

명세서

발명의 명칭: 음성을 이용한 안면 검출 방법

기술분야

- [1] 본 발명은 음성을 이용한 안면 검출 방법에 관한 것이다. 구체적으로, 본 발명은 수신된 음성데이터를 기초로 미리 정해진 메시지와 관련된 구간을 도출하고, 도출된 구간을 기준으로 추출된 영상데이터의 영상 프레임에서 안면 이미지를 검출하는 방법에 관한 것이다.

[2]

배경기술

- [3] 이 부분에 기술된 내용은 단순히 본 실시예에 대한 배경 정보를 제공할 뿐 종래기술을 구성하는 것은 아니다.
- [4] 최근 스마트 디바이스와 네트워크의 발전, 그리고 다양한 네트워크 서비스의 발달로 인하여 종래 대면으로 이루어지던 은행업무를 포함하는 여러 업무들이 온라인/무선을 이용한 비대면 업무처리 형태로 전환되었다. 이때, 비대면 업무처리 중 사용자에게 대한 본인인증이 필요한 경우, 사용자의 실시간 영상으로부터 사용자의 안면을 추출하여 미리 등록된 사용자의 사진과 비교하는 안면 검출 방법이 널리 사용되고 있다.
- [5] 종래의 안면 검출 방법은 녹화된 전체 영상에 대해 디코딩을 실행하고, 디코딩된 녹화영상의 모든 프레임에 대해서 최적의 얼굴포즈가 존재하는 특정 프레임을 탐색하는 방식을 취하고 있어, 안면 검출에 대해 상당한 시간과 리소스를 필요로 하였다.
- [6] 또한, 종래의 다른 안면 검출 방법은 녹화영상의 모든 프레임을 추출하고, 추출된 모든 프레임에 대해 안면 검출 알고리즘을 실행함으로써, 안면 검출에 이용되는 리소스가 급격하게 증가되는 문제점이 있었다.
- [7] 따라서, 적은 시간과 리소스를 이용하여 동일한 효과를 얻을 수 있는 안면 검출 방법에 대한 니즈가 존재하였다.

[8]

발명의 상세한 설명

기술적 과제

- [9] 본 발명의 목적은, 주파수 영역으로 변환한 음성데이터를 이용하여 미리 정해진 메시지와 관련된 구간을 도출하고, 도출된 구간에 대응되는 영상데이터에서 미리 정해진 기준을 만족하는 영상 프레임을 도출하고, 도출된 영상 프레임에서 안면 이미지를 검출하는 방법을 제공하는 것이다.
- [10] 또한, 본 발명의 다른 목적은, 미리 학습된 딥러닝 모듈을 이용하여 미리 정해진 메시지와 가장 관련도 높은 음성데이터의 구간을 도출하고, 도출된 구간에 대응되는 영상데이터에서 미리 정해진 기준을 만족하는 영상 프레임을

도출하고, 도출된 영상 프레임에서 안면 이미지를 검출하는 방법을 제공하는 것이다.

- [11] 본 발명의 목적들은 이상에서 언급한 목적으로 제한되지 않으며, 언급되지 않은 본 발명의 다른 목적 및 장점들은 하기의 설명에 의해서 이해될 수 있고, 본 발명의 실시예에 의해 보다 분명하게 이해될 것이다. 또한, 본 발명의 목적 및 장점들은 특허 청구 범위에 나타낸 수단 및 그 조합에 의해 실현될 수 있음을 쉽게 알 수 있을 것이다.

[12]

기술적 해결방법

- [13] 본 발명의 일 실시예에 따른 안면 검출 방법은, 사용자 단말과 연계된 서버에서 수행되는 안면 검출 방법에 있어서, 상기 사용자 단말로부터 영상데이터와 음성데이터를 수신하는 단계, 상기 수신된 음성데이터를 기초로 미리 정해진 메시지와 관련된 제1 구간을 도출하는 단계, 상기 도출된 제1 구간을 기초로 제2 구간을 설정하는 단계, 상기 제2 구간에 대응되는 상기 영상데이터의 일부를 추출하는 단계, 상기 추출된 영상데이터에서 미리 정해진 기준을 만족하는 영상 프레임을 도출하는 단계 및 상기 도출된 영상 프레임에 포함된 안면 이미지를 검출하는 단계를 포함한다.
- [14] 또한, 상기 제1 구간을 도출하는 단계는, 상기 음성데이터를 미리 정해진 시간단위마다 주파수 영역으로 변환한 스펙트로그램(spectrogram)을 생성하는 단계와, 상기 미리 정해진 메시지를 포함하는 음성데이터의 주파수 패턴을 생성하는 단계와, 상기 스펙트로그램에서 상기 주파수 패턴과 유사도가 가장 높은 구간을 상기 제1 구간으로 선정하는 단계를 포함할 수 있다.
- [15] 또한, 상기 스펙트로그램을 생성하는 단계는, 상기 미리 정해진 시간단위로 설정된 제1 윈도우에 해당하는 제1 음성데이터를 주파수 영역으로 변환한 제1 스펙트럼을 생성하고, 상기 미리 정해진 시간단위로 설정되며, 상기 제1 윈도우와 다른 제2 윈도우에 해당하는 제2 음성데이터를 주파수 영역으로 변환한 제2 스펙트럼을 생성하고, 상기 제1 스펙트럼과 상기 제2 스펙트럼을 병합하여 상기 스펙트로그램을 생성하는 것을 포함한다.
- [16] 또한, 상기 제1 윈도우와 상기 제2 윈도우는, 상기 음성데이터의 시간영역에서 일부 오버랩될 수 있다.
- [17] 또한, 상기 제1 구간을 도출하는 단계는, 상기 음성데이터를 미리 정해진 시간단위의 구간으로 샘플링하는 단계와, 미리 정해진 메시지를 포함하는 음성 패턴을 생성하는 단계와, 딥러닝 모듈을 이용하여 상기 샘플링된 구간별 음성데이터와, 상기 음성 패턴을 기초로 구간별 음성 유사도를 추출하는 단계와, 상기 음성 유사도가 미리 정해진 기준치보다 높은 구간을 상기 제1 구간으로 선정하는 단계를 포함할 수 있다.
- [18] 또한, 상기 딥러닝 모듈은, 상기 샘플링된 구간별 음성데이터 및 상기 음성

- 패턴을 입력 노드로 하는 입력 레이어와, 상기 음성 유사도를 출력 노드로 하는 출력 레이어와, 상기 입력 레이어와 상기 출력 레이어 사이에 배치되는 하나 이상의 히든 레이어를 포함하고, 상기 입력 노드와 상기 출력 노드 사이의 노드 및 에지의 가중치는 상기 딥러닝 모듈의 학습 과정에 의해 업데이트될 수 있다.
- [19] 또한, 상기 제2 구간은, 상기 음성데이터 내에서 상기 제1 구간보다 시계열적으로 후순위에 위치할 수 있다.
- [20] 또한, 상기 제2 구간의 일부는, 상기 제1 구간에 오버랩될 수 있다.
- [21] 또한, 상기 영상 프레임을 도출하는 단계는, 상기 제2 구간에 대해, 미리 정해진 주기를 이용하여 하나 이상의 프레임을 도출하거나, 상기 제2 구간에서 각 프레임의 옵티컬 플로우(Optical flow)가 기준치보다 작은 프레임을 도출하는 것을 포함할 수 있다.
- [22] 또한, 상기 안면 이미지를 검출하는 단계는, 상기 도출된 각 프레임에 대한 안면 랜드마크를 도출하고, 상기 도출된 랜드마크를 기초로 안면 정렬을 위한 보정을 수행하고, 상기 보정된 이미지에서 특징점을 추출하는 것을 포함할 수 있다.
- [23] 본 발명의 다른 실시예에 따른 안면 검출 방법은, 사용자 단말과 연계된 영상데이터와 음성데이터를 수신하는 단계, 상기 수신된 음성데이터를 기초로 미리 정해진 메시지와 관련된 구간을 도출하는 단계, 상기 도출된 구간을 기준으로, 미리 정해진 범위의 상기 영상데이터의 일부를 추출하는 단계, 상기 추출된 영상데이터에서 미리 정해진 기준을 만족하는 영상 프레임을 도출하는 단계 및 상기 도출된 영상 프레임에 포함된 안면 이미지를 검출하는 단계를 포함한다.
- [24] 또한, 상기 구간을 도출하는 단계는, 상기 음성데이터를 미리 정해진 시간단위마다 주파수 영역으로 변환한 스펙트로그램을 생성하는 단계와, 상기 미리 정해진 메시지를 포함하는 음성데이터의 주파수 패턴을 생성하는 단계와, 상기 스펙트로그램에서 상기 주파수 패턴과 유사도가 가장 높은 구간을 상기 구간으로 선정하는 단계를 포함할 수 있다.
- [25] 또한, 상기 제1 사용자는 상기 원본 메일의 담당자이고, 상기 제2 사용자는 상기 담당자의 관리자일 수 있다.
- [26] 또한, 상기 구간을 도출하는 단계는, 상기 음성데이터를 미리 정해진 시간단위의 구간으로 샘플링하는 단계와, 미리 정해진 메시지를 포함하는 음성 패턴을 생성하는 단계와, 딥러닝 모듈을 이용하여 상기 샘플링된 구간별 음성데이터와, 상기 음성 패턴을 기초로 구간별 음성 유사도를 추출하는 단계와, 상기 음성 유사도가 미리 정해진 기준치보다 높은 구간을 상기 구간으로 선정하는 단계를 포함할 수 있다.
- [27]

