: TPS의 파라메터를 추정하여 적용하는 현재의 GMM 모듈은 타겟 사용자 이미지가 팔장을 끼거나 들고 있는 등의 의상의 기본 세팅과 크게 차이나는 포즈를 취하



(a) Segmentation for body shape



(b) Un-replaced cloth area

(c) Mask for same BG colored cloth

(d) Large target pose

Fig. 3 Key Issues in CP-VTON (Improvement in the Proposed System)

는 경우 이를 반영하는 것이 근본적으로 어렵다. (Fig. 3 (d))

참고로 그럼에도 불구하고 해당 논문에서 제시된 결과들이 그럴듯한 모습을 보이는 것은 1) 사용된 의상이 반팔로 사용자의 자세에 민감하지 않고, 2) 큰 포즈를 취한 예들이 상대적으로 적으며, 3) 의상이 단색 위주이기 때문에 TON의 In-painting 기법에 의하여 오류가 감추어졌기 때문이다.

3. 개선된 CP-VTON: CP-VTON+

앞서 지적한 사항 중 문제점 5를 제외한 문제점 들을 해결하는 전체 구조를 수정되는 요소를 표시하여 Fig. 4과 같이 제시하였다. 각 수정 요소에 대한 설명은 다음과 같다.

- 수정 1: 사람 표현을 수정하기 위하여 목 부분의 Label인 Skin을 추가하였고, 또한 머리카락 등을 제외하고 사용자 실루엣을 구성하였다 (Fig. 3. (a) 우측).
- 수정 2: TON 적용시 교체되는 의상을 제외한 영역의 입력을 추가하였다 (Fig. 3. (b) 우측).
- 수정 3: 합성마스크가 교체 의상에 집중이 될 수 있도록 Loss 함수를 식 (2)와 같이 변경 하였다.

$$L = c_1|L_0 - L_{GT}| + c_2 L_{VGG} + c_3|M_{GT} - M_0| \quad (2)$$

- 수정 4: 트레이닝과 테스트 과정에서 의상의 mask를 추가로 입력하여 영역을 구분할 수 있도록 하였다.

4. 실험 결과

실험에 사용한 이미지는 VITON과 CP-VTON에서 사용한 데이터셋을 기반으로 사용하였다. 성능비교는 의상 변형 결과와 최종 블렌딩 결과를 따로 평가하였다. 의상 변형 결과의 대표적인 예는 Fig. 5와 같다. CP-VTON+의 결과가 좀 더

