Journal of Korean Society of Industrial and Systems Engineering
https://doi.org/10.11627/jksie.2021.44.4.193

Vol. 44, No. 4 : 193-207, December 2021
ISSN : 2005-0461 / EISSN : 2287-7975

# Development of Dataset Evaluation Criteria for Learning Deepfake Video

Rayng-Hyung Kim · Tae-Gu Kim†

Department. of Industiral & Management Enginering, Hanbat National University

# 딥페이크 영상 학습을 위한 데이터셋 평가기준 개발

김량형 · 김태구†

한밭대학교 산업경영공학과

As Deepfakes phenomenon is spreading worldwide mainly through videos in web platforms and it is urgent to address the issue on time. More recently, researchers have extensively discussed deepfake video datasets. However, it has been pointed out that the existing Deepfake datasets do not properly reflect the potential threat and realism due to various limitations. Although there is a need for research that establishes an agreed-upon concept for high-quality datasets or suggests evaluation criterion, there are still handful studies which examined it to-date. Therefore, this study focused on the development of the evaluation criterion for the Deepfake video dataset. In this study, the fitness of the Deepfake dataset was presented and evaluation criterions were derived through the review of previous studies. AHP structuralization and analysis were performed to advance the evaluation criterion. The results showed that Facial Expression, Validation, and Data Characteristics are important determinants of data quality. This is interpreted as a result that reflects the importance of minimizing defects and presenting results based on scientific methods when evaluating quality. This study has implications in that it suggests the fitness and evaluation criterion of the Deepfake dataset. Since the evaluation criterion presented in this study was derived based on the items considered in previous studies, it is thought that all evaluation criterions will be effective for quality improvement. It is also expected to be used as criteria for selecting an appropriate deefake dataset or as a reference for designing a Deepfake data benchmark. This study could not apply the presented evaluation criterion to existing Deepfake datasets. In future research, the proposed evaluation criterion will be applied to existing datasets to evaluate the strengths and weaknesses of each dataset, and to consider what implications there will be when used in Deepfake research.

Keywords : Deepfake, Dataset, Video, Evaluation criteria, AHP

## 1. 서 론

딥페이크(Deepfake)는 딥러닝(deep learning) 기법에 의한 조작된 이미지나 영상을 말한다. 최근 딥페이크는 'n번방', '지인능욕' 등 디지털 성범죄로 악용되면서 우리 사회에 중요한 이슈로 떠올랐다.

딥페이크 악용이 급증할 수 있었던 배경에는 Faceplay, Face2Face 등 오픈 소스가 큰 몫을 하였기 때문이다. 오픈 소스를 통해 전문지식 없이도 누구나 쉽게 제작할 수 있고 제작된 영상을 SNS와 같은 웹 플랫폼을 통해 빠르게 확산시킬 수 있기 때문이다. 디지털 카메라와 영상/이미지 편집 프로그램 기술이 나날이 정교해지고 발전하는 상황에서 딥페이크의 해악은 더욱 심화될 것으로 예견된다.

Received 24 November 2021; Finally Revised 21 December 2021; Accepted 24 December 2021
† Corresponding Author : taegu.kim@hanbat.ac.kr

딥페이크 악용에 대한 우려가 높아지면서 학계에서는 딥페이크를 효과적으로 탐지할 수 있는 방법론 개발이 한창이다[2, 5, 6, 16]. 대부분의 딥페이크 탐지 모델은 특정 딥페이크 영상 데이터셋을 기반으로 생성되며 학습 데이터에 존재하는 시각적 아티팩트(artifact)에 초점이 맞추어져 있다. 따라서 딥페이크 영상 제작과정에서 얼굴합성이나 얼굴교환 방법에 따라 모델의 성능이 크게 좌우된다. 다시 말해, 딥페이크 탐지 모델의 성능은 학습 데이터에 제약되기 때문에 학습 데이터의 품질은 무엇보다도 중요하다고 할 수 있다.

그러나 선행연구에서는 현재까지 공개된 딥페이크 영상 데이터셋들이 품질 측면에서 여러 한계점(예: 소규모 데이터, 낮은 시각적 품질, 다양성 미확보 등)이 존재하여 데이터 가용과 일반화하기에 충분하지 못하다고 지적하고 있다[13, 28, 29, 33].

일반적으로 인공지능 기법을 활용한 학습모델은 학습 데이터(training data)를 바탕으로 모델링되기 때문에 모델 성능은 학습 데이터에 의해 결정되는 특성이 있어 학습 데이터의 품질은 매우 중요하다. 이에 따라 현재 인공지능 학습용 데이터의 품질과 평가기준에 대한 가이드라인의 필요성이 중요시되고 있으며 산업 및 학계에서는 이에 대한 논의가 한창 진행 중이다[9, 11, 19, 20, 34].

UCI[36]나 Kaggle[37]과 같은 인공지능 학습용 데이터셋은 연구용으로 널리 활용되어지면서 데이터셋의 가용성과 일반화가 확보되었다고 볼 수 있을 것이다. 그러나 딥페이크 영상 데이터셋의 경우, 불과 3년 전부터 공개되었기 때문에 기존의 인공지능 학습용 데이터셋의 역사와는 비교될 정도로 짧다.

딥페이크 영상 데이터셋은 데이터 품질관리 성숙단계로 보면, 데이터 품질에 대해 문제점과 필요성을 부분적으로 인지하고 이를 해결하려고 노력하는 단계인 1단계(도입)'에 머물러 있다(<Table 1> 참고).

<Table 1> Data Quality Management Growth Stage

| Step | Contents |
|---|---|
| **Step 1** (Introduction) | Partial awareness and implementation of the issues and requirements of data quality management |
| **Step 2** (Standardization) | Quantify the basis(process, solution, etc.) for data quality management |
| **Step 3** (Intergration) | Perform consistent data structure quality management from integration perspective |
| **Step 4** (Quantification) | Data quality management through statistical techniques or quantitative methods |
| **Step 5** (Optimization) | Follow-up management through continuous improvement/promotion and evaluation from a company-wide perspective |

Sorce: KOREA Data Agency[11].

이에 따라 딥페이크 영상 데이터셋도 인공지능 학습용 데이터셋과 같이 데이터 품질 강화를 위한 기준 수립이 필요할 것으로 보인다. 그러나 선행연구에서는 구체적인 품질 방안이나 평가기준을 마련한 연구는 거의 없는 실정이다[16].

이에 본 연구는 딥페이크 영상 데이터셋의 평가기준을 마련하는데 연구의 초점을 맞추고자 한다. 본 연구에서는 평기기준 마련을 위해 선행연구를 검토한다. 그리고 이를 토대로 딥페이크 영상 데이터셋의 품질 기준 적합성을 제안하고 구체적인 평가항목을 도출한다.

평가기준의 고도화를 위해 본 연구에서는 전문가 의견을 수렴한다. 이를 위해 선행연구 검토를 통해 도출된 평가항목들을 AHP 계층화하고 최종적인 평가기준을 마련한다.

## 2. 선행연구 검토

### 2.1 기존 딥페이크 영상 데이터셋 검토

현재까지 공개된 딥페이크 영상 데이터셋은 FaceForensics++, Celeb-DF, DeeperForensics-1.0 등 9여 개 정도로 확인된다. 딥페이크 영상 데이터셋은 딥페이크 이미지 데이터셋에 비해 그 수는 적지만, 전 세계 여러 조직에서 데이터 구축과 벤치마크가 이루어지고 있어 빠르게 증가하고 있는 상황이다.

기존 딥페이크 영상 데이터셋들은 각각의 특징을 지니고 있다. 선행연구에서는 이 특징을 기준으로 딥페이크 영상 데이터셋을 세대별로 분류하고 연구를 진행하였다[2, 16, 27]. 세대별 분류는 한 세대 내에서의 공통된 특징이 무엇인지 파악가능하고, 다음 세대로 넘어가면서 시간적 경과에 따른 세대별 차이가 무엇인지 파악할 수 있어 분석에 용이하다. 이에 본 연구에서도 선행연구와 같이 기존 딥페이크 영상 데이터셋들을 세대별로 분류하고 문헌검토를 진행하고자 한다. 기존 딥페이크 영상 데이터셋들은 3세대로 분류 가능하며 그 내용은 <Table 2>와 같다.

1세대 딥페이크 영상 데이터셋에는 UADFV, Deepfake-TIMIT, FaceForensics++가 있다. 1세대의 특징을 살펴보면 데이터 규모는 영상수 5,000개 미만이며, 영상 대부분이 인터넷(예: YouTube 등)을 출처로 하고 있어 등장인물의 콘텐츠 동의권이 없다는 것이 특징이다. 1세대는 연구를 위한 자료제공을 목적으로 처음 공개되었다는 점에서 의의가 있다. 그러나 제한적인 데이터 규모나 낮은 품질, 등장인물의 초상권 문제가 한계점으로 남는다.