발명의 효과

- [28] 본 발명의 안면 검출 방법은, 주파수 영역으로 변환한 음성데이터를 이용하여 미리 정해진 메시지와 관련된 구간을 도출하고, 도출된 구간에 대응되는 영상데이터에 포함된 프레임 내에서 안면 이미지를 검출함으로써, 정면으로 정렬된 최적의 안면 이미지를 빠르게 탐색할 수 있다. 이에 따라, 본 발명은 안면 검출에 소요되는 시간을 단축시켜 사용자의 안면 검출 속도를 향상시키고, 시스템에 인가되는 부하를 감소시킬 수 있다.
- [29] 또한, 본 발명의 안면 검출 방법은, 미리 학습된 딥러닝 모듈을 이용하여 미리 정해진 메시지와 가장 관련도 높은 음성데이터의 구간을 도출하고, 도출된 구간에 대응되는 영상데이터에 포함된 프레임 내에서 정면으로 정렬된 최적의 안면 이미지를 검출함으로써, 정면으로 정렬된 최적의 안면 이미지를 빠르게 탐색할 수 있다. 이를 통해, 본 발명은 안면 검출의 정확도를 높이고, 안면 검출에 필요한 시간과 리소스를 감소시킬 수 있다.
- [30] 상술한 내용과 더불어 본 발명의 구체적인 효과는 이하 발명을 실시하기 위한 구체적인 사항을 설명하면서 함께 기술한다.

[31]

도면의 간단한 설명

- [32] 도 1은 본 발명의 실시예에 따른 안면 검출 방법을 수행하는 시스템을 설명하기 위한 개념도이다.
- [33] 도 2는 본 발명의 몇몇 실시예에 따른 안면 검출 방법을 기초로 안면 유사도를 산출하는 과정을 설명하기 위한 도면이다.
- [34] 도 3은 본 발명의 몇몇 실시예에 따른 안면 검출 방법을 설명하기 위한 순서도이다.
- [35] 도 4는 도 3의 S220 단계에 따른 제1 구간을 도출하는 방법의 일 예를 설명하기 위한 순서도이다.
- [36] 도 5는 도 4의 S321 단계에서 스펙트로그램을 생성하는 몇몇 예시를 설명하기 위한 도면이다.
- [37] 도 6은 도 4의 안면 검출 방법을 통해 생성된 스펙트로그램을 설명하기 위한 도면이다.
- [38] 도 7은 도 3의 S220 단계에 따른 제1 구간을 도출하는 방법의 다른 예를 설명하기 위한 도면이다.
- [39] 도 8은 도 7의 안면 검출 방법에서 이용되는 딥러닝 모듈을 개략적으로 설명하기 위한 블록도이다.
- [40] 도 9는 도 8의 딥러닝 모듈의 구성을 도시한 도면이다.
- [41] 도 10은 도 3의 S250 단계 및 S260 단계에 대한 몇몇 예시를 설명하기 위한 순서도이다.
- [42] 도 11은 본 발명의 몇몇 실시예에 따른 안면 검출 방법을 수행하는 시스템의 하드웨어 구현을 설명하기 위한 도면이다.

[43]

발명의 실시를 위한 형태

[44]

본 명세서 및 특허청구범위에서 사용된 용어나 단어는 일반적이거나 사전적인 의미로 한정하여 해석되어서는 아니된다. 발명자가 그 자신의 발명을 최선의 방법으로 설명하기 위해 용어나 단어의 개념을 정의할 수 있다는 원칙에 따라, 본 발명의 기술적 사상과 부합하는 의미와 개념으로 해석되어야 한다. 또한, 본 명세서에 기재된 실시예와 도면에 도시된 구성은 본 발명이 실현되는 하나의 실시예에 불과하고, 본 발명의 기술적 사상을 전부 대변하는 것이 아니므로, 본 출원시점에 있어서 이들을 대체할 수 있는 다양한 균등물과 변형 및 응용 가능한 예들이 있을 수 있음을 이해하여야 한다.

[45]

본 명세서 및 특허청구범위에서 사용된 제1, 제2, A, B 등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 상기 구성요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다. 예를 들어, 본 발명의 권리 범위를 벗어나지 않으면서 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다. '및/또는'이라는 용어는 복수의 관련된 기재된 항목들의 조합 또는 복수의 관련된 기재된 항목들 중의 어느 항목을 포함한다.

[46]

본 명세서 및 특허청구범위에서 사용된 용어는 단지 특정한 실시 예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 출원에서 "포함하다" 또는 "가지다" 등의 용어는 명세서상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[47]

다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해서 일반적으로 이해되는 것과 동일한 의미를 가지고 있다.

[48]

일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥 상 가지는 의미와 일치하는 의미를 가지는 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[49]

또한, 본 발명의 각 실시예에 포함된 각 구성, 과정, 공정 또는 방법 등은 기술적으로 상호 간 모순되지 않는 범위 내에서 공유될 수 있다.

[50]

이하에서는, 도 1 내지 도 11을 참조하여 본 발명의 실시예에 따른 안면 검출 방법 및 이를 수행하는 시스템에 대해 자세히 설명하도록 한다.

[51]

[52]

도 1은 본 발명의 실시예에 따른 안면 검출 방법을 수행하는 시스템을 설명하기

위한 개념도이다.

- [53] 도 1을 참조하면, 본 발명의 실시예에 따른 시스템은, 금융사 서버(100), 사용자 단말(200) 및 상담원 단말(300)을 포함한다.
- [54] 금융사 서버(100)(이하, 서버)는 사용자 단말(200)과 상담원 단말(300) 간의 영상통화를 중개하며, 영상통화 데이터를 이용하여 사용자의 신원확인 또는 본인인증을 수행할 수 있다. 이때, 서버(100)는 안면 검출 방법을 이용하여 영상통화에서 사용자의 안면 이미지를 추출하고, 추출된 안면 이미지를 이용하여 사용자의 신원확인 또는 본인인증을 수행할 수 있다.
- [55] 다만, 서버(100)에서 수행되는 안면 검출 방법이 위의 동작에 국한되는 것은 아니며, 다양한 실시예에서 응용되어 수행될 수 있음은 자명하나, 이하에서는 설명의 편의를 위하여 영상통화에서 사용자의 본인인증을 수행하는 것을 예로 들어 설명하도록 한다.
- [56] 서버(100)는 안면 검출 방법의 수행주체로써 동작할 수 있다. 구체적으로, 서버(100)는 사용자 단말(200)로부터 영상통화 데이터를 수신할 수 있다. 이때, 영상통화 데이터는 사용자의 목소리를 녹음한 음성데이터 및 사용자의 얼굴을 촬영한 영상데이터를 포함할 수 있다.
- [57] 이어서, 서버(100)는 수신된 음성데이터를 기초로 미리 정해진 메시지와 관련된 특정 구간(이하, 제1 구간)을 도출할 수 있다.
- [58] 이때, 서버(100)는 사용자의 음성데이터를 주파수 영역으로 변환하는 과정을 통해 생성한 스펙트로그램(spectrogram), 또는 딥러닝 모듈(Deep learning module)을 이용하여, 미리 정해진 메시지를 포함하는 음성 패턴과 유사한 음성데이터 구간을 도출할 수 있다.
- [59] 여기에서, 스펙트로그램(spectrogram)은 소리나 파동을 시각화하여 파악하기 위한 도구로, 파형(waveform)과 스펙트럼(spectrum)의 특징이 조합된 그래프를 의미한다. 파형(waveform) 그래프에서는 시간축의 변화에 따른 진폭 축의 변화가 나타나고, 스펙트럼(spectrum)에서는 주파수 축의 변화에 따른 진폭 축의 변화가 나타나는 반면, 스펙트로그램에서는 시간축과 주파수 축의 변화에 따라 진폭의 차이를 인쇄 농도 또는 표시 색상의 차이로 나타내게 된다.
- [60] 본 발명의 일 실시예에서, 서버(100)는 음성데이터의 스펙트로그램을 이용하여 제1 구간을 도출할 수 있다.
- [61] 구체적으로, 서버(100)는 음성데이터를 미리 정해진 시간단위마다 주파수 영역으로 변환한 스펙트로그램을 생성한다. 이어서, 서버(100)는 미리 정해진 메시지(예를 들어, “카메라 정면에 얼굴을 향해 주세요”)를 포함하는 음성데이터의 주파수 패턴을 생성한다.
- [62] 이어서, 서버(100)는 생성된 주파수 패턴과 가장 유사한 스펙트로그램 내의 구간을 제1 구간으로 설정할 수 있다. 이때, 제1 구간은 시간축을 기준으로 설정될 수 있다. 스펙트로그램을 이용하여 음성데이터 구간을 도출하는 과정은 도 4 내지 도 6을 통해 자세히 설명하도록 한다.