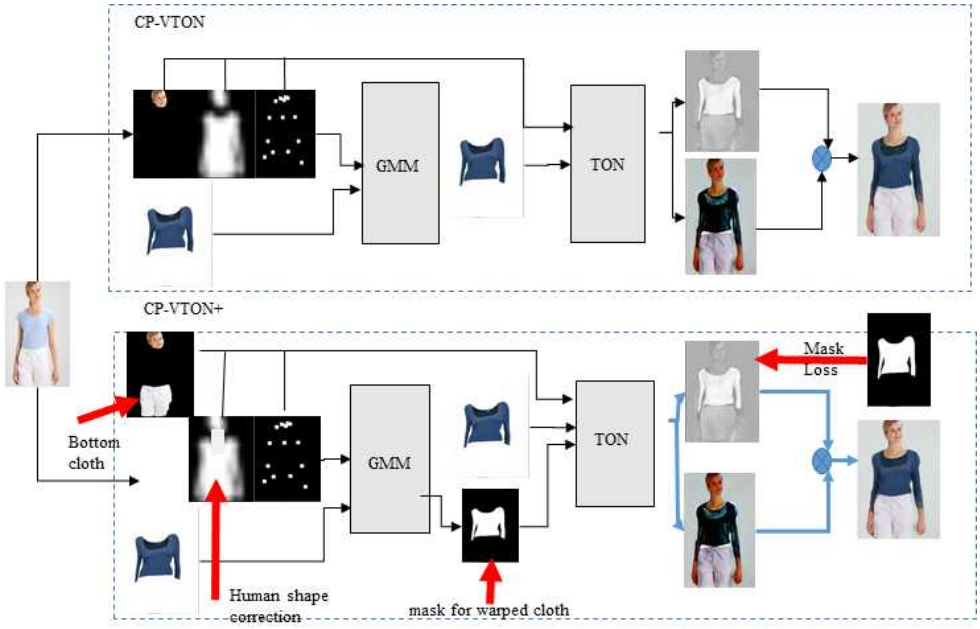


Fig. 4 Enhanced Proposed Algorithm: CP-VTON+

정상에 가까운 형태를 유지하고 있다. 그러나 여전히 변형이 자연스럽지 못한 결과가 눈에 뜨인다. 전체 테스트 데이터에 대하여 적용 후 IOU를 비교한 결과 0.78과 0.75로 오히려 CP-VTON이 약간 앞서는 것을 보였다. 이러한 결과가 나오는 이유는 잘못된 사용자 실루엣이기는 하지만, 동일의상, 즉 정답 정보가 적용되기 때문이라고 보인다. 주의할 점은 실제 해당 네트워크가 사용될 응용은 동일한 의상이 아닌 경우이며, 이 경우를 주관적으로 비교 하였을 때, 제안된 방식이 보다 좋은 결과를 보였다(Fig. 6). 이를 객관적인 수치로 평가하기 위해서는 동일한 모델이 동일한 자세로 두 개의 다른 옷을 착용하고 있는 데이터 셋이 필요하다.

다음으로 가상 착용의 최종결과를 SSIM (Structural similarity)과 LPIPS (Learned perceptual image patch similarity, Zhang et al., 2018), IS (Inception score)을 기준으로 확인하였다. SSIM은 사람 시각 시스템이 이미지에서 구조 정보에 민감한 점을 이용하여 PSNR (Peak Signal to Noise Ratio)보다 주관적인 화질을 잘 평가하는 것으로 알려진 객관적 평가기준으로 원본 이미지 x와 왜곡 이미지 y의 밝기, 콘트라스트, 구조를 비교한다. LPIPS는 신경망 모델에서 추출되는 특성을 이용하여 학습에 의하여 사람의 인지적 특성에 맞도록 유사도를 평가하는 새로운 모델이다.

IS 는 GAN 출력을 참고 이미지가 없이 자체적으로 얼마나 자연스러운가 하는 점을 평가하는 알고리즘이다. 두 가지 점을 평가하는데 하나는 출력결과가 얼마나 다양한가 하는 점과 출

Fig. 5 Warped Cloths through GMM Network

Fig. 6 Comparison and Analysis of IoU Result for Same Cloth and New Cloth

Fig. 7 Comparison of Final Results with CP-VTON Results

력이 기존에 학습된 데이터 대비 얼마나 유사점이 많은가 하는 점으로 값이 큰 것이 좋은 결과를 의미한다.

LPIPS는 0,1397 대 0.1263으로, SSIM은 0.7798 대 0.8076으로 약간 앞서는 결과를 얻었다. Inception Score (Baratt et al., 2018)의 결과는 2.7417대 2.76으로 모두 약간 높은 결과를 보였다. 육안으로 볼 때 CP-VTON+의 결과가 의상의 로고나 무늬가 보다 선명하게 보이는 것을 확인할 수 있다 (Fig. 7).