2세대는 1세대의 한계점을 극복하고자 여러 노력을 기울였다는 점에서 의의가 있다. 2세대 딥페이크 영상 데이

<Table 2> Introduction of Deepfake Datasets

| Dataset | | Scale(Real:Fake) | Release Date | Source | Research |
|---|---|---|---|---|---|
| 1 Generation | UADFV | 49:49 | 2018.11 | YouTube | [15] |
| | Deepfake-TIMIT | 320:320 | 2018.12 | Actors | [12] |
| | FaceForensics++ | 1,000:4,000 | 2019.01 | YouTube | [22] |
| 2 Generation | DDD | 363:3,068 | 2019.09 | Actors | [4] |
| | DFDC-preview | 1,131:4,113 | 2019.10 | Actors | [1] |
| | Celeb-DF | 590:5,639 | 2019.11 | YouTube | [16] |
| 3 Generation | DFDC | 48,190:104,500 | 2019.10 | Actors | [2] |
| | DeeperForensics-1.0 | 10,000:50,000 | 2020.05 | Actors | [7] |
| | KoDF | 175,766 | 2021.06 | Actors | [13] |

터셋은 2019년 후반기에 출시되었으며 DDD(Deepfakes Detection Dataset), DFDC-preview, Celeb-DF가 있다. 2세대에서는 소규모 데이터의 한계점을 극복하고자 이전 세대 대비 규모가 10배 이상 증대되었다[2]. 또한 등장인물의 초상권 문제를 해결하기 위해 실험환경 조성을 통한 영상제작이 이루어지기 시작한다.

실험환경은 제작자가 연출의도를 반영하여 연출에 필요한 모든 제반사항을 인위적으로 세팅해야 함을 의미하기도 한다. 따라서 실험환경이 얼마나 실제 시나리오와 같이 구현할 수 있는지가 중요하다. 그러나 2세대에서는 실험환경의 조성이 막 시작되었을 뿐 그 중요성에 대해서는 크게 인식되지는 않았다.

3세대는 2019년 후반부터 현재까지 공개된 딥페이크 영상 데이터셋들을 포함한다. 3세대에는 DFDC, DeeperForensics-1.0, KoDF가 있다. 2세대에서 초상권 문제의 해결방안으로 실험환경 조성이 이루어졌다는데 의미를 두었다면, 3세대에서는 탐지 모델의 일반화(external validity)와 데이터셋의 가용성을 높이기 위해 실제 시나리오와 같은 구체적인 실험설계가 이루어졌다는 점에서 의의가 있다.

3세대의 모든 딥페이크 영상 데이터셋들은 실험환경 조성을 통해 영상제작이 이루어졌고, 데이터는 10,000개 이상의 대규모 데이터셋이라는 점에서 큰 특징이 있다. 또한 등장인물의 포즈나 표정, 조명기법 등 실험환경의 통제 여부가 중요시되고, 다양한 영상소스나 변조방법을 반영하는 등 다양성 확보의 노력이 돋보인다.

## 2.2 평가기준 도출을 위한 연구방향

이상의 내용을 정리하면 딥페이크 영상 데이터셋은 연구를 위한 데이터 제공을 목적으로 만들어졌으나 한계점으로 인해 데이터 가용성과 품질 측면에서 문제가 있었음을 알 수 있다. 차기 연구에서 기존의 한계점을 극복하면서 데이터 품질 개선에 노력을 기울였으나 이는 부분적으로 인지하여 개선되었을 뿐 아직도 개선되어야 할 부분은 많다[7, 13].

딥페이크 영상 데이터셋의 궁극의 목표는 실제 딥페이크 대해 잘 수행되는 탐지 모델 개발에 도움이 되는 것이다. 따라서 본 연구는 향후 딥페이크 영상 데이터셋 품질 관리 기준 수립과 방향성을 제시하기 위해 딥페이크 영상 데이터셋의 품질 기준 적합성(fitness)을 <Table 3>과 같이 제안하고자 한다.

본 연구에서 제안된 품질 기준 적합성은 6가지 품질 특성 차원으로 구성되었으며 인공지능 학습용 데이터 품질 관리 가이드라인[19]을 바탕으로 하였다.

인공지능 학습용 데이터 품질관리 가이드라인에서는 학습 데이터의 품질이 전체 품질 수준을 좌우하기 때문에 학습 데이터의 품질 확보가 중요하다고 말하며, 데이터셋 구축시 신뢰성, 충분성, 다양성, 윤리성, 다양성, 사실성, 공평성이 고려되어야 한다고 강조한다.

딥페이크 영상 데이터셋도 인공지능 학습용 데이터셋과 같이 품질관리 기준의 적용이 가능하다. 다만 딥페이크 영상 데이터셋의 구축 목적과 특성이 고려된 품질 기준이 필요하다. 이에 본 연구는 인공지능 학습용 데이터 품질관리 가이드라인에서 제시한 데이터 품질관리 특성을 준용하여 딥페이크 영상 데이터셋의 품질 특성에 맞게 수정 적용하였다. 또한 딥페이크 영상 데이터셋이 지향점을 고려하여 '일반화 가능성'도 추가하였다(<Table 3> 참고).

본 연구는 제안된 품질 기준 적합성을 중심으로 평가기준 도출을 위한 구체적인 접근방법에 대해 논의한다. 이를 위해 본 연구에서는 데이터 품질에 영향을 미치는 요인을 중심으로 접근하고자 한다. 데이터 품질에 영향을 미치는 요인에는 부(-)(negative)의 요인과 정(+)(positive)의 요인이 존재한다. 부(-)의 요인은 데이터 품질을 저하시키는 요인이며 본 연구에서는 기존 연구의 한계점으로 정의된다. 정(+)의 요인은 데이터 품질을 향상시키는 요인이며 본 연구에서는 품질 개선을 위한 선행연구의 노력들로 정의된다. 딥페이크 데이터 벤치마크 연구에서 보여지는 노력들은 매우 다양하며 각 연구에서 주요 공헌점으로 제시되고 있다. 따라서 평가기준 도출시 유용할 것으로 기대된다.

<Table 3> Fitness of Deepfake Video Dataset

| Quality Characteristics | Contents |
|---|---|
| ① Reliability | It must be obtained from a reliable source[7, 19]. |
| ② Sufficiency | The training data should be provided in sufficient quantity to have a positive effect on the robustness and performance of the training model[7, 13, 16, 19, 22]. |
| ③ Ethics | The data must be obtained in a legal and ethical manner[2, 13, 19]. |
| ④ Diversity | In actual scenarios, various forgery and forgery techniques are used for various faces. In order to reflect the potential threat of learning data, diversity must be secured in the range where data characteristic information is useful for learning[7, 13, 19]. |
| ⑤ Reality(Potential Threat) | Deepfake image dataset is created under artificial environment and conditions. Therefore, it should contain the realism and potential threat well by reflecting the environment and characteristics like the actual scenario[7]. |
| ⑥ Generalizability | The ultimate goal of Deepfake image datasets is to help generalize detection models[13, 22, 33]. |

본 연구는 선행연구 검토를 통해 기존 연구의 한계점을 탐색하였다. 기존 연구의 한계점은 크게 3가지로 정리된다. 첫째, 선행연구에서 가장 많이 지적된 한계점은 소규모 데이터(small-scale)이다. 딥페이크 탐지 연구에서 딥러닝이 괄목할 만한 성능을 보여주면서 최근 관련 연구에서는 딥러닝을 기반으로 한 탐지 모델들이 많이 개발되고 있는 상황이다. 일반적으로 딥 뉴럴 네트워크(deep neural network)는 학습해야 할 파라미터(parameter)가 많기 때문에 학습을 위한 충분한 양이 요구된다. 그러나 데이터 양이 충분하지 못한 경우 과적합 문제(overfitting)로 인해 탐지성능이 떨어지는 문제점이 있다[32]. 이에 선행연구에서는 최근 연구의 추세를 고려할 때 소규모 데이터는 과적합 문제의 발생과 학습모델의 일반화에 악영향을 미칠 것이라고 지적하고 있다[2, 7, 13, 16, 22].

현재 GAN 기반의 영상 기술 발달과 관련 오픈 소스의 공개로 대량의 영상 제작이 가능해지면서 향후 딥페이크 데이터 벤치마크 연구에서는 종래에 비해 더욱 큰 규모의 데이터 구축이 이루어질 것으로 전망된다. 이에 데이터 품질과 관련하여 데이터 규모가 갖는 의미는 더욱 중요해질 것으로 판단된다.

둘째, 낮은 시각적 품질(low visual quality)이 한계점으로 지적된다. 기존 딥페이크 영상 데이터셋에서는 시각적 품질이 너무 낮거나 육안으로 확인될 정도의 시각적 아티팩트가 존재한다. 선행연구에서는 이러한 결함으로 인해 기존 딥페이크 영상 데이터셋이 실제 시나리오와 차이가 크고 사실성이 결여되어 있다고 지적하였다[7, 13, 16].