- [63]
- [64] 또한, 본 발명의 다른 실시예에서, 서버(100)는 미리 학습된 딥러닝 모듈을 이용하여 제1 구간을 도출할 수 있다.
- [65] 구체적으로, 서버(100)는 음성데이터를 미리 정해진 시간단위의 구간으로 샘플링한다. 이어서, 서버(100)는 미리 정해진 메시지(예를 들어, “카메라 정면에 얼굴을 향해 주세요”)를 포함하는 음성 패턴을 생성할 수 있다. 이어서, 서버(100)는 미리 학습된 딥러닝 모듈을 이용하여 샘플링된 음성데이터와, 생성된 음성 패턴을 비교하여 구간별 음성 유사도를 산출할 수 있다. 이때, 음성 유사도를 산출하는 알고리즘은 다양하게 변형되어 이용될 수 있으며, 해당 알고리즘에 대한 자세한 설명은 통상의 기술자에게 널리 알려져 있는 바, 이에 대한 자세한 설명은 여기에서 생략하도록 한다.
- [66] 이어서, 서버(100)는 유사도가 미리 정해진 기준치보다 높은 구간을 제1 구간으로 선정할 수 있다. 딥러닝 모듈을 이용하여 음성데이터 구간을 도출하는 과정은 도 7 내지 도 9를 이용하여 후술하도록 한다.
- [67] 이어서, 서버(100)는 도출된 제1 구간을 기준으로 제2 구간을 도출할 수 있다. 이때, 제2 구간은 제1 구간과 다른 위치에 배치될 수 있으며, 미리 정해진 메시지의 종류에 따라 상대적인 위치가 다르게 설정될 수 있다.
- [68] 예를 들어, “카메라 정면에 얼굴을 향해 주세요”라는 미리 정해진 메시지를 기준으로 제2 구간이 도출되는 경우, 제2 구간은 음성데이터 내에서 제1 구간보다 시계열적으로 뒤에(즉, 후순위에) 위치할 수 있다.
- [69] 다른 예로, “얼굴 검사를 완료하였습니다.”라는 미리 정해진 메시지를 기준으로 제2 구간이 도출되는 경우, 제2 구간은 음성데이터 내에서 제1 구간보다 시계열적으로 앞에 위치할 수 있다.
- [70]
- [71] 이어서, 서버(100)는 도출된 구간(미리 정해진 메시지와 관련된 구간; 즉, 제2 구간)을 기준으로 영상데이터의 일부를 추출하고, 추출된 영상데이터에 포함된 영상 프레임을 도출할 수 있다.
- [72] 이때, 서버(100)는 도출된 구간에 대해 다양한 방법으로 영상 프레임을 도출할 수 있다.
- [73] 예를 들어, 서버(100)는 일정 시간 간격(예를 들어, 1/n 프레임 간격)으로 영상 프레임을 도출할 수 있다. 다른 예로, 서버(100)는 도출된 구간의 옵티컬 플로우(Optical flow)가 기준치보다 작은 프레임을 도출할 수 있다. 여기에서, 옵티컬 플로우란, 카메라에 의해 촬영되어 입력되는 시간적으로 다른 2개의 영상데이터로부터 그 영상에 나타나는 외견상 움직임을 벡터로 나타낸 것을 말한다. 다만, 이는 영상 프레임을 도출하는 몇몇 예시에 불과하고, 본 발명은 다양한 방법을 통해 영상 프레임이 도출될 수 있음은 물론이다. 이어서, 서버(100)는 도출된 영상 프레임에서 안면 이미지를 검출할 수 있다. 영상 프레임 도출 및 안면 이미지를 검출하는 방법은 도 10에서 자세히 설명하도록 한다.

- [74] 이어서, 서버(100)는 도출된 안면 이미지를 이용하여, 사용자의 신원확인 또는 본인인증의 절차를 수행할 수 있다.
- [75] 본 발명에서 서버(100)와 사용자 단말(200)은 서버-클라이언트 시스템으로 구현될 수 있다. 구체적으로, 서버(100)는 각 사용자 계정에 대해 음성데이터, 영상데이터 및 미리 입력받은 안면 이미지(예를 들어, 신분증 이미지 또는 과거에 검출된 안면 이미지 등)를 분류하여 저장 및 관리할 수 있고, 금융정보 제공 및 영상통화 등과 관련된 다양한 서비스를 사용자 단말(200)에 설치된 단말 어플리케이션을 통해 제공할 수 있다.
- [76] 이때, 단말 어플리케이션은 음성데이터 및 영상데이터를 수신하기 위한 전용 어플리케이션이거나, 웹 브라우징 어플리케이션일 수 있다. 여기에서, 전용 어플리케이션은 사용자 단말(200)에 내장된 어플리케이션이거나, 어플리케이션 배포 서버로부터 다운로드 되어 사용자 단말(200)에 설치된 어플리케이션일 수 있다.
- [77] 사용자 단말(200)은 유무선 통신 환경에서 어플리케이션을 동작시킬 수 있는 통신 단말기를 의미한다. 도 1에서 사용자 단말(200)은 휴대용 단말기의 일종인 스마트폰(smart phone)으로 도시되었지만, 본 발명이 이에 한정되는 것은 아니며, 상술한 바와 같이 금융 어플리케이션을 동작시킬 수 있는 장치에 제한없이 적용될 수 있다. 예를 들어, 사용자 단말(200)은 퍼스널 컴퓨터(PC), 노트북, 태블릿, 휴대폰, 스마트폰, 웨어러블 디바이스(예를 들어, 워치형 단말기) 등의 다양한 형태의 전자 장치를 포함할 수 있다.
- [78] 또한, 도면 상에는 하나의 사용자 단말(200)만을 도시하였으나, 본 발명이 이에 한정되는 것은 아니며, 서버(100)는 복수의 사용자 단말(200)과 연동하여 동작할 수 있다.
- [79] 부가적으로, 사용자 단말(200)은 사용자의 입력을 수신하는 입력부, 비주얼 정보를 디스플레이 하는 디스플레이부, 외부와 신호를 송수신하는 통신부, 사용자의 얼굴을 촬영하는 카메라부, 사용자의 음성을 디지털 데이터로 변환하는 마이크부, 및 데이터를 프로세싱하고 사용자 단말(200) 내부의 각 유닛들을 제어하며 유닛들 간의 데이터 송/수신을 제어하는 제어부를 포함할 수 있다. 이하, 사용자의 명령에 따라 제어부가 사용자 단말(200) 내부에서 수행하는 명령은 사용자 단말(200)이 수행하는 것으로 통칭한다.
- [80] 한편, 상담원 단말(300)은 서버(100)와 상호 연계되어 동작하며, 사용자 단말(200)과 영상통화를 수행하는 상대방이 될 수 있다. 도면에 명확하게 도시하지는 않았으나, 서버(100)는 복수의 상담원 단말(300)과 연계되어 동작하며, 사용자 단말(200)로부터 영상통화요청이 수신되는 경우, 복수의 상담원 단말(300) 중 어느 하나를 선택하여 영상통화를 요청한 사용자 단말(200)과 매칭시킬 수 있다.
- [81] 서버(100)는 매칭된 사용자 단말(200)과 상담원 단말(300)에 상호 영상통화를 수행할 수 있도록 중계하는 역할을 수행한다. 이때, 서버(100)는 사용자