5. 결 론

본 논문에서는 최근 발표된 이미지 기반 딥러닝 기반 가상착용기술 중 가장 좋은 성능을 보이는 CP-VTON의 성능을 세부적으로 분석하고 이들의 문제점 5가지를 도출하였다. 이 5개의 문제들 중 문제 5 '3차원 변형'을 제외한 4가지 경우의 해결방안을 제시하고 이를 네트워크에 적용하였다. 앞선 논문 (Tuan et al., 2019)에서 지적한 바와 같이 CP-VTON 논문에서 저자들이 주장하는 바와 달리, 의상의 가려짐이 거의 없고, 자세가 변형이 거의 없는 경우에 한해서만 실용적인 의미가 있을 것으로 보인다. 그렇지 않은 범위의 경우에는 3D 방식 등의 새로운 알고리즘이 필요할 것으로 보인다. 하지만, 대부분의 실용화에서 사용된 알고리즘이 모든 경우를 해결하는 것은 아니듯, 제안된 방식의 한계점을 정확히 인식하고 입력을 제한할 수 있다면 이 논문에서 제안하는 방식으로 향상된 결과를 얻을 수 있을 것으로 판단된다.

References

Ahn H. (2018a). Online Virtual Try On using Mannequin Cloth Pictures, *Journal of the Korea Industrial Information Systems Research*, 23(6), 29 - 38.

Ahn H. (2018b). Image-based Virtual Try-On System, *Journal of Korean Computer Game Society*, 31(3), 37-45.

Barratt, S., and Sharma, R. (2018). *A Note on the Inception Score*, arXiv preprint arXiv:1801.01973.

Cao, Z., Simon, T., Wei, S. E., and Sheikh, Y. (2017). Realtime Multi-person 2d Pose Estimation using Part Affinity Fields, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7291-7299.

Han, X., Wu, Z., Wu, Z., Yu, R., and Davis, L. S. (2018). Viton: An Image-based Virtual Try-on Network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7543-7552.

Jaderberg, M., Simonyan, K., and Zisserman, A. (2015). Spatial Transformer Networks. Proceedings of Advances in Neural Information Processing Systems, pp. 2017-2025.

Liang, X., Gong, K., Shen, X., and Lin, L. (2018). Look into Person: Joint Body Parsing & Pose Estimation Network and a New Benchmark. *IEEE Transactions on PAMI*, 41(4), 871-885.

Raj, A., Sangkloy, P., Chang, H., Lu, J., Ceylan, D., and Hays, J. (2018). Swapnet: Garment Transfer in Single View Images. *Proceedings of the European Conference on Computer Vision,* pp. 666-682.

Tuan, T., Rahman, M., and Ahn, H. (2019). Performance Evaluation of VTON Algorithms using a Pair of Cloth and Human Image, *Journal of the Korea Industrial Information Systems Research*, 24(6), 24 - 30.

Wang, B., Zheng, H., Liang, X., Chen, Y., Lin, L., and Yang, M. (2018). Toward Characteristic-preserving Image-based Virtual Try-on Network. *Proceedings of the European Conference on Computer Vision,* pp. 589-604.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). The Unreasonable Effectiveness of Deep Features as a Perceptual Metric, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 586-595.

미나르 마드올 라호만
(Matiur Rahman Minar)

- BUET (방글라데시) 컴퓨터공학과 학사 (2015)
- Automation Solutionz Inc., (캐나다) 2014-2018, 방글라데시 원격근무 프로그래머
- 서울과학기술대학교 전기정보공학과 (2019-현재) 석사과정
- 관심분야 : 컴퓨터비젼, 딥러닝, 데이터마이닝

따이 트안 투안
(Thai Thanh Tuan)

- HUTECH (베트남) 컴퓨터공학과 학사 (2010)
- HCMVNU (베트남) 컴퓨터공학 석사 (2015)
- 서울과학기술대학교 전기정보공학과 (2017-현재) 박사과정
- 관심분야 : 컴퓨터비젼, 딥러닝, 데이터마이닝

안 희준 (Heejune Ahn)

- 종신회원
- KAIST 전기정보공학과 박사 (2000)
- (주) LG전자 차세대단말연구소 선임연구원(1998-2002)
- (주) Tmax 소프트 책임연구원 (2002-2004)
- 서울과학기술대학교 전기정보공학과 (2004-현재) 정교수
- 관심분야 : 컴퓨터비젼, 컴퓨터 통신, 데이터마이닝