Yu et al.[33]는 시각적 특징을 기반으로 하는 탐지 연구들(visual feature-based detection)에서 높은 탐지성능을 연구결과로 제시하고 있지만, 실제로 잘 만들어진 딥페이크 영상은 이러한 시각적 특징이 거의 없거나 매우 적기 때문에 이는 결함에 의한 편향(bias)된 결과이며 해당 데이터셋에서 개발된 탐지 모델은 실전에서 큰 제약이 있을 것이라고 지적하였다. 물론 시각적 품질이 낮은 경우 인간의 딥페이크 판별력이 떨어져 유효할 수도 있다. 그러나 이는 영상이 너무 흐릿하거나 이미지가 뭉개져 육안 판별이 어려움에 연유한 결과임으로 실제로 잘 만들어진 딥페이크 영상만이 우리사회에 큰 파급효과가 있다는 점을 고려하면 낮은 시각적 품질은 데이터 품질 관점에서 중요한 의미를 갖지 않을 수 있다[22].

셋째, 등장인물의 콘텐츠 동의권 미확보(non contents agreeing subjects)가 한계점으로 지적된다. 1세대 데이터셋의 공개 이후 등장인물의 공식적 동의가 없으면 초상권 문제와 이로 인한 데이터 가용성에 제약이 있을 것이라고 지적되어 오면서 2세대 이후부터는 실험환경 조성을 통한 영상 제작이 해결방안으로 떠오르게 되었다[7, 13]. 원래 실험환경의 조성은 등장인물의 동의권 확보를 목적으로 시작되었지만, 실험환경이 실제 시나리오와 얼마나 유사하게 조성될 수 있는지 주요 이슈로 인식되면서 향후 데이터셋 구축 전략과 방향전환의 계기가 된다는 점에서 의의가 있다.

이상으로 기존 연구의 한계점에 대해 살펴보았다. 다음에서는 데이터 품질 향상을 위한 선행연구의 노력들을 고찰한다. 선행연구의 노력들은 크게 3가지로 정리된다.

첫째, 제작과정 및 실험환경의 통제 여부(control)를 들 수 있다. 실험환경이 실제 시나리오와 얼마나 유사한지가 중요시되면서 제작과정과 실험환경 통제 여부는 데이터 품질에 영향을 미치는 중요한 요인이 된다. 2세대부터 최근까지 실험환경 조성을 통한 영상 제작이 딥페이크 데이터 벤치마크 연구에서 주요 연구트렌드가 되었지만, 몇몇 선행연구에서는 실제 시나리오와 같은 자연스러운 데이터 분포를 위해 인터넷에서 영상을 가져와 얼굴합성 작업을 하는 노력도 존재한다[16, 35]. 또한 원본 영상 제작시 여러 각도에서 카메라 및 조명을 설정하여 실험참가자에게 포즈, 제스처, 감정표현을 자연스럽게 또는 대본을 따라서 표현하라고 요청하는 등 실험환경의 통제를 통해 다양한 연출을 하고자 노력하였다[7, 13].

둘째, 선행연구에서는 데이터셋의 가용성과 일반화를 높이기 위해 다양성(diversity)을 확보하고자 노력하였다. 데이터 설계시 남녀 성비 및 인종(흑, 황, 백, 갈)을 균등하게 조정하여 인물의 다양성을 확보하고자 하였으며, 야외에서 자연스러운 배경을 토대로 촬영을 진행하거나 전문 스튜디오에서 촬영을 진행하는 등 배경의 다양성도 확보하고자 노력하였다[2, 7, 13]. 또한 잠재적인 위협성을 반영하기 위해 다수의 딥페이크 기법을 사용하거나 변조의 유형과 강도를 다양하게 적용하는 등 변조방법의 다양성도 함께 확보하고자 노력하였다.

셋째, 선행연구에서는 딥페이크 영상 데이터셋을 구축하고 일반화가 가능할지 타당성 있는 근거를 제시하고자 노력하였다. 그 일환이 바로 자제검증(validation)이다. 선행연구의 자체검증에서는 선행연구의 탐지 모델(예: XceptionNet 등)을 적용하여 탐지성능을 제시하거나 데이터셋 간 탐지성능을 비교하는 방식이 가장 많이 채택되었다[13, 16].

선행연구에서의 탐지 모델 개발과정과 탐지성능 평가방법은 연구에 따라 다르며 차별점이 있다. Jiang et al.[7]은 데이터 품질 평가시 학습셋 및 테스트셋뿐만 아니라 별도의 은닉 테스트셋(hidden test set)을 두어 검증을 진행하였고, Dolhansky et al.[2]는 학습셋에서 탐지 모델을 만들고 테스트셋을 공개 테스트(public test) 및 비공개 테스트(private test)로 구분하여 검증을 진행하였다. Kwon et al.[13]은 동일한 데이터셋에서 탐지 모델을 만들어 검증을 진행하고, 또한 학습셋과 테스트셋을 다르게 하여 검증한 다음, 이들 결과를 비교하였다. 일반적으로 딥페이크 영상 데이터셋은 학습셋 및 테스트셋 간의 분포가 유사하기 때문에 이로 인해 편향이 발생할 수 있다. 따라서 선행연구들은 이와 같은 노력들을 통해 극복하고자 하였다[13, 21].

선행연구에서는 탐지성능 평가방법 이외에 여러 방식을 통해 데이터 품질을 평가하였다. Roccler et al.[22]은 딥페이크 영상을 사람이 눈으로 직접 확인하여 딥페이크 여부를 판별하는 과정을 거쳤고, Li et al.[16]는 원본과 딥페이크 영상의 구조적 유사성을 조사하기 위해 SSIM(structural similarity index measure)을 활용하기도 하였다.

# 3. 연구방법

## 3.1 중요 평가항목 도출

앞 장에서는 딥페이크 영상 데이터셋의 품질 기준 적합성을 제안하고 기존 연구의 한계점과 품질 개선을 위한 선행연구의 노력들에 대해 고찰해 보았다.

본 연구는 선행연구 검토를 토대로 딥페이크 영상 데이터셋의 품질 기준 적합성에 부합하는 구체적인 평가항목을 도출한다. 본 연구에서 도출된 평가항목은 얼굴표현방법, 데이터 특성, 변조방법, 자체검증, 결함으로 구성된다 (<Table 4> 참고).

<Table 4> Matching of Fitness and Evaluation Criteria

| Evaluation Criteria | | Fitness |
|---|---|---|
| Facial Expression | Identity Diversity | Diversity |
| | Expression Diversity | Diversity |
| | Pose Diversity | Diversity |
| | Illumination | Diversity |
| Data Characteristics | Source | Reliability, Ethics |
| | Scale | Sufficiency |
| | Video Length | Sufficiency |
| | Visual Quality | Sufficiency |
| | Deepfake Ratio | Sufficiency |
| | Image Compression Method | Reality, Generalizability |
| Perturbation Method | Perturbation Diversity | Diversity, Reality (Potential Threat) |
| | Deepfake Technique Diversity | Diversity, Reality (Potential Threat) |
| | Image pre-Proccesing | Reality, Generalizability |
| | post-Proccessing | Reality, Generalizability |
| Validation | Detection Rate | Generalizability |
| | Eyes Level Assessment | Generalizability |
| | Image Quality Assessment | Generalizability |
| Defect | Geometry Defect | Reality(Potential Threat) |
| | Disproportionate Shadows | Reality(Potential Threat) |
| | Incomplete Light Reflection | Reality(Potential Threat) |
| | Color Mismatch | Reality(Potential Threat) |

Note: Left is Evalution Criteria, and right matches Fitness to Evaluation Criteria.

### 3.1.1 얼굴표현방법(Facial Expression)

얼굴은 사람의 개성과 정체성을 나타내고 표정과 포즈를 통해 그 사람의 의사를 전달하는 커뮤니케이션 역할을 한다. 얼굴표현은 그 사람만의 고유하고 독특한 정보를 반영하고 있다. 또한 딥페이크 영상에서 실제 보여지는 부분이 얼굴이므로 인간의 식별영역에 있는 평가기준이기에 중요하다고 볼 수 있다. 얼굴표현방법의 하위항목은 인물의 정체성을 나타내는 객체다양성, 감정표현을 나타내는 표정다양성, 인물이 취하고 있는 포즈를 나타내는 포즈다양성, 그리고 영상의 배경을 담당하는 조명기법으로 구성된다.

(1) 객체다양성(Identity Diversity)

- 누구인지 식별할 수 있는 얼굴을 나타내며 인물이 갖고 있는 연령, 성별, 인종 등의 얼굴정보를 포함

한다. 객체다양성의 평가기준으로 데이터에서 등장하는 얼굴 수, 성별, 인종 등 빈도나 분포 등을 통해 나타낼 수 있다.

(2) 표정다양성(Expression Diversity)
- 감정표현을 의미하며 중립, 분노, 행복, 슬픔, 놀라움, 경멸, 혐오 등의 얼굴표현으로 나타난다. 대상자가 자연스러운 감정표현을 하는지, 또는 정형화된 감정표현 대본에 따른 감정표현인지의 여부로 구분될 수 있다. 표정다양성의 평가기준은 감정상태의 통제 여부이다. 영상 제작시 실험참가자가 자연스로운 표정을 짓도록 하는지, 과도한 표정이나 경직된 표정을 통제하기 위해 대본에 의해 표정을 짓도록 하는지 등이 예로 볼 수 있다.