- 단말(200)과 상담원 단말(300) 간의 영상통화의 내역을 저장 관리할 수 있다.
- [82] 한편, 통신망(400)은 서버(100), 사용자 단말(200) 및 상담원 단말(300)을 연결하는 역할을 수행한다. 즉, 통신망(400)은 사용자 단말(200) 또는 상담원 단말(300)이 서버(100)에 접속한 후 데이터를 송수신할 수 있도록 접속 경로를 제공하는 통신망을 의미한다. 통신망(400)은 예컨대 LANs(Local Area Networks), WANs(Wide Area Networks), MANs(Metropolitan Area Networks), ISDNs(Integrated Service Digital Networks) 등의 유선 네트워크나, 무선 LANs, CDMA, 블루투스, 위성 통신 등의 무선 네트워크를 망라할 수 있으나, 본 발명의 범위가 이에 한정되는 것은 아니다.
- [83] 이하에서는, 본 발명의 실시예에 따른 시스템에서 수행되는 안면 검출 방법에 대해 구체적으로 살펴보도록 한다.
- [84]
- [85] 도 2는 본 발명의 몇몇 실시예에 따른 안면 검출 방법을 기초로 안면 유사도를 산출하는 과정을 설명하기 위한 도면이다.
- [86] 도 2를 참조하면, 서버(100)는 사용자 단말(200)로부터 수신한 영상통화 데이터(VC) 중 음성데이터(SD)를 이용하여 사용자의 음성을 분석하여, 영상데이터(VD) 중 일부에 해당하는 특정 구간을 추출한다(S110).
- [87] 구체적으로, 서버(100)는 영상통화가 진행되는 사용자 단말(200)로부터 영상데이터(VD) 및 음성데이터(SD)를 포함하는 영상통화 데이터(VC)를 실시간으로 수신할 수 있다. 서버(100)는 수신된 음성데이터(SD)를 분석하여 미리 정해진 메시지(예를 들어, “카메라 정면에 얼굴을 향해 주세요” 또는 “얼굴 촬영이 완료되었습니다.”)와 관련된 구간을 도출할 수 있다.
- [88] 이때, 서버(100)는 스펙트로그램 또는 딥러닝 모듈을 이용하여 미리 정해진 메시지와 관련된 구간을 도출할 수 있다. 이에 대한 자세한 설명은 도 4 내지 도 6 및 도 7 내지 도 9에서 자세히 설명하도록 한다.
- [89] 이어서, 서버(100)는 추출된 음성데이터(SD)의 특정 구간에 해당하는 영상데이터(VD)에서, 샘플링을 통해 특정 프레임을 추출한다(S120).
- [90] 여기에서, 서버(100)는 도출된 특정 구간을 기준으로, 미리 정해진 범위의 영상데이터(VD)의 일부 구간을 추출할 수 있다. 서버(100)는 추출된 영상데이터(VD)에서 미리 정해진 기준을 만족하는 몇몇 영상 프레임을 도출할 수 있다.
- [91] 예를 들어, 서버(100)는 추출된 영상데이터(VD)에 대해 일정 시간 간격으로 프레임을 샘플링하거나, 옵티컬 플로우가 기준치보다 작은 영상 프레임을 도출하여 샘플링 할 수 있다.
- [92] 다른 예로, 서버(100)는 추출된 영상데이터(VD)에 대해 포즈 검출 알고리즘을 동작시킬 수 있다. 포즈 검출 알고리즘에 의해 미리 정해진 포즈가 검출된 경우, 서버(100)는 포즈 검출 알고리즘을 종료하고 검출된 포즈와 관련된 영상 프레임을 추출할 수 있다.

- [93] 다만, 이는 영상 프레임을 도출하는 몇몇 예시에 불과하며, 본 발명이 이에 제한되는 것은 아니다.
- [94] 이어서, 서버(100)는 추출된 영상 프레임에서 사용자의 안면을 검출한다(S130). 서버(100)는 미리 학습된 딥러닝 모델(예를 들어, MTCNN, Retinaface, 또는 Blazeface)을 이용하여 사용자의 안면을 검출할 수 있다. 사용자의 안면은 영상 프레임 내에서 바운딩 박스를 이용하여 검출될 수 있다. 이때, 서버(100)에서 사용되는 딥러닝 모델은 다양하게 변형되어 사용될 수 있다.
- [95] 이어서, 서버(100)는 추출된 사용자의 안면을 정렬한다(S140).
- [96] 구체적으로, 서버(100)는 추출된 안면에 대한 안면 랜드마크를 검출할 수 있다. 이때, 안면 랜드마크란 눈, 코, 입, 턱선 및 콧대와 같은 안면의 특징을 구성하는 부분을 뜻한다. 이어서, 서버(100)는 검출된 안면 랜드마크를 기초로 안면을 정렬할 수 있다. 예를 들어, 서버(100)는 눈과 눈 사이에 직선을 형성하고, 해당 직선과 가로 수평선 사이의 각도를 측정하여 반대각도만큼 안면 이미지를 회전시키는 방법을 이용할 수 있다. 다만, 이는 하나의 예시에 불과하며 본 발명이 이에 한정되는 것은 아니다.
- [97] 이어서, 서버(100)는 정렬된 안면의 특징점을 추출한다(S150).
- [98] 이어서, 서버(100)는 추출된 안면의 특징점을 이용하여 안면의 유사도를 산출한다(S160). 이때, 서버(100)는 추출된 안면 특징점을 실수 벡터로 표현할 수 있으며, 미리 저장된 사용자의 신분증 이미지에서 추출된 특징점과 비교하는 과정을 통하여 안면 유사도를 산출할 수 있다. 이렇게 산출된 안면 유사도는, 사용자 얼굴의 동일성 판단에 이용될 수 있다.
- [99] 이하에서는, 본 발명의 몇몇 실시예에 따른 안면 검출 방법에서 제1 구간 및 제2 구간을 도출하는 과정에 대해 자세히 설명하도록 한다.
- [100]
- [101] 도 3은 본 발명의 몇몇 실시예에 따른 안면 검출 방법을 설명하기 위한 순서도이다.
- [102] 도 3을 참조하면, 서버(100)는 영상통화를 통해 영상데이터 및 음성데이터를 수신한다(S120).
- [103] 이어서, 서버(100)는 수신된 음성데이터를 기초로, 미리 정해진 메시지와 관련된 제1 구간을 도출한다(S220).
- [104] 예를 들어, 서버(100)는 수신한 음성데이터에서 “카메라 정면에 얼굴을 향해주세요”라는 미리 정해진 메시지가 출력되는 구간을 제1 구간으로 설정할 수 있다. 이때, 서버(100)는 음성데이터를 주파수 영역으로 변환한 스펙트로그램 또는 미리 학습된 딥러닝 모듈을 이용하여 미리 정해진 메시지와 관련된 제1 구간을 도출할 수 있다.
- [105] 이어서, 서버(100)는 도출된 제1 구간을 기초로 제2 구간을 설정한다(S230).
- [106] 예를 들어, 서버(100)는 도출된 제1 구간의 종료지점부터 약 10초동안의 구간 또는 제1 구간의 종료지점부터 “얼굴 촬영이 완료되었습니다.”라는 메시지가

포함된 부분까지의 구간을 제2 구간으로 설정할 수 있다. 다만, 이는 하나의 예시일 뿐, 본 발명이 이에 제한되는 것은 아니다.

- [107] 여기서, 제2 구간은 음성데이터 내에서 제1 구간보다 시계열적으로 후순위에 위치할 수 있고, 제2 구간의 일부는 제1 구간에 오버랩 될 수 있음은 물론이다.
- [108] 이어서, 서버(100)는 제2 구간에 대응되는 영상데이터의 일부를 추출한다(S240).
- [109] 이어서, 서버(100)는 추출된 영상데이터에서 미리 정해진 기준을 만족하는 영상 프레임을 도출한다(S250). 이때, 서버(100)는 영상데이터에 대해 미리 설정된 일정 시간을 주기(예를 들어, 1/n)마다 영상 프레임을 도출하거나, 오픈컬 플로우를 이용하여 영상 프레임을 도출할 수 있다.
- [110] 이어서, 서버(100)는 도출된 영상 프레임에 포함된 안면 이미지를 검출한다(S260). 이때, 서버(100)는 미리 학습된 딥러닝 모델(예를 들어, MTCNN, Retinaface, 또는 Blazeface)을 이용하여 사용자의 안면을 검출할 수 있고, 사용자의 안면은 영상 프레임 내에서 바운딩 박스를 이용하여 검출될 수 있다. 다만, 본 발명이 이에 한정되는 것은 아니며, 서버(100)에서 사용되는 딥러닝 모델은 다양하게 변형되어 사용될 수 있음은 물론이다.
- [111] 이하에서는 본 발명의 일 실시예에 따른 스펙트로그램을 이용하여 제1 구간을 도출하는 안면 검출 방법에 대해 설명하도록 한다.
- [112]
- [113] 도 4는 도 3의 S220 단계에 따른 제1 구간을 도출하는 방법의 일 예를 설명하기 위한 순서도이다.
- [114] 도 4를 참조하면, S210 단계에 이어서, 서버(100)는 음성데이터를 특정 시간단위 마다 주파수 영역으로 변환하여 스펙트로그램을 생성한다(S321).
- [115] 구체적으로, 서버(100)는 사용자 단말(200)로부터 수신한 음성데이터에 대해 미리 정해진 시간단위를 기초로 분할할 수 있다. 이어서, 서버(100)는 분할된 복수의 음성데이터를 각각 주파수 영역으로 변환하여 복수 개의 스펙트럼을 생성하고, 생성된 복수 개의 스펙트럼을 시간 순으로 병합하여 스펙트로그램을 생성할 수 있다.
- [116] 이어서, 서버(100)는 미리 정해진 메시지(예를 들어, “카메라 정면에 얼굴을 향해 주세요”)를 포함하는 음성데이터의 주파수 패턴을 생성한다(S323). 이때, 서버(100)는 미리 정해진 메시지가 포함된 음성데이터의 샘플을 변환하여, 미리 정해진 메시지에 대응되는 주파수 패턴을 생성할 수 있다.
- [117] 이어서, 서버(100)는 S321 단계에서 생성된 스펙트로그램과, S323 단계에서 생성된 주파수 패턴을 비교하여, 상기 주파수 패턴과 가장 유사한 시간영역 상의 제1 구간을 도출한다(S325).
- [118] 이때, 서버(100)는 스펙트로그램에서 미리 정해진 시간단위 별로 주파수 패턴과의 유사도를 도출할 수 있다. 이어서, 서버(100)는 스펙트로그램에서 주파수 패턴과 유사도가 가장 높은 구간을 제1 구간으로 선택할 수 있다.

- [119]
- [120] 도 5는 도 4의 S321 단계에서 스펙트로그램을 생성하는 몇몇 예시를 설명하기 위한 도면이다.
- [121] 도 5를 참조하면, (a11)은 미리 정해진 시간단위의 윈도우로 분할된 음성데이터를 나타내고, (a12)는 (a11)에서 분할된 음성데이터를 주파수 영역으로 변환한 스펙트럼을 시계열적으로 이어 붙여 만들어진 스펙트로그램을 나타낸다.
- [122] 이때, 서버(100)는 STFT(Short Time Fourier Transform, 국소 푸리에 변환)를 이용하여 음성데이터를 주파수 영역으로 변환할 수 있다. 여기에서, STFT란, 데이터에서 시간에 대해 구간을 짧게 나눈 후, 나누어진 여러 구간의 데이터에 대해 푸리에 변환을 실시하여 단위시간에 따른 주파수 분포를 이미지화 하는 방법이다.
- [123] 구체적으로, 서버(100)는 사용자 단말(200)로부터 수신한 음성데이터를 미리 정해진 시간단위로 나눌 수 있다. 이하에서는, 설명의 편의를 위해 미리 정해진 시간단위를 3.3초라고 가정하고 설명하도록 한다.
- [124] 예를 들어, (a11)를 참조하면, 서버(100)는 10초 길이의 음성데이터를 3.3초 단위로 나눌 수 있다. 이때, 서버(100)는 음성데이터의 0초 내지 3.3초에 해당하는 구간을 제1 윈도우(W11)로 설정할 수 있고, 3.4초 내지 6.6초에 해당하는 구간을 제2 윈도우(W12)로 설정할 수 있다. 또한, 서버(100)는 6.8초 내지 10초에 해당하는 구간을 제3 윈도우(W13)로 설정할 수 있다. 여기에서, 윈도우의 가로길이(Window length)는 미리 정해진 시간단위이다. 즉, 제1 윈도우 내지 제3 윈도우(W11 내지 W13)의 가로길이는 3.3초일 수 있다.
- [125] 이어서, 서버(100)는 제1 윈도우 내지 제3 윈도우(W11 내지 W13)를 주파수 영역으로 변환하여 각각의 스펙트럼을 생성할 수 있다. 구체적으로, 서버(100)는 제1 윈도우(W11)에 해당하는 제1 음성데이터를 주파수 영역으로 변환하여 제1 스펙트럼(S11)을 생성할 수 있다. 이어서, 서버(100)는 제2 윈도우(W12)의 제2 음성데이터를 변환하여 제2 스펙트럼(S12)을 생성하고, 제3 윈도우(W13)의 제3 음성데이터를 변환하여 제3 스펙트럼(S13)을 생성할 수 있다.
- [126] 이어서, 서버(100)는 생성된 제1 스펙트럼 내지 제3 스펙트럼(S11 내지 S13)을 시계열순으로 병합하여 음성데이터에 대한 스펙트로그램(a12)을 생성할 수 있다.
- [127]
- [128] 한편, 서버(100)는 음성데이터에 대해 오버랩(Overlap)된 윈도우를 적용한 STFT 분석을 수행할 수 있다. 이때, 복수의 윈도우는 음성데이터의 시간영역에서 오버랩 될 수 있으며, 오버랩되는 길이는 미리 설정되거나, 윈도우의 비율로 특정될 수 있다.
- [129] 예를 들어, (a21)을 참조하면, 서버(100)는 10초 길이의 음성데이터를 3.3초 단위로 나눌 수 있다. 서버(100)는 음성데이터의 0초 내지 3.3초에 해당하는

구간을 제1 윈도우(W21)로 설정할 수 있다.

- [130] 이어서, 서버(100)는 제1 윈도우(W21)에 오버랩 되는 제2 윈도우(W22)를 설정할 수 있다. 이때, 제2 윈도우(W22)는 2.2초 내지 5.5초에 해당하는 구간에 위치할 수 있다.
- [131] 또한, 서버(100)는 제2 윈도우(W22)에 오버랩 되는 제3 윈도우(W23)와, 제3 윈도우(W23)에 오버랩 되는 제4 윈도우(W24)를 설정할 수 있다.
- [132] 이어서, 서버(100)는 제1 윈도우 내지 제4 윈도우(W21 내지 W24)를 주파수 영역으로 변환하여, 각각의 스펙트럼(S21 내지 S24)을 생성할 수 있다.
- [133] 이어서, 서버(100)는 생성된 복수의 스펙트럼(S21 내지 S24)을 시계열순으로 병합하여 음성데이터에 대한 스펙트로그램(a22)을 생성할 수 있다.
- [134] 이때, 각각의 스펙트럼은 일측에 배치된 윈도우와 오버랩되는 시간구간을 뺀 나머지 구간에 배치될 수 있다. 예를 들어, 제1 윈도우(W21)의 단위시간은 0초 내지 3.3초이나, 일측에 위치하는 제2 윈도우(W22)와 오버랩되는 구간을 뺀, 0초 내지 2.2초에 해당하는 위치에 변환된 제1 스펙트럼(S21)이 배치될 수 있다.
- [135] 또한, 생성된 스펙트로그램(a22)을 살펴보면, 각 스펙트럼은 양쪽에 위치한 각 스펙트럼들의 주파수 영역과 일부 겹치는 것을 확인할 수 있다.
- [136] 이렇게 시간영역에서 오버랩되는 윈도우를 이용함으로써, 본 발명은 제1 구간을 더 세밀하게 도출할 수 있어, 미리 정해진 메시지와 매칭되는 구간을 도출하는데 있어 정확도를 향상시킬 수 있다.
- [137]
- [138] 도 6은 도 4의 안면 검출 방법을 통해 생성된 스펙트로그램을 설명하기 위한 도면이다.
- [139] 도 6을 참고하면, 서버(100)는 사용자 단말(200)로부터 수신한 음성데이터에 대해 전술한 도 5의 과정을 통하여 스펙트로그램을 생성할 수 있다.
- [140] 서버(100)는 생성된 스펙트로그램에서, 미리 정해진 메시지를 포함하는 음성데이터에 관한 주파수 패턴과, 유사도가 가장 높은 구간을 도출할 수 있다. 예를 들어, 서버(100)는 스펙트로그램을 미리 정해진 구간별로 구분하고, 구분된 각 구간에 대한 스펙트럼과 주파수 패턴 간의 유사도를 산출할 수 있다.
- [141] 이어서, 서버(100)는 산출된 유사도가 제일 높은 스펙트럼이 속한 구간을 제1 구간으로 선정할 수 있다.
- [142] 추가적으로, 서버(100)는 제1 구간을 도출하는데 있어, 로그 멜 스펙트로그램(Log mel spectrogram) 또는 림로사(LibROSA)를 이용할 수 있다. 다만, 이는 하나의 예시에 불과하며, 제1 구간을 도출하기 위한 다양한 알고리즘이 이용될 수 있음은 물론이다.
- [143] 이하에서는 본 발명의 다른 실시예에 따른 딥러닝 모듈을 이용하여 제1 구간을 도출하는 안면 검출 방법에 대해 설명하도록 한다.
- [144]
- [145] 도 7은 도 3의 S220 단계에 따른 제1 구간을 도출하는 방법의 다른 예를

설명하기 위한 도면이다.