(3) 포즈다양성(Pose Diversity)
- 대상자가 취하고 있는 자세나 포즈를 의미한다. 포즈다양성의 평가기준은 포즈 상태에의 통제 여부이다. 실험참가자의 포즈가 통제되었는지 또는 자연스럽게 포즈를 취하고 있는지로 구분된다.

(4) 조명기법(Illumination)
- 조명에 따라 얼굴 윤곽, 눈, 코, 볼 등에서 반사광이나 그림자가 발생함으로써 시각 특징(visual feature)에 큰 영향을 줄 수 있다. 특히 얼굴 변조 시 입사조명이 잘못되거나 부정확한 추정이 이루어질 경우, 음영 아티팩트를 초래하여 데이터 품질이 떨어지는 결과를 초래한다. 조명기법은 대상자에게 일관되게 비추어져 있는지, 상하 좌우 등 여러 각도에서 조명을 비추고 있는지, 또는 자연광인지, 조명 모델을 통해 인위적으로 생성한 것인지의 여부 등 자연광 및 조명설정 여부가 평가기준이 된다.

### 3.1.2 데이터 특성(Data Characteristics)

딥페이크 영상 데이터셋의 제작과정, 규모, 화질수준 등 물리적인 데이터 특성을 반영한 평가기준이다. 데이터 특성의 하위항목은 생성방법, 데이터규모, 단위영상길이, 화질수준, 딥페이크비율, 영상압축으로 구성된다.

(1) 생성방법(Source)
- Dolhansky et al.[2]은 원본 영상의 품질이 딥페이크 영상 데이터셋의 품질을 결정하는 중요한 요인이라고 하였다. 1세대 데이터셋들은 원본 영상을 YouTube에서 가져와 딥페이크 영상 제작을 진행하였는데, 선행연구에서는 그 결과물에 대해 YouTube 영상의 얼굴교환 결과가 상대적으로 부자연스럽거나 너무 인공적이라고 평가하였다. 이와 대비하여 Roccler et al.[18]이나 Li et al.[16]는 실제 시나리오의 특성을 반영하기 위해 원본 영상을 모두 YouTube에서 가져오기도 하였다. 이처럼 영상의 제작과정과 생성방법이 어떠한 과정으로 이루어졌는지가 딥페이크 영상 데이터셋 품질에 중요한 영향을 미칠 것으로 판단된다. 생성방법의 평가기준은 인터넷 출처 여부, 직접 제작 여부, 이 둘을 혼합한 경우의 여부가 된다.

(2) 데이터규모(Scale)
- 많은 선행연구에서는 데이터의 규모가 클수록 고품질로 인식하고 있다. 특히 최근 탐지 모델들이 딥러닝을 기반으로 하고 있기 때문에 학습 데이터는 충분한 양이 요구된다. 또한 소규모 데이터의 경우 과적합 문제나 탐지성능이 떨어지는 문제점을 낳기도 한다. 데이터규모의 평가기준은 영상수, 클립수(clip), 프레임수(frame) 등으로 나타낼 수 있다.

(3) 단위영상길이(Video Length)
- 평균 영상당 재생시간을 의미한다. 딥페이크 영상을 제작하는 과정에서 편집과정을 거치게 된다. 이때 대부분의 영상들은 필요한 부분만 추출하여 클립 형태로 만들어진다. 단위영상길이는 데이터 크기와 비례한다. 단위영상의 길이가 너무 짧거나 길면 학습비용이 커지거나 과적합 문제가 발생하기 때문에 균형있는 영상길이가 요구될 것이다. 단위영상길이의 평가기준은 재생시간이며 일반적으로 초단위로 나타낸다.

(4) 화질수준(Visual Quality)
- 화질수준은 해상도, 초당 프레임수(FPS), 비트레이트(bit rate) 등으로 구성되며 이 값들이 높을수록 영상은 고화질로 분류된다. 딥페이크 영상의 화질수준이 높을수록 보는 이로 하여금 시각적 품질이 좋게 인식될 수 있으며, 높은 화질수준에서 다운그레이드가 쉽기 때문에 입력 대상 얼굴을 수용할 때 크기 조정이나 회전 등 작업에 용이하다. 다만 현실에서 딥페이크 영상은 고화질과 저화질 모두 존재하며, 특히 사람이 눈으로 딥페이크 여부를 판별하고자 할 때 저화질에서 인간의 판별력이 떨어질 수 있음을 고려할 필요가 있다. 또한 화질수준이 높을수록 학습모델의 시간적 비용도 높아지기 때문에 화질수준과 학습비용의 균형을 요하게 된다.

(5) 딥페이크비율(Deepfake Ratio)
- 전체 영상수에서 원본과 딥페이크 영상수의 비율을

의미한다. 가령, 1(실제) : 10(딥페이크) 비율로 나타낼 수 있다.

(6) 영상압축(Image Compression Method)
- 영상압축과 양자화 파라미터는 영상의 품질을 좌우하는 항목이다. 일반적으로 양자화 파라미터(QP) 값을 높이면 영상압축률이 높아지며 저품질로 분류된다. 영상압축의 평가기준으로는 압축이 어떻게 이루어졌는지(H.264, HEVC 등)와 양자화 파라미터 값(예: 고화질은 23, 저화질은 40으로 설정)으로 나타낸다.

### 3.1.3 변조방법(Perturbation Method)

선행연구에서는 실제 딥페이크의 변조와 같은 효과를 반영하기 위해서 변조방법이 다양화될 것을 강조하고 있다[7, 13]. 다양한 변조방법이 포함된 데이터셋에서 탐지 모델이 개발되어야 모델의 강건함(robustness)과 성능을 보장할 수 있을 것이며, 궁극적으로 탐지 모델의 일반화가 가능해져 딥페이크 영상 데이터셋의 품질 기준 적합성에 도달할 수 있을 것이다. 이에 변조방법은 평가기준으로 채택된다. 변조방법의 하위항목은 변조방식의 다양성, 테크닉(기법)의 다양성, 영상전역처리 여부, 후처리 여부로 구성된다.

(1) 변조방식의 다양성(Perturbation Diversity)
- 잠재적인 변조 정도가 데이터에 잘 반영되어 있는지를 나타낸다. 변조방식에는 복제 이동(copy & moving), 스플라이싱(splicing), 객체 제거(object remove), 모핑(morping) 등 수많은 방식들을 포함한다. 평가기준은 얼마나 많은 변조방식이 사용되었는지의 여부로 정의된다.

(2) 테크닉(기법)의 다양성(Deepfake Technique Diversity)
- 테크닉의 다양성은 변조에 사용된 구체적인 기법이 무엇인지를 의미한다. 초기 딥페이크 영상 데이터셋에서는 단일한 기법에 의존하여 딥페이크 영상을 만든 경우가 대부분이다. 이에 선행연구에서는 실제 시나리오와 같은 위협성을 반영하기 위해서 다양한 기법을 사용하여 딥페이크 영상제작이 이루어질 필요가 있음을 강조하였다. 최근 컴퓨터 비전 분야에서 GAN 기반의 기술이 괄목할 만한 성과를 보이면서 악성 딥페이크 영상들도 GAN 기반으로 제작되어지고 있다. 테크닉(기법)에는 컴퓨터 그래픽 기반 접근방식(예: Face2Face, FaceSwap)과 학습 기반 접근 방식(예: Deepfakes, NeuralTextures) 등 다수의 기법들이 존재한다[25, 26]. 평가기준은 얼마나 많은 기법을 활용하였는지의 여부로 정의된다.

(3) 영상전역처리 여부(Image pre-Proccesing)
- 영상의 시각적 품질을 높이기 위해서는 원본 및 타겟 영상의 품질이 중요하다. 이를 위해 색상 채도 변경, 색상 대비 변경, 가우스 블러(Gaussian Blur), 명암 변화, 잡음(noise) 추가 등 다양한 영상 전역처리를 통해 품질 개선이 이루어질 수 있다. 딥페이크 영상 데이터셋 설계시 어떠한 영상 전역 처리과정을 거쳤는지 구체적인 기술(記述)이 필요하며 이에 대한 평가가 중요해질 것으로 판단된다. 평가기준은 결과물을 위해 어떠한 과정의 영상처리를 하였는지에 대한 여부와 그 과정의 적절성으로 정의된다.