- [146] 도 7을 참고하면, 서버(100)는 음성데이터를 특정 시간단위의 구간으로 샘플링한다(S421). 구체적으로, 서버(100)는 샘플링 모듈에 사용자 단말(200)로부터 수신한 음성데이터를 입력할 수 있다. 샘플링 모듈은 입력된 음성데이터를 기초로 미리 설정된 특정 시간단위로 음성데이터를 구간별로 나누어 출력할 수 있다.
- [147] 이어서, 서버(100)는 미리 정해진 메시지를 포함하는 음성 패턴을 생성한다(S423). 서버(100)는 미리 정해진 메시지(예를 들어, “카메라 정면에 얼굴을 향해 주세요”)가 포함된 음성데이터의 일부를 음성 패턴으로 설정할 수 있다.
- [148] 이어서, 서버(100)는 딥러닝 모듈을 이용하여, 샘플링된 구간 별 음성데이터와 음성 패턴을 기초로 구간 별 음성 유사도를 추출한다(S425). 이때, 딥러닝 모듈의 입력 노드에는 샘플링된 구간 별 음성데이터 및 음성 패턴이 입력되고, 출력 노드에는 음성 유사도가 출력될 수 있다.
- [149] 이어서, 서버(100)는 딥러닝 모듈에서 출력된 음성 유사도가 미리 정해진 기준치보다 높은 구간을 도출하여 제1 구간으로 설정한다(S427). 이때, 서버(100)는 음성 유사도가 미리 정해진 기준치보다 높은 구간 중 음성 유사도가 가장 높은 구간을 제1 구간으로 도출할 수 있다.
- [150]
- [151] 도 8은 도 7의 안면 검출 방법에서 이용되는 딥러닝 모듈을 개략적으로 설명하기 위한 블록도이다.
- [152] 구체적으로, 도 8을 참조하면, 딥러닝 모듈(DM)은 구간 별 음성데이터 및 음성 패턴을 입력받고, 이에 대한 출력으로 구간 별 음성 유사도를 출력할 수 있다.
- [153] 이때, 구간 별 음성데이터는 샘플링 모듈(SM)에 의해 생성될 수 있다. 샘플링 모듈(SM)은 사용자 단말(200)로부터 입력받은 음성데이터를 미리 설정된 구간별로 나누어지도록 샘플링할 수 있다. 샘플링 모듈(SM)을 통해 출력된 구간 별 음성데이터는 딥러닝 모듈(DM)에 입력될 수 있다. 또한, 음성 패턴은 미리 정해진 메시지(예를 들어, “카메라 정면에 얼굴을 향해 주세요”)가 포함된 음성데이터를 의미한다.
- [154] 딥러닝 모듈(DM)은 빅데이터를 기초로 학습된 인공지능망을 이용하여, 음성 패턴에 대한 구간 별 음성데이터의 유사도(즉, 구간 별 음성 유사도)를 도출할 수 있다.
- [155] 딥러닝 모듈(DM)은 입력된 데이터를 기초로 도출된 별도의 파라미터에 대한 매핑 데이터를 이용하여 인공지능망 학습을 수행할 수 있다. 딥러닝 모듈(DM)은 학습 인자로 입력되는 파라미터들에 대하여 머신 러닝(machine learning)을 수행할 수 있다. 이때, 서버(100)의 메모리에는 머신 러닝에 사용되는 데이터 및 결과 데이터 등이 저장될 수 있다.
- [156] 보다 자세히 설명하자면, 머신 러닝(Machine Learning)의 일종인 딥러닝(Deep

Learning) 기술은 데이터를 기반으로 다단계로 깊은 수준까지 내려가 학습하는 것이다.

- [157] 딥러닝(Deep learning)은, 단계를 높여가면서 복수의 데이터들로부터 핵심적인 데이터를 추출하는 머신 러닝(Machine Learning) 알고리즘의 집합을 나타낸다.
- [158] 딥러닝 모듈(DM)은 공지된 다양한 딥러닝 구조를 이용할 수 있다. 예를 들어, 딥러닝 모듈(DM)은 CNN(Convolutional Neural Network), RNN(Recurrent Neural Network), DBN(Deep Belief Network), GNN(Graph Neural Network) 등의 구조를 이용할 수 있다.
- [159] 구체적으로, CNN(Convolutional Neural Network)은 사람이 물체를 인식할 때 물체의 기본적인 특징들을 추출한 다음 뇌 속에서 복잡한 계산을 거쳐 그 결과를 기반으로 물체를 인식한다는 가정을 기반으로 만들어진 사람의 뇌 기능을 모사한 모델이다.
- [160] RNN(Recurrent Neural Network)은 자연어 처리 등에 많이 이용되며, 시간의 흐름에 따라 변하는 시계열 데이터(Time-series data) 처리에 효과적인 구조로 매 순간마다 레이어를 쌓아올려 인공신경망 구조를 구성할 수 있다.
- [161] DBN(Deep Belief Network)은 딥러닝 기법인 RBM(Restricted Boltzman Machine)을 다층으로 쌓아 구성되는 딥러닝 구조이다. RBM(Restricted Boltzman Machine) 학습을 반복하여 일정 수의 레이어가 되면, 해당 개수의 레이어를 가지는 DBN(Deep Belief Network)이 구성될 수 있다.
- [162] GNN(Graphic Neural Network, 그래픽 인공신경망, 이하, GNN)는 특정 파라미터 간 매핑된 데이터를 기초로 모델링된 모델링 데이터를 이용하여, 모델링 데이터 간의 유사도와 특징점을 도출하는 방식으로 구현된 인공신경망 구조를 나타낸다.
- [163] 한편, 딥러닝 모듈(DM)의 인공신경망 학습은 주어진 입력에 대하여 원하는 출력이 나오도록 노드간 연결선의 웨이트(weight)를 조정(필요한 경우 바이어스(bias) 값도 조정)함으로써 이루어질 수 있다. 또한, 인공신경망은 학습에 의해 웨이트(weight) 값을 지속적으로 업데이트시킬 수 있다. 또한, 인공신경망의 학습에는 역전파(Back Propagation) 등의 방법이 사용될 수 있다.
- [164] 한편, 서버(100)의 메모리에는 머신 러닝으로 미리 학습된 인공신경망(Artificial Neural Network)이 탑재될 수 있다.
- [165] 딥러닝 모듈(DM)은 도출된 파라미터에 대한 모델링 데이터를 입력 데이터로 하는 머신 러닝(machine learning) 기반의 개선 프로세스 추천 동작을 수행할 수 있다. 이때, 인공신경망의 머신 러닝 방법으로는 준지도학습(semi-supervised learning)과 지도학습(supervised learning)이 모두 사용될 수 있다. 또한, 딥러닝 모듈(DM)은 설정에 따라 학습 후 구간 별 음성 유사도를 출력하기 위한 인공신경망 구조를 자동 업데이트하도록 제어될 수 있다.
- [166] 추가적으로, 도면에 명확하게 도시하지는 않았으나, 본 발명의 다른 실시예에서, 딥러닝 모듈(DM)의 동작은 서버(100) 또는 별도의 클라우드

서버(미도시)에서 실시될 수 있다. 이하에서는, 전술한 본 발명의 실시예에 따른 딥러닝 모듈(DM)의 구성에 대해 살펴보도록 한다.

[167]

[168] 도 9는 도 8의 딥러닝 모듈의 구성을 도시한 도면이다.

[169] 도 9를 참조하면, 딥러닝 모듈(DM)은 구간 별 음성데이터 및 음성 패턴을 입력노드로 하는 입력 레이어(input)와, 구간 별 음성 유사도를 출력노드로 하는 출력 레이어(Output)와, 입력 레이어와 출력 레이어 사이에 배치되는 M 개의 히든 레이어를 포함한다.

[170] 여기서, 각 레이어들의 노드를 연결하는 에지(edge)에는 가중치가 설정될 수 있다. 이러한 가중치 혹은 에지의 유무는 학습 과정에서 추가, 제거, 또는 업데이트 될 수 있다. 따라서, 학습 과정을 통하여, k개의 입력노드와 i개의 출력노드 사이에 배치되는 노드들 및 에지들의 가중치는 업데이트될 수 있다.

[171] 딥러닝 모듈(DM)이 학습을 수행하기 전에는 모든 노드와 에지는 초기값으로 설정될 수 있다. 그러나, 누적하여 정보가 입력될 경우, 노드 및 에지들의 가중치는 변경되고, 이 과정에서 학습인자로 입력되는 파라미터들(즉, 구간 별 음성데이터 및 음성 패턴)과 출력노드로 할당되는 값(즉, 구간 별 음성 유사도) 사이의 매칭이 이루어질 수 있다.

[172] 추가적으로, 클라우드 서버(미도시)를 이용하는 경우, 딥러닝 모듈(DM)은 많은 수의 파라미터들을 수신하여 처리할 수 있다. 따라서, 딥러닝 모듈(DM)은 방대한 데이터에 기반하여 학습을 수행할 수 있다.

[173] 딥러닝 모듈(DM)을 구성하는 입력노드와 출력노드 사이의 노드 및 에지의 가중치는 딥러닝 모듈(DM)의 학습 과정에 의해 업데이트될 수 있다. 또한, 딥러닝 모듈(DM)에서 출력되는 파라미터는 구간 별 음성 유사도 외에도 다양한 데이터로 추가 확장될 수 있음은 물론이다.

[174] 이어서, 서버(100)는 제1 구간을 기준으로 제2 구간을 설정하고, 제2 구간에 대응되는 영상데이터의 일부를 추출할 수 있다. 이에 대한 자세한 설명은 전술하였으므로, 중복되는 설명은 생략하도록 한다.

[175] 이하에서는, 추출된 영상데이터에서 미리 설정된 기준을 만족하는 영상 프레임을 추출하고, 도출된 영상 프레임에 포함된 안면 이미지를 검출하는 방법에 대한 몇몇 예시에 대해 설명하도록 한다.

[176]

[177] 도 10은 도 3의 S250 단계 및 S260 단계에 대한 몇몇 예시를 설명하기 위한 순서도이다.

[178] 도 10을 참조하면, 본 발명의 일 실시예에서, 서버(100)는 제2 구간에 대한 영상데이터에 대하여, 일정 시간 간격(예를 들어, 1/n 프레임 간격)으로 영상 프레임을 도출할 수 있다(S551).