(4) 후처리 여부(post-Proccessing)
- 최근 출시된 딥페이크 영상 데이터셋들은 데이터 규모가 큰 것이 특징이다. 또한 대부분의 탐지 연구들에서 딥러닝을 기반으로 모델링하기 때문에 충분한 양의 데이터가 요구된다. 따라서 앞으로 출시될 딥페이크 영상 데이터셋들도 대규모일 것으로 전망된다. 대규모 데이터의 경우 인공지능 기법 등을 활용한 컴퓨터 연산에 의해 제작될 가능성이 크다. 이에 따라 얼굴합성이나 얼굴교환 후 예기치 못한 품질의 문제가 발생할 수 있다. 시각적 품질을 높이기 위해 후처리가 필요하지만 대규모 데이터에 적용하여 시간이나 노력 등 적은 비용으로 가성비 있는 품질을 만들어 내는 것이 향후 과제로 남을 것이다. 이에 후처리 여부는 평가기준으로 채택될 수 있다. 후처리는 필터링(예: 엠보싱(embossing), 블러링(blurring), 샤프닝(sharpening) 등), 잡음 제거(예: 가우시안 잡음 분포(Gaussian noise model), 에지 보전 잡음 제거 필터(edge-preserving nosie removal filter), 양방향 필터(bilateral filter), 미디언 필터(median filter) 등), 그레이스케일(grayscale) 변환 등을 예로 들 수 있다. 평가기준은 어떠한 후처리 과정이 이루어졌는지에 대한 여부와 시간비용, 노력 등이 고려되어 가성비가 높은 작업인지, 또한 과정이 적절하였는가 등으로 정의된다.

### 3.1.4 자체검증(Validation)

자체검증은 데이터셋 구축 후 딥페이크 영상 데이터셋의 품질 기준 적합성을 달성할 수 있는지 그리고 타당성 있는 근거를 제시하고 있는지에 대한 내용을 포함한다. 자체검증의 하위항목은 탐지율, 육안식별수준, 이미지측정평가로 구성된다.

(1) 탐지율(Detection Rate)
- 선행연구에서 데이터셋의 일반화의 평가지표로 가장 많이 제시된 것이 탐지율이다. 선행연구에서는 기존의 탐지 모델을 적용하여 만들어진 데이터셋에 적용하여 탐지성능을 평가하는 방식으로 자체검증이 이루어졌다. 이 밖에도 데이터셋 간의 탐지성능 비교 등 수많은 방식들이 존재한다. 일반적으로 탐지율은 ROC 커브나 정확도(accuracy) 등 수량적 지표를 통해 제시되기 때문에 정량적 평가가 가능한 장점이 있다. 또한 연구자는 탐지율을 통해 데이터셋 설계가 딥페이크 영상 데이터셋의 품질 기준 적합성에 부합하는지를 검토할 수도 있다.

(2) 육안식별수준(Eyes Level Assessment)
- 딥페이크 영상을 눈으로 확인하고 그 여부를 구별 및 판단하는 과정으로 정의된다. 딥페이크 판단 여부를 결정하는 사람은 전문가 평가나 일반인 평가로 구분될 수 있다. 이 항목은 데이터 가용성과 자동화된 딥페이크 탐지 방법의 효용성을 나타낼 기준이 될 수 있다. Rossler et al.[22]의 연구에서는 영상의 품질이 낮을수록 사람의 딥페이크 탐지율이 낮아진다는 결과가 제시되기도 하였다. 평가기준은 딥페이크 영상을 눈으로 확인하고 얼마나 실제와 유사한가로 정의된다. 평가방식은 리커드 척도를 활용하거나 판별인원의 기준선 이상(가령, 50/100명 이상) 등의 방식을 이용할 수 있다.

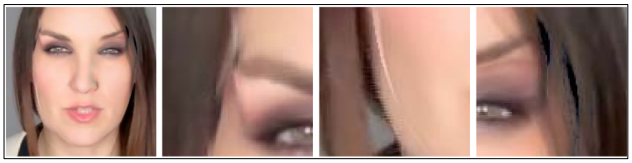
(3) 이미지측정평가(Image Quality Assessment, IQA)
- 이미지측정평가는 정량적 측정도구를 활용하여 데이터 품질을 평가하는 것을 말한다. 원본과 딥페이크 영상의 구조적 유사성을 측정하여 연구자가 정한 기준에 있으면 이미지측정평가가 높은 것으로 평가될 수 있다. 평가기준으로 SSIM(Structural Similarity Index Measure), MS-SSIM, IW-SSIM, CW-SSIM, FSIM, GSM, AKD(Average Keypoint Distance), FRQA, VIF(Visual Information Fidelity), NRQA, 선명도, Blur Matrix, BIQI, NIQE, PSNR 등 다양한 측정도구를 통해 평가가 가능하다.

### 3.1.5 결함(Defect)

결함은 데이터 품질과 데이터 가용성을 떨어뜨리는 요인이다. 따라서 결함은 데이터 품질을 위해 완전한 해결이 요구된다. 본 연구에서는 DF-TIMIT와 DFDC 데이터셋에서 발견되는 결함의 예를 <Figure 1>로 나타내었다. DF-TIMIT의 얼굴사진을 보면 합성경계면 사이에 시각적 아티팩트가 다수 발견되는 것을 볼 수 있다. 또한 DFDC에서는 시각적 아티팩트뿐 아니라 기하학적인 결함으로 인해 얼굴을 인식하지 못할 만큼의 저품질을 보여주기도 한다. 결함은 여러 유형으로 존재한다. 본 연구에서 결함의 하위항목은 기하학적 결함, 불균형 그림자 여부, 반사광 미처리, 색상불일치로 구성하였다.


<Figure 1> Examples of Defect

(1) 기하학적 결함(Geometry Defect)
- 얼굴 변조를 위해서는 얼굴 기하학을 추정해야한다. 기하학적 추정이 완벽하게 이루어지지 않으면 코, 얼굴 윤곽 등에서 아티팩트가 발생하며, 머리카락으로 가려진 일부 부위에서는 '구멍'이 난 것처럼 아티팩트가 발생하게 된다. 평가기준은 기하학적 결함 여부이다.



Sorce: FaceForensics++[22].

<Figure 2> Example of Geometry Defect

(2) 불균형 그림자 여부(Disproportionate Shadows)
- 콧구멍 등 얼굴 일부분에 그림자가 있어야 할 부분에 그림자가 없거나 불균형하게 그림자가 남아있는 경우를 말한다. 평가기준은 불균형 그림자 여부이다.

Sorce: FaceForensics++[22].

<Figure 3> Example of Disproportionate Shadows

(3) 반사광 미처리(Incomplete Light Reflection)
- 이마, 볼, 턱, 코 등 변조 후 반사광이 나타나지 않는 경우를 말한다. 평가기준은 반사광 미처리 여부이다.

(4) 색상불일치(Color Mismatch)
- 기존 딥페이크 영상 데이터셋에서 색상불일치가 많이 발견되어 데이터의 가용성 문제가 제기되었다. 홍채색이 다르게 나타나는 현상을 이색성(heterochromia)라 하는데 이는 실제로 굉장히 희박한 확률로 나타난다. 딥페이크 영상에서 색상불일치의 심각도는 다양하기 때문에 식별 가능한 색상불일치의 여부는 평가기준이 될 수 있다.

Sorce: (top) Karras et al.[8], (bottom) Wang et al.[30].

<Figure 4> Examples of Color Mismatch

<Table 5> Deepfake Dataset Evaluation Criteria

| Primary Evaluation criteria | Secondary Evaluation criteria | Factor | Evaluation Method | Rationale | Reference |
|---|---|---|---|---|---|
| ① Facial Expression | Identity Diversity | number of fake | frequency | - Confirm the personality or identity of the face | [2, 7, 13] |
| | | skin color | frequency, Gini's coefficient, entropy | | |
| | | age | frequency, distribution | | |
| | | gender | frequency, ratio, Gini's coefficient | | |
| | Expression Diversity | | emotional state | - emotional expression(neutral, anger, happiness, sadness, surprise, contempt, etc.)<br>- neutral expression, use the script | [7, 13] |
| | Pose Diversity | | pose state | - posture, position, pose | [7] |
| | Illumination | | natural light, illumination setting | - Reflected light/shadow caused by lighting<br>- This affects the visual feature | [7, 13] |
| ② Data Characteristics | Source | - actor<br>- source | internet, create, mix | - Original video quality have a decisive effect on Deepfake video dataset quality | [2] |
| | Scale | - number of total videos<br>- total play time | number of videos/clips/frames | - Detection performance depend on data scale<br>- In prior studies, the large scale have been perceived as high quality dataset | [2, 7, 16] |
| | Video Length | average play time per video | unit: second | - Meaning of average play time per video<br>- play time in seconds | [2, 16] |
| | Visual Quality | resolution | 1920×1080 | - The higher the value of the relevant item, such as resolution, the higher the quality<br>- The higher the resolution of the synthetic face, the better the visual quality, and the ease of operation such as resizing and rotating when accommodating the input target face<br>- Consideration of time cost of learning model | [2, 16] |
| | | number of frame per second | 4fps, 8fps etc. | | |
| | | bitrate | 192kbps per sec, 26Mbps per sec, | | |
| | Deepfake Ratio | ratio of Real/Fake | ratio(%) | - Ratio of original and Deepfake videos<br>ex) 1(real) : 10(Deepfake) | - |
| | Image Compression Method | image compression technique/level | - What image compression technique was used<br>- quantization parameter | - Items that influence video quality<br>- The higher the quantization parameter(QP), the lower the quality | [16, 22] |
| ③ Perturbation Method | Perturbation Diversity | | copy & moving, splicing, object removal, etc. | - Items that indicate whether potential threats are well reflected<br>- Whether various perturbation methods are adopted | [17] |
| | Deepfake Technique Diversity | | - What technique was used<br>- DFAE, FSGAN, NTH, StyleGAN etc.<br>- Number of techniques | - Items that indicate whether potential threats are well reflected<br>- Adoption of various Deepfake techniques | [2, 7, 16] |