[179] 이때, 서버(100)는 영상 프레임의 도출을 위한 프레임 도출주기를 미리 설정할 수 있다. 예를 들어, 도출주기가 10으로 설정된 경우, 서버(100)는 제2 구간의

영상데이터에 포함된 10개의 영상 프레임 마다 1개의 영상 프레임을 도출할 수 있다. 다만, 이는 하나의 예시에 불과하며, 영상 프레임의 도출주기는 가변되거나, 랜덤하게 형성될 수 있음은 물론이다.

- [180] 한편, 본 발명의 다른 실시예에서, 서버(100)는 제2 구간에 대한 영상데이터에 대하여, 영상데이터의 옵티컬 플로우가 기준치 보다 작은 영상 프레임을 도출한다(S553).
- [181] 예를 들어, 서버(100)는 제2 구간의 영상데이터에서 제1 프레임과 제2 프레임을 추출하고, 각 영상 프레임 내에서 하나 이상의 특징점을 기준으로 벡터 형식의 옵티컬 플로우를 추출할 수 있다. 이때, 서버(100)는 벡터의 절대값을 계산하여 옵티컬 플로우의 크기를 산출할 수 있다. 이어서, 서버(100)는 산출된 옵티컬 플로우의 크기가 미리 설정된 기준치보다 작은 경우, 해당 옵티컬 플로우를 포함하는 영상 프레임을 도출할 수 있다.
- [182] 다만, 이는 영상 프레임을 도출하는 몇몇 예시에 불과하고, 본 발명이 위 방법에 제한되는 것은 아니다.
- [183] 이어서, 서버(100)는 추출된 영상 프레임에서 사용자의 안면 이미지를 검출한다. 서버(100)는 미리 학습된 딥러닝 모델(예를 들어, MTCNN, Retinaface, 또는 Blazeface)을 이용하여 사용자의 안면 이미지를 검출할 수 있다. 사용자의 안면 이미지는 영상 프레임 내에서 바운딩 박스를 이용하여 검출될 수 있다. 이때, 서버(100)에서 사용되는 딥러닝 모델은 다양하게 변형되어 사용될 수 있다.
- [184] 이어서, 서버(100)는 도출된 각 영상 프레임에 대한 안면 랜드마크를 도출한다(S561). 예를 들어, 서버(100)는 영상 프레임에 표시된 안면에서 눈, 코, 입, 턱선 또는 콧대를 도출할 수 있다.
- [185] 이어서, 서버(100)는 도출된 랜드마크를 기초로 안면 정렬을 위한 보정을 수행한다(S563). 예를 들어, 서버(100)는 도출된 랜드마크 중 좌측 눈의 시작부분과 우측 눈의 시작부분을 선으로 연결하여 직선을 생성할 수 있다. 이어서, 서버(100)는 생성된 직선과 수평기준선 사이의 각도를 측정할 수 있다. 서버(100)는 측정된 각도와 동일한 크기의 반대각도로 도출된 안면 이미지를 회전시킴으로써, 안면 이미지를 정렬할 수 있다. 다만, 이는 하나의 예시에 불과하고, 본 발명이 위의 방법에 한정되는 것은 아니다.
- [186] 이어서, 서버(100)는 안면 정렬을 위한 보정이 수행된 이미지에서 특징점을 추출한다(S565). 이때, 특징점은 이미 공개된 다양한 알고리즘에 의해 추출될 수 있으므로, 여기에서 자세한 설명은 생략하도록 한다.
- [187] 이어서, 서버(100)는 사용자의 신분증 이미지에서 추출된 특징점과 보정된 이미지에서 추출된 특징점을 비교함으로써 안면 유사도를 산출할 수 있다. 이렇게 산출된 안면 유사도는, 사용자 얼굴의 동일성 판단에 이용될 수 있다.
- [188]
- [189] 도 11은 본 발명의 몇몇 실시예에 따른 안면 검출 방법을 수행하는 시스템의 하드웨어 구현을 설명하기 위한 도면이다.

- [190] 도 11을 참조하면, 본 발명의 몇몇 실시예들에 따른 안면 검출 방법을 수행하는 서버(100)는 전자 장치(1000)로 구현될 수 있다. 전자 장치(1000)는 컨트롤러(1010), 입출력 장치(1220, I/O), 메모리 장치(1230, memory device), 인터페이스(1040) 및 버스(1250, bus)를 포함할 수 있다. 컨트롤러(1010), 입출력 장치(1020), 메모리 장치(1030) 및/또는 인터페이스(1040)는 버스(1050)를 통하여 서로 결합될 수 있다. 버스(1050)는 데이터들이 이동되는 통로(path)에 해당한다.
- [191] 구체적으로, 컨트롤러(1010)는 CPU(Central Processing Unit), MPU(Micro Processor Unit), MCU(Micro Controller Unit), GPU(Graphic Processing Unit), 마이크로프로세서, 디지털 신호 프로세서, 마이크로컨트롤러, 어플리케이션 프로세서(AP, application processor) 및 이들과 유사한 기능을 수행할 수 있는 논리 소자들 중에서 적어도 하나를 포함할 수 있다.
- [192] 입출력 장치(1020)는 키패드(keypad), 키보드, 터치스크린 및 디스플레이 장치 중 적어도 하나를 포함할 수 있다. 메모리 장치(1030)는 데이터 및/또는 프로그램 등을 저장할 수 있다.
- [193] 인터페이스(1040)는 통신 네트워크로 데이터를 전송하거나 통신 네트워크로부터 데이터를 수신하는 기능을 수행할 수 있다. 인터페이스(1040)는 유선 또는 무선 형태일 수 있다. 예컨대, 인터페이스(1040)는 안테나 또는 유무선 트랜시버 등을 포함할 수 있다. 도시하지 않았지만, 메모리 장치(1030)는 컨트롤러(1010)의 동작을 향상시키기 위한 동작 메모리로서, 고속의 디램 및/또는 에스램 등을 더 포함할 수도 있다. 메모리 장치(1030)는 내부에 프로그램 또는 어플리케이션을 저장할 수 있다.
- [194] 사용자 단말(200)은 개인 휴대용 정보 단말기(PDA, personal digital assistant) 포터블 컴퓨터(portable computer), 웹 타블렛(web tablet), 무선 전화기(wireless phone), 모바일 폰(mobile phone), 디지털 뮤직 플레이어(digital music player), 메모리 카드(memory card), 또는 정보를 무선환경에서 송신 및/또는 수신할 수 있는 모든 전자 제품에 적용될 수 있다.
- [195] 또는, 본 발명의 실시예들에 따른 서버(100) 및 사용자 단말(200)은 각각 복수의 전자 장치(1000)가 네트워크를 통해서 서로 연결되어 형성된 시스템일 수 있다. 이러한 경우에는 각각의 모듈 또는 모듈의 조합들이 전자 장치(1000)로 구현될 수 있다. 단, 본 실시예가 이에 제한되는 것은 아니다.
- [196] 추가적으로, 서버(100)는 워크스테이션(workstation), 데이터 센터, 인터넷 데이터 센터(internet data center(IDC)), DAS(direct attached storage) 시스템, SAN(storage area network) 시스템, NAS(network attached storage) 시스템 및 RAID(redundant array of inexpensive disks, or redundant array of independent disks) 시스템 중 적어도 하나로 구현될 수 있으나, 본 실시예가 이에 제한되는 것은 아니다.
- [197] 또한, 서버(100)는 사용자 단말(200)을 이용하여 네트워크를 통해서 데이터를 전송할 수 있다. 네트워크는 유선 인터넷 기술, 무선 인터넷 기술 및 근거리 통신

기술에 의한 네트워크를 포함할 수 있다. 유선 인터넷 기술은 예를 들어, 근거리 통신망(LAN, Local area network) 및 광역 통신망(WAN, wide area network) 중 적어도 하나를 포함할 수 있다.