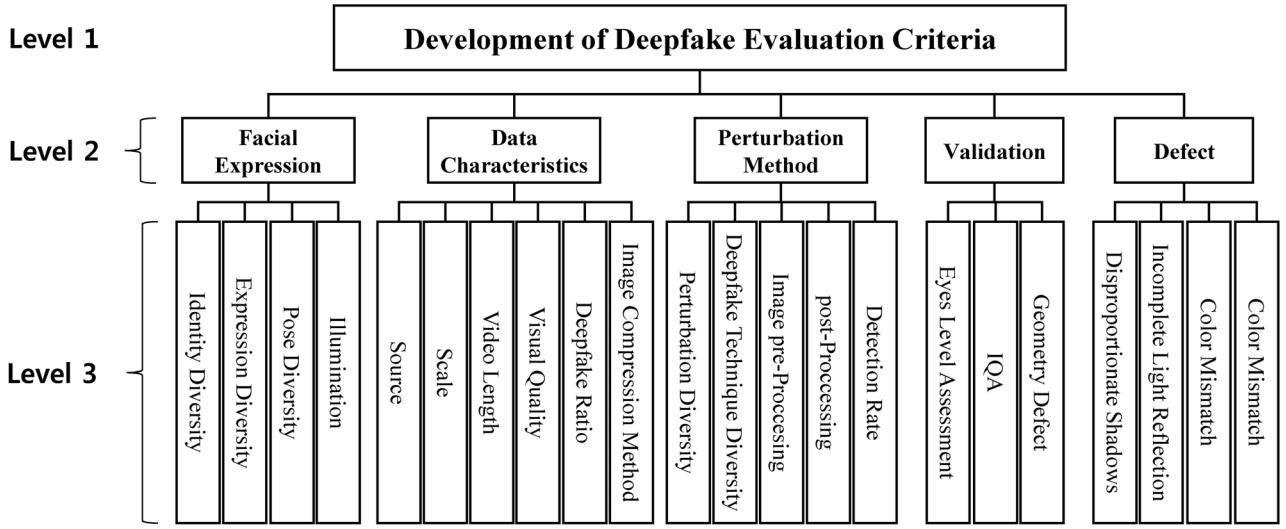
<Table 5> Deepfake Dataset Evaluation Criteria(Continued)

| Primary Evaluation criteria | Secondary Evaluation criteria | Factor | Evaluation Method | Rationale | Reference |
|---|---|---|---|---|---|
| ③ Perturbation Method | Image pre-Proccesing | saturation | True / False | - Whether to pre-process images for natural Deepfake images | [16] |
| | | hue | True / False | | |
| | | Gaussian Blur | True / False | | |
| | | brightness | True / False | | |
| | | add noise | True / False | | |
| | post-Proccessing | fittering | embossing blurring, sharpening, etc. | - Whether to post-process images for natural Deepfake images | [2] |
| | | elimination noise | Gaussian noise model, edge-preserving nosie removal filter, etc. | | |
| | | grey scale change | | | |
| ④ Validation | Detection Rate | (Based on existing research on the scope of actual evaluation) | detetion ratio(%), ROC, ACC, etc. | - Detection performance results by applying the detection model of previous studies | [7, 12, 16, 22] |
| | Eyes Level Assessment | expert, ordinary people | -Likert 5-point scale<br>-Above or below the standard of a certain number of people<br>ex) 50/100 or more | - A person sees a Deepfake and identifies and judges for Deepfake<br>- Based on data availability and effectiveness of automated Deepfake detection methods | [7, 13, 16, 22] |
| | Image Quality Assessment | | SSIM, MS-SSIM, IW-SSIM, CW-SSIM, FSIM, GSM, AKD, FRQA,BIQI, NIQE, PSNR, etc. | - Quantitative assessment<br>- If the structural similarity between original and Deepfake is measured and it is within the criteria set by the researcher, the IQA is evaluated as high | [13, 16] |
| ⑤ Defect | Geometry Defect | | True / False | - For face synthesis, it is necessary to estimate the face geometry<br>- If the geometric estimation is not made perfectly, artifacts occur in the nose and face contours, and in some areas covered with hair, artifacts occur as if there was a 'hole' | [16, 18] |
| | Disproportionate Shadows | | True / False | - It refers to a case where there is no shadow or the shadow remains disproportionately on the part of the face where there should be a shadow, such as the nostrils | [16] |
| | Incomplete Light Reflection | | True / False | - Refers to a case in which reflected light does not appear after modulating the forehead, cheeks, chin, nose, etc | [16] |
| | Color Mismatch | | True / False | - Many color inconsistencies are found in the existing dataset, which raises many problems with the availability of the data<br>- Since the severity of color mismatch in Deepfake images varies, whether or not there is an identifiable color mismatch is an evaluation criteria | [16, 18] |

## 3.2 평가기준 고도화

본 연구에서는 딥페이크 영상 데이터셋 품질 강화를 위한 기준 수립을 위해 선행연구를 검토하여 품질 기준 적합성에 따른 평가항목을 도출하였다.

본 연구는 도출된 평가항목에 대한 평가기준 고도화를 위해 전문가 의견을 수렴하고자 한다. 전문가 의견 수렴에는 AHP 기법을 활용한다.




<Figure 5> AHP Structure

AHP(Analytic Hierarchy Process)는 의사결정 사안이 복잡하고 여러 평가기준으로 이루어졌을 때 의사결정 사안을 계층화하여 의견을 수렴하는 의사결정기법이다[23]. AHP는 목표를 설정하고 의사결정의 대안들을 위계적 구조로 계층화한다. 각 계층에서는 대안들 간의 쌍대비교(pairwise comparison)가 이루어지고 가중치를 산출함으로써 대안의 우선순위 또는 중요도를 도출할 수 있다. AHP는 평가의 일관성 파악이나 의견수렴이 매우 용이해 다양한 실무분야에서 활용되고 있다[10, 31]. 이에 따라 AHP 기법이 본 연구의 방법으로 적절할 것으로 판단된다.

본 연구에서 도출된 평가항목은 <Figure 5>와 같이 AHP 계층화하였다. 1계층에서는 딥페이크 영상 데이터셋 평가기준 개발을 목표로 설정하였고, 2계층에서는 얼굴표현방법, 데이터 특성, 변조방법, 자체검증, 결함으로 구성하였다. 3계층에서는 2계층의 하위요인들로 구성하였다.

## 4. 분석결과

본 연구는 딥페이크 영상 학습을 위한 데이터셋 평가기준을 개발하기 위해 딥페이크, 컴퓨터 비전, 기계학습/AI 등 관련 경험 및 학식을 갖춘 전문가들에게 설문지를 배부하였다. 설문기간은 2021년 9월부터 10월까지 약 2개월에 걸쳐 진행하였고 총 20부의 응답지를 받았다. 응답의 일관성을 위해 CR(Consistency Ratio)값이 Level 1 CR값이 0.2 이상인 응답지는 분석에서 제외하였고, Level 2 CR값이 0.2 이상이 둘 이상 존재하는 응답지도 분석에서 제외하였다(<Table 6> 참고)[10]. 본 연구에서는 총 16부의 응답지만이 AHP 분석에 활용되었다.

<Table 6> Consistency Ratio

| Respo-ndent | Level 1 | Level 2 | | | | |
|---|---|---|---|---|---|---|
| | Deepfake Dataset Evaluation criteria | Facial Expres-sion | Data Charac-teristic | Perturb-ation Method | Valid-ation | Defect |
| 1 | 0.007 | 0.040 | 0.114 | 0.029 | 0.000 | 0.014 |
| 2 | 0.010 | 0.063 | 0.020 | 0.004 | 0.017 | 0.000 |
| 3 | 0.048 | 0.137 | 0.133 | 0.071 | **0.281** | 0.029 |
| 4 | 0.114 | 0.260 | 0.083 | 0.044 | 0.196 | 0.044 |
| 5 | 0.092 | 0.183 | 0.126 | 0.062 | 0.000 | **0.261** |
| 6 | 0.114 | 0.260 | 0.083 | 0.044 | 0.196 | 0.044 |
| 7 | 0.104 | 0.067 | 0.106 | 0.079 | 0.129 | 0.046 |
| 8 | 0.135 | 0.151 | 0.080 | 0.060 | 0.023 | 0.066 |
| 9 | 0.048 | 0.137 | 0.136 | 0.071 | **0.281** | 0.029 |
| 11 | 0.070 | **0.649** | 0.068 | 0.016 | 0.028 | 0.037 |
| 12 | 0.073 | 0.040 | 0.182 | **0.276** | 0.004 | 0.076 |
| 13 | 0.078 | 0.168 | 0.077 | 0.000 | 0.017 | 0.000 |
| 14 | 0.095 | 0.028 | 0.086 | 0.000 | 0.051 | 0.070 |
| 17 | 0.105 | 0.058 | **0.294** | 0.121 | 0.095 | 0.015 |
| 18 | 0.116 | 0.162 | 0.090 | 0.012 | 0.129 | 0.000 |
| 19 | 0.092 | 0.100 | 0.073 | 0.185 | 0.000 | 0.058 |

본 연구에서는 3계층 가중치 산출시 CR값이 0.2 이상인 항목은 제외하여 계산하였다. 예를 들면, 얼굴표현방법의 3계층의 경우 11번 응답지의 CR값이 0.649로 0.2 이상인 것을 볼 수 있다. 본 연구는 얼굴표현방법의 하위요인에 대해 모두 가중치를 산출하고 최종적으로 얼굴표현 방법 가중치 계산을 위해 산술평균을 하였는데, 이때 11번 응답

지만 제외하여 산술평균하였다. 나머지 모든 3계층의 요인에 대해서도 같은 방식을 적용하였다.

응답자의 인구통계학적 특성은 다음과 같다. 먼저 직무 관련성 및 경험을 살펴보면, 기계학습/AI 분야가 56.3%로 가장 많은 비중을 차지하였고 그 다음으로 영상처리 및 컴퓨터 비전 분야 25.0%, 딥페이크 연구/활용 분야 18.3%의 순으로 나타났다. 또한 응답자들은 모두 연구직종 종사자로 나타났으며, 업력은 2년 이내가 62.5%로 가장 많았고 그 다음으로 순으로 2~5년 이내가 31.2%, 5~10년 이내가 6.3%로 나타났다.

AHP 분석결과, 2계층에서는 얼굴표현방법(1위)이 가장 중요한 것으로 나타났다. 그 다음으로 자체검증(2위), 데이터 특성(3위), 결함(4위), 변조방법(5위) 순으로 중요한 것으로 나타났다.

3계층의 각 요인별 중요도를 살펴보면, 얼굴표현방법에서는 포즈다양성(1위)과 조명기법(2위)이 중요한 것으로 나타났고, 데이터 특성에서는 생성방법(1위)과 영상압축(2위)이, 변조방법에서는 후처리 여부(1위), 영상전역처리 여부(2위)가, 자체검증에서는 탐지율(1위)과 이미지측정 평가(2위)가 결함에서는 반사광 미처리(1위), 불균형 그림자 여부(2위)가 중요한 것으로 나타났다.

본 연구의 결과는 선행연구에서 소규모 데이터나 낮은 화질수준 등에 대해 한계점으로 지적하면서 중요하게 생각해 온 것과는 대비되는 결과이다.

본 연구의 결과를 정리하면 얼굴표현방법, 자체검증, 데이터 특성이 데이터 품질 결정에 중요한 것으로 나타났다. 얼굴은 사람의 정체성을 나타내는 중요한 정보역할을 한다. 얼굴정보는 실험환경 조성시 성별, 연령, 인종을 구성할 때 이를 결정하는 기준이 되며 표정과 포즈의 연출을 통해 시청자에게 무엇을 보여주려는지 제작자의 의도를 표현할 수 있다. 또한 딥페이크 영상 제작시 얼굴과 표현방법은 원본 영상의 품질을 결정짓고 이는 최종적으로 딥페이크 영상 데이터셋의 품질을 결정짓게 된다. 본 연구에서 전문가 의견은 이상의 내용을 고려한 결과라고 볼 수 있다.

제2장 선행연구 검토에서 다양성의 확보가 중요하며 이를 강조해 왔음을 확인하였다. 이는 탐지모델의 강건성을 높이고 종국적으로 데이터셋의 일반화를 도모하기 위한 것이다. 본 연구에서 얼굴표현방법은 딥페이크 영상 데이터셋 품질 기준 적합성에서 다양성과 연관이 높다. 전문가 의견은 선행연구와 같이 다양성 확보가 가장 중요하다고 보았으며, 이에 따라 다양성 특성을 가장 많이 반영한 얼굴표현방법이 중요하게 나타난 것으로 해석된다.

2계층에서 자체검증도 중요한 것으로 나타났다. 선행연구에서는 딥페이크 영상 데이터셋 구축 후 일반화 가능성을 평가하기 위해 타당성 있는 근거를 제시하고자 하였다.

<Table 7> AHP Analysis Results

| Level 2 | Weight | Ranking | Level 3 | Weight |
|---|---|---|---|---|
| **① Facial Expression** | **0.2400** | **1** | Identity Diversity | 0.046 |
| | | | Expression Diversity | 0.033 |
| | | | **Pose Diversity** | **0.088** |
| | | | Illumination | 0.079 |
| ② Data Characteristics | 0.2050 | 3 | **Source** | **0.063** |
| | | | Scale | 0.024 |
| | | | Video Length | 0.040 |
| | | | Visual Quality | 0.014 |
| | | | Deepfake Ratio | 0.016 |
| | | | Image Compression Method | 0.046 |
| ③ Perturbation Method | 0.1298 | 5 | Perturbation Diversity | 0.021 |
| | | | Deepfake Technique Diversity | 0.025 |
| | | | Image pre-Proccesing | 0.033 |
| | | | **post-Proccessing** | **0.043** |
| ④ Validation | 0.2299 | 2 | **Detection Rate** | **0.088** |
| | | | Eyes Level Assessment | 0.063 |
| | | | Image Quality Assessment | 0.080 |
| ⑤ Defect | 0.1953 | 4 | Geometry Defect | 0.030 |
| | | | Disproportionate Shadows | 0.051 |
| | | | **Incomplete Light Reflection** | **0.072** |
| | | | Color Mismatch | 0.047 |

본 연구에서 전문가들은 선행연구에서와 같이 데이터 품질 평가를 위해 타당성 있는 근거 제시가 필요하며 이에 따라 자체검증이 매우 중요하다고 의견을 모았다. 특히 탐지율과 이미지측정평가의 중요도가 높게 나온 점은 데이터 품질 평가시 과학적 방법을 토대로 한 타당한 근거 제시가 강조되는 것으로 해석할 수 있다.

데이터 특성의 3계층 하위요인을 살펴보면, 생성방법이 가장 중요한 것으로 나타났다. 전문가 의견은 신뢰 있는 출처로부터 데이터 획득 및 생성이 되어야하며 또한 이 과정에서 연구윤리를 준수하는 것도 중요하다고 보았다.

또한 데이터 특성의 3계층 하위요인에서 생성방법 다음으로 중요한 항목은 영상압축으로 나타났다. 실제로 유포되는 딥페이크 영상들은 대부분이 압축되어 있다. 실제 딥페이크의 위협을 극복하기 위해서는 탐지방법론의 개발시 영상압축에 대한 해결능력이 요구될 것이다. 따라서 연구자료로 사용되는 딥페이크 영상 데이터셋은 사실성와 일반화 가능성이 제고될 필요가 있다. 전문가 의견은 이상의 내용을 반영된 것으로 해석된다.

## 5. 결 론

최근 빅데이터 시대에 접어들면서 인공지능 학습용 데이터셋의 데이터 품질 강화와 기준 수립에 대한 논의가 중요해지고 있다. 딥페이크 악용이 사회적 이슈로 급부상하면서 딥페이크 데이터셋은 학계에 큰 공헌이 되고는 있으나, 아직 도입 단계에 있어 데이터 품질 강화와 기준 수립에 대한 논의는 없는 실정이다. 이에 본 연구는 딥페이크 영상 데이터셋의 평가기준을 마련하고자 연구를 진행하였다.

본 연구에서는 현재까지 공개된 딥페이크 영상 데이터셋을 세대별로 분류하고 각 데이터셋의 특징을 분석하였다. 또한 선행연구 검토를 통해 딥페이크 영상 데이터셋의 품질 기준 적합성을 제안하고 이에 부합하는 구체적 평가기준을 도출하고자 하였다. 이를 위해 본 연구에서는 기존 연구의 한계점과 품질 개선을 위한 선행연구들의 노력을 고찰하였으며 이를 토대로 얼굴표현방법, 데이터 특성, 변조방법, 자체검증, 결함을 평가기준으로 도출하였다. 평가기준의 고도화를 위해 AHP를 활용하여 전문가 의견을 수렴하였다.

AHP 분석결과, 얼굴표현방법, 자체검증, 데이터 특성은 데이터 품질 결정에 중요한 것으로 나타났다. 본 연구에서 전문가 의견은 영상의 품질 결정과 다양성 확보가 중요하므로 얼굴표현방법이 가장 중요하다고 보았으며, 품질평가시 과학적 방법을 기반으로 한 결과를 제시함으로써 데이터셋의 일반화 가능성을 높이는 것도 중요하다고 보았다. 또한 윤리적 문제가 제기되지 않고 신뢰 있는 출처로부터 데이터 확보와 구축이 이루어질 필요가 있다고 보았다.