- [198] 무선 인터넷 기술은 예를 들어, 무선랜(Wireless LAN: WLAN), DMNA(Digital Living Network Alliance), 와이브로(Wireless Broadband: Wibro), 와이맥스(World Interoperability for Microwave Access: Wimax), HSDPA(High Speed Downlink Packet Access), HSUPA(High Speed Uplink Packet Access), IEEE 802.16, 롱 텀 에볼루션(Long Term Evolution: LTE), LTE-A(Long Term Evolution-Advanced), 광대역 무선 이동 통신 서비스(Wireless Mobile Broadband Service: WMBS) 및 5G NR(New Radio) 기술 중 적어도 하나를 포함할 수 있다. 단, 본 실시예가 이에 제한되는 것은 아니다.
- [199] 근거리 통신 기술은 예를 들어, 블루투스(Bluetooth), RFID(Radio Frequency Identification), 적외선 통신(Infrared Data Association: IrDA), UWB(Ultra-Wideband), 지그비(ZigBee), 인접 자장 통신(Near Field Communication: NFC), 초음파 통신(Ultra Sound Communication: USC), 가시광 통신(Visible Light Communication: VLC), 와이 파이(Wi-Fi), 와이 파이 다이렉트(Wi-Fi Direct), 5G NR (New Radio) 중 적어도 하나를 포함할 수 있다. 단, 본 실시예가 이에 제한되는 것은 아니다.
- [200] 네트워크를 통해서 통신하는 서버(100)는 이동통신을 위한 기술표준 및 표준 통신 방식을 준수할 수 있다. 예를 들어, 표준 통신 방식은 GSM(Global System for Mobile communication), CDMA(Code Division Multi Access), CDMA2000(Code Division Multi Access 2000), EV-DO(Enhanced Voice-Data Optimized or Enhanced Voice-Data Only), WCDMA(Wideband CDMA), HSDPA(High Speed Downlink Packet Access), HSUPA(High Speed Uplink Packet Access), LTE(Long Term Evolution), LTEA(Long Term Evolution-Advanced) 및 5G NR(New Radio) 중 적어도 하나를 포함할 수 있다. 단, 본 실시예가 이에 제한되는 것은 아니다.
- [201]
- [202] 정리하면, 본 발명의 안면 검출 방법은 주파수 영역으로 변환한 음성데이터를 이용하여 미리 정해진 메시지와 관련된 구간을 도출하거나, 미리 학습된 딥러닝 모듈을 이용하여 미리 정해진 메시지와 가장 관련도 높은 음성데이터의 구간을 도출할 수 있다. 이어서, 본 발명은 도출된 구간에 대응되는 영상데이터에 포함된 프레임 내에서 안면 이미지를 검출함으로써, 정면으로 정렬된 최적의 안면 이미지를 빠르게 탐색할 수 있다.
- [203] 이에 따라, 본 발명은 안면 검출에 소요되는 시간을 단축시켜 사용자의 안면 검출 속도를 향상시키고, 안면 검출의 정확도를 높일 수 있으며, 시스템에 인가되는 부하를 감소시킬 수 있다.
- [204]
- [205] 이상의 설명은 본 실시예의 기술 사상을 예시적으로 설명한 것에 불과한

것으로서, 본 실시예가 속하는 기술 분야에서 통상의 지식을 가진 자라면 본 실시예의 본질적인 특성에서 벗어나지 않는 범위에서 다양한 수정 및 변형이 가능할 것이다. 따라서, 본 실시예들은 본 실시예의 기술 사상을 한정하기 위한 것이 아니라 설명하기 위한 것이고, 이러한 실시예에 의하여 본 실시예의 기술 사상의 범위가 한정되는 것은 아니다. 본 실시예의 보호 범위는 아래의 청구범위에 의하여 해석되어야 하며, 그와 동등한 범위 내에 있는 모든 기술 사상은 본 실시예의 권리범위에 포함되는 것으로 해석되어야 할 것이다.

청구범위

- [청구항 1] 사용자 단말과 연계된 서버에서 수행되는 안면 검출 방법에 있어서,
 상기 사용자 단말로부터 영상데이터와 음성데이터를 수신하는 단계;
 상기 수신된 음성데이터를 기초로 미리 정해진 메시지와 관련된 제1 구간을 도출하는 단계;
 상기 도출된 제1 구간을 기초로 제2 구간을 설정하는 단계;
 상기 제2 구간에 대응되는 상기 영상데이터의 일부를 추출하는 단계;
 상기 추출된 영상데이터에서 미리 정해진 기준을 만족하는 영상 프레임을 도출하는 단계; 및
 상기 도출된 영상 프레임에 포함된 안면 이미지를 검출하는 단계를 포함하는 안면 검출 방법.
- [청구항 2] 제1 항에 있어서,
 상기 제1 구간을 도출하는 단계는,
 상기 음성데이터를 미리 정해진 시간단위마다 주파수 영역으로 변환한 스펙트로그램(spectrogram)을 생성하는 단계와,
 상기 미리 정해진 메시지를 포함하는 음성데이터의 주파수 패턴을 생성하는 단계와,
 상기 스펙트로그램에서 상기 주파수 패턴과 유사도가 가장 높은 구간을 상기 제1 구간으로 선정하는 단계를 포함하는 안면 검출 방법.
- [청구항 3] 제2 항에 있어서,
 상기 스펙트로그램을 생성하는 단계는,
 상기 미리 정해진 시간단위로 설정된 제1 윈도우에 해당하는 제1 음성데이터를 주파수 영역으로 변환한 제1 스펙트럼을 생성하고,
 상기 미리 정해진 시간단위로 설정되며, 상기 제1 윈도우와 다른 제2 윈도우에 해당하는 제2 음성데이터를 주파수 영역으로 변환한 제2 스펙트럼을 생성하고,
 상기 제1 스펙트럼과 상기 제2 스펙트럼을 병합하여 상기 스펙트로그램을 생성하는 것을 포함하는 안면 검출 방법.
- [청구항 4] 제3 항에 있어서,
 상기 제1 윈도우와 상기 제2 윈도우는, 상기 음성데이터의 시간영역에서 일부 오버랩되는 안면 검출 방법.
- [청구항 5] 제1 항에 있어서,
 상기 제1 구간을 도출하는 단계는,

상기 음성데이터를 미리 정해진 시간단위의 구간으로 샘플링하는 단계와,
 미리 정해진 메시지를 포함하는 음성 패턴을 생성하는 단계와,
 딥러닝 모듈을 이용하여 상기 샘플링된 구간별 음성데이터와, 상기 음성 패턴을 기초로 구간별 음성 유사도를 추출하는 단계와,
 상기 음성 유사도가 미리 정해진 기준치보다 높은 구간을 상기 제1 구간으로 선정하는 단계를 포함하는
 안면 검출 방법.

[청구항 6] 제5 항에 있어서,
 상기 딥러닝 모듈은,
 상기 샘플링된 구간별 음성데이터 및 상기 음성 패턴을 입력 노드로 하는 입력 레이어와,
 상기 음성 유사도를 출력 노드로 하는 출력 레이어와,
 상기 입력 레이어와 상기 출력 레이어 사이에 배치되는 하나 이상의 히든 레이어를 포함하고,
 상기 입력 노드와 상기 출력 노드 사이의 노드 및 에지의 가중치는 상기 딥러닝 모듈의 학습 과정에 의해 업데이트되는
 안면 검출 방법.

[청구항 7] 제1 항에 있어서,
 상기 제2 구간은, 상기 음성데이터 내에서 상기 제1 구간보다 시계열적으로 후순위에 위치하는
 안면 검출 방법.

[청구항 8] 제7 항에 있어서,
 상기 제2 구간의 일부는, 상기 제1 구간에 오버랩되는
 안면 검출 방법.

[청구항 9] 제1 항에 있어서,
 상기 영상 프레임을 도출하는 단계는,
 상기 제2 구간에 대해, 미리 정해진 주기를 이용하여 하나 이상의 프레임을 도출하거나,
 상기 제2 구간에서 각 프레임의 옵티컬 플로우(Optical flow)가 기준치보다 작은 프레임을 도출하는 것을 포함하는
 안면 검출 방법.

[청구항 10] 제1 항에 있어서,
 상기 안면 이미지를 검출하는 단계는,
 상기 도출된 각 프레임에 대한 안면 랜드마크를 도출하고,
 상기 도출된 랜드마크를 기초로 안면 정렬을 위한 보정을 수행하고,
 상기 보정된 이미지에서 특징점을 추출하는 것을 포함하는
 안면 검출 방법.

- [청구항 11] 사용자 단말과 연계된 서버에서 수행되는 안면 검출 방법에 있어서,
 상기 사용자 단말로부터 영상데이터와 음성데이터를 수신하는 단계;
 상기 수신된 음성데이터를 기초로 미리 정해진 메시지와 관련된 구간을
 도출하는 단계;
 상기 도출된 구간을 기준으로, 미리 정해진 범위의 상기 영상데이터의
 일부를 추출하는 단계;
 상기 추출된 영상데이터에서 미리 정해진 기준을 만족하는 영상
 프레임을 도출하는 단계; 및
 상기 도출된 영상 프레임에 포함된 안면 이미지를 검출하는 단계를
 포함하는
 안면 검출 방법.
- [청구항 12] 제11 항에 있어서,
 상기 구간을 도출하는 단계는,
 상기 음성데이터를 미리 정해진 시간단위마다 주파수 영역으로 변환한
 스펙트로그램(spectrogram)을 생성하는 단계와,
 상기 미리 정해진 메시지를 포함하는 음성데이터의 주파수 패턴을
 생성하는 단계와,
 상기 스펙트로그램에서 상기 주파수 패턴과 유사도가 가장 높은 구간을
 상기 구간으로 선정하는 단계를 포함하는
 안면 검출 방법.
- [청구항 13] 제11 항에 있어서,
 상기 구간을 도출하는 단계는,
 상기 음성데이터를 미리 정해진 시간단위의 구간으로 샘플링하는
 단계와,
 미리 정해진 메시지를 포함하는 음성 패턴을 생성하는 단계와,
 딥러닝 모듈을 이용하여 상기 샘플링된 구간별 음성데이터와, 상기 음성
 패턴을 기초로 구간별 음성 유사도를 추출하는 단계와,
 상기 음성 유사도가 미리 정해진 기준치보다 높은 구간을 상기 구간으로
 선정하는 단계를 포함하는
 안면 검출 방법.