선행연구에서 인공지능 학습용 데이터셋의 품질 측정에 대한 방법과 해석은 매우 다양하다. 일반적으로 실무에서 많이 활용되고 있는 지표는 한국데이터산업진흥원[11] 및 한국지능정보사회진흥원[19]에서 개발한 인공지능 학습용 데이터 품질관리 지표이다. 여기서는 데이터 품질 확보 관점의 지표(적합성, 정확성, 유효성)와 데이터 절차 관점의 지표(준비성, 완전성, 유용성)로 구성된다. 이들 지표는 데이터의 생애주기를 기반으로 각 단계별로 품질평가 지표로 제시되었다는 점에서 의의가 있다. 따라서 일반적인 데이터셋 구축시에 활용도가 높을 것으로 사료된다.

다만, 딥페이크 영상 데이터셋은 도입단계의 상황이고 데이터 구축 목적과 구체적인 평가기준이 필요함을 고려할 경우, 기존 인공지능 학습용 데이터셋의 품질관리 지표의 적용은 추상적일 수 있으며 전체 생애주기를 기반으로 평가하기에는 시기상조일 것으로 사료된다. 이러한 점에서 본 연구가 제시하는 딥페이크 영상 데이터셋 평가기준이 차별점이 있다고 할 수 있을 것이다.

본 연구에서 제시된 평가기준은 선행연구에서 고려되었던 항목들을 토대로 도출되었기에 모든 평가기준들이 품질 개선에 유효할 것으로 생각된다. 또한 적절한 데이터셋 선택 기준으로 활용되거나 딥페이크 데이터 벤치마크 설계에 참고자료로 사용할 수 있을 것으로 기대된다.

본 연구에서 제시된 평가기준은 기존 딥페이크 영상 데이터셋에 적용하지 못한 한계점이 있다. 향후 연구에서는 제시된 평가기준을 기존 딥페이크 영상 데이터셋에 적용하여 각 데이터셋의 장단점을 평가하고, 딥페이크 연구에 사용될 때 어떠한 부분에서 함의점이 있을지 고찰하고자 한다.

## Acknowledgement

This research was supported by the research fund of Hanbat National University in 2021.

## References

[1] Dolhansky, B., Howes, R., Pflaum, B., Baram, N., and Ferrer, C. C., The Deepfake detection challenge (dfdc) preview dataset, *arXiv preprint arXiv:1910.08854.*, 2019.

[2] Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., and Ferrer, C.C., The Deepfake detection challenge (dfdc) dataset, *arXiv preprint arXiv:2006.07397*, 2020.

[3] Đorđević, M., Milivojević, M., and Gavrovska, A.,

Deepfake video analysis using SIFT features, In *2019 27th Telecommunications Forum (TELFOR), IEEE*, 2019, November, pp. 1-4.

[4] Dufour, N. and Gully, A., *Contributing data to deep-fake detection research*, 2019. URL https://ai.googleblog.com /2019/09/contributing-data-to-Deepfake-detection.html.

[5] Feng, K., Wu, J., and Tian, M., A Detect method for Deepfake video based on full face recognition, In *2020 IEEE International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA), IEEE*, 2020, November, Vol. 1, pp. 1121-1125.

[6] Guarnera, L., Giudice, O., and Battiato, S., Fighting Deepfake by exposing the convolutional traces on images, *IEEE Access*, 2020, Vol. 8, pp. 165085-165098.

[7] Jiang, L., Li, R., Wu, W., Qian, C., and Loy, C. C., Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection, In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2889-2898.

[8] Karras, T., Aila, T., Laine, S., and Lehtinen, J., Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[9] Kim, H. S. A study on the data quality management evaluation model. *Journal of the Korea Convergence Society*, 2020, Vol. 11, No. 7, pp. 217-222.

[10] Kim, T.G., Kim, D.H., Lee, D.J., Kim, K.T., and Moon, S.D.S., Development of Checklist to Promote Commercialization of R&D for Railway Transportation using AHP method, *Journal of the Korea Management Engineers Society*, 2017, Vol. 22, No. 1, pp. 77-87.

[11] KOREA Data Agency, *Guideline for data quality management(Ver 2.1)*, 2006.

[12] Korshunov, P., and Marcel, S., Vulnerability assessment and detection of Deepfake videos, In *2019 International Conference on Biometrics (ICB), IEEE*, 2019, June, pp. 1-6.

[13] Kwon, P., You, J., Nam, G., Park, S., and Chae, G., KoDF: A Large-scale Korean Deepfake Detection Dataset, *arXiv preprint arXiv:2103.10094*, 2021.

[14] Le, T. N., Nguyen, H. H., Yamagishi, J., and Echizen, I. OpenForensics: Large-Scale Challenging Dataset For Multi-Face Forgery Detection And Segmentation In-The-Wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10117-10127.

[15] Li, Y., and Lyu, S., Exposing Deepfake videos by detecting face warping artifacts, *ArXiv Preprint ArXiv:1811.00656*, 2018.

[16] Li, Y., Yang, X., Sun, P., Qi, H., and Lyu, S., Celeb-df: A large-scale challenging dataset for Deepfake forensics, In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3207-3216.

[17] Mahfoudi, G., Tajini, B., Retraint, F., Morain-Nicolier, F., Dugelay, J.L., and Marc, P.I.C., DEFACTO: Image and face manipulation dataset, In *2019 27th European Signal Processing Conference (EUSIPCO), IEEE*, 2019, September, pp. 1-5.

[18] Matern, F., Riess, C., and Stamminger, M., Exploiting visual artifacts to expose Deepfakes and face manipulations, In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), IEEE*, 2019, January, pp. 83-92.

[19] National Information Society Agency, *Guideline for artificial intelligence learning data quality management*, 2020.

[20] Park, G. and Kim, C.J., Quality Characteristics of Public Open Data, *Journal of Digital Convergence*, 2015, Vol. 13, No. 10, pp. 135-146.

[21] Ramadhani, K.N. and Munir, R., A Comparative Study of Deepfake Video Detection Method, In *2020 3rd International Conference on Information and Communications Technology (ICOIACT), IEEE*, 2020, November, pp. 394-399.

[22] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Nießner, M., Faceforensics++: Learning to detect manipulated facial images, In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1-11.

[23] Saaty, T.L., Axiomatic foundation of the analytic hierarchy process, *Management Science*, 1986, Vol. 32, No. 7, pp. 841-855.

[24] Tayseer, M., Mohammad, J., Ababneh, M., Al-Zoube, A., and Elhassan, A., Digital Forensics and Analysis of Deepfake Videos, In *11th International Conference on Information and Communication Systems (ICICS)*, 2020.

[25] Thies, J., Zollhöfer, M., and Nießner, M., Deferred neural rendering, *Image synthesis using neural textures. ACM Transactions on Graphics (TOG)*, 2019, Vol. 38, No. 4, pp. 1-12.

[26] Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., and Nießner, M., Face2face: Real-time face capture and

reenactment of rgb videos, In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2387-2395.

[27] Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., and Ortega-Garcia, J., Deepfakes and beyond: A survey of face manipulation and fake detection, *Information Fusion*, 2020, Vol. 64, pp. 131-148.

[28] Verdoliva, L., Media Forensics and Deepfakes: An Overview, *IEEE Journal of Selected Topics in Signal Processing*, 2020, Vol. 14, No. 5, pp. 910-932.

[29] Wang, R., Juefei-Xu, F., Ma, L., Xie, X., Huang, Y., Wang, J., and Liu, Y., Fakespotter: A simple yet robust baseline for spotting ai-synthesized fake faces, *ArXiv Preprint ArXiv:1909.06122*, 2019.

[30] Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., and Catanzaro, B., High-resolution image synthesis and semantic manipulation with conditional gans, In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8798–8807.

[31] Yang, J.Y. and Lee, S.R., A Study on the Priority-Gap Measurement of Performance Factors Before and After Introduction of Electronic Price Information System in Retail Stores using IT-BSC and AHP, *Information Systems Review*, 2020, Vol. 22, No. 2, pp. 53-76.

[32] Youm, G.Y. and Kim, M.C., SAR image target recognition research trend using deep learning, *The Journal of The Korean Institute of Communication Sciences*, 2017, Vol. 34, No. 7, pp. 31-39.

[33] Yu, P., Xia, Z., Fei, J., and Lu, Y., A Survey on Deepfake Video Detection, *IET Biometrics*, 2016, October.

[34] Yun, C.H., Shin, H.Y., Choo, S.Y., and Kim, J.I., An evaluation study on artificial intelligence data validation methods and open-source frameworks, *Journal of Korea Multimedia Society*, 2021, Vol. 24, No. 10, pp. 1403–1413.

[35] Zi, B., Chang, M., Chen, J., Ma, X., and Jiang, Y.G., WildDeepfake: A challenging real-world dataset for Deepfake detection, In *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, October, pp. 2382-2390.

[36] UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets.php.

[37] Kaggle, https://www.kaggle.com/.

**ORCID**

Rayng-Hyung Kim | https://orcid.org//0000-0002-7676-1055
Tae-Gu Kim | https://orcid.org//0000-0003-0246-3546