마이크로 서비스 구조 기반 실시간 지능형 비디오 컨텐츠 제공 서비스 개발

유미선, 문재원

전자부품연구원

altjs543@keti.re.kr, jwmoon@keti.re.kr

Development of intelligent video web service based on Micro-service architecture

Miseon Yu, Jaewon Moon

Information and Media Research Center,

Korea Electronics Technology Institute

요약

IoT 산업과 인공지능 기술의 발전으로 다양한 데이터를 분석하여 서비스에 쉽게 활용할 수 있게 되었다. 이에 대해 클라우드 기반으로 된 분석 기술이 주로 발전하였으나, 개인 정보 노출 위험성 및 네트워크 종속성 문제를 해결하기 위해 최근에는 엣지 기반으로 분석하고 클라우드와 협업하는 기술 연구가 활발하게 진행되고 있다. 리소스가 제한적인 엣지 디바이스 기반 환경에서 원활한 서비스를 제공하기 위해서는 서비스의 기능을 목적별로 최소화하여 독립적이고 경량화된 어플리케이션을 엣지에 배포하고 실행되게 해야 한다. 마이크로서비스 설계 기법은 이를 해결 할 수 있는 대표적인 방법으로 대두되고 있다. 본 논문에서는 여러 마이크로 서비스의 결과를 전달 받아 최종적으로 적합한 결과를 재생하는 컨텐츠 제공 서비스 구조를 제안하고 구현 결과를 소개하였다. 높은 데이터 처리 성능을 요구하는 영상 처리 서비스를 제공함에 있어 제안하는 방법을 활용하여 엣지 디바이스 활용 효율성을 높이고 보다 만족도 높은 컨텐츠 제공 서비스를 제공할 수 있다.

1. 서론

IoT 산업과 인공지능 기술의 발전으로 다양한 데이터를 분석하여 서비스에 쉽게 활용할 수 있게 되었다. 이에 대해 클라우드 기반으로 된 분석 기술이 주로 발전하였으나, 개인 정보 노출 위험성 및 네트워크 종속성 문제를 해결하기 위해 최근에는 엣지 기반으로 분석하고 클라우드와 협업하는 기술 연구가 활발하게 진행되고 있다. 하지만, 리소스가 제한적인 엣지 디바이스 기반 환경의 원활한 서비스를 위해선 서비스의 기능을 목적 별로 최소화하여 독립적이고 경량화된 어플리케이션을 엣지에 배포하고 실행되게 해야 한다. 마이크로서비스 설계 기법은 이를 해결 할 수 있는 대표적인 방법으로 대두되고 있다. 마이크로 서비스 구조(Micro-service architecture)는 경량으로 최소화된 기능을 각각 개발하고 이를 독립적으로 배포하여 서로 다양한 경량 프로토콜을 통해 상호작용하게 하는 s/w 설계 구조로 가볍고 개발 및 업데이트를 기능 별로 독립적으로 할 수 있기에 유지 보수가 쉽고 배포가 빠르다는 장점이 있다.[1]

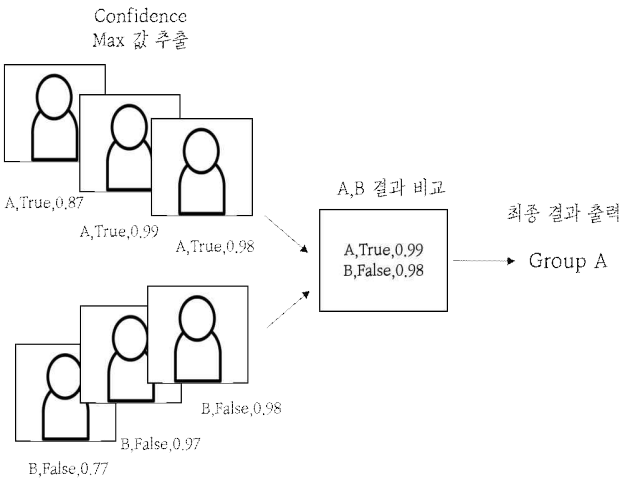
이런 상황에서, 제공되는 영상을 실시간으로 분석하고 결과를 즉각적으로 출력해주는 엣지 기반 서비스 구조 또한 마이크로 서비스 기반으로 설계되어야 한다. 특히, 동영상 스트림은 아주 빠른 시간 내에 다양한 정보를 보내기 때문에, 출력 단은 어떤 정보를 추출하고 결과를 어떻게 출력할 지에 대한 마이크로 서비스 구조 기반의 정의가 필요하다. 따라서 본 연구는 기존의 모노리틱(Monolithic) 방식이 아닌 여러 마이크로 서비스들이 협업하여 제공된 실시간 영상 분석 결과를 효율적으로 처리하고 컨텐츠를 제공하는 서비스 구조를 제안하고 그 구현 결과를 소개한다.

2. 마이크로서비스 기반 컨텐츠 제공 시스템

본 논문에서 구현하고자 하는 서비스는 동영상 서비스들을 한데 모아 사용자에게 출력을 제공하는 실시간 반응 시스템(Realtime Monitoring System)이다. 전 단은 스트림에서 사람의 얼굴을 추출하고 이에 대한 정보를 분석하는 모듈들로 구성되어 있으며, 이는 각각 비디오 스트림에서 얼굴 영역 검출 모듈(Face Detector), 얼굴 영역 특징 추출 모듈(Feature Extractor), 특징을 분석하여 멤버 여부를 추론하는 멤버 판단 모듈(Member Verifier)로 총 3개의 마이크로 서비스 모듈로 이루어져있다. 실시간 반응 시스템은 앞서 언급한 서비스들과 상호작용하여 영상에 등장한 사람에 맞는 화면을 보여준다. 그렇기에 전 단에서 보내주는 결과들과 상호작용 하며 시간 지연 없이 반응하는 것이 매우 중요하다. 기존의 모노리틱 시스템에서는 각 서비스들이 비독립적으로 연결되어 이를 순차적으로 처리하지만 마이크로 서비스 구조에서는 데이터의 상호작용이 추가적으로 필요하기 때문이다. 본 구현에서, 실시간 반응 시스템의 서버가 모듈들로부터의 응답을 기다렸다가 응답이 도착하면 바로 이에 대한 반응을 실시하도록 소켓, Rest API등의 다양한 프로토콜의 조합을 사용하였다.

실시간 반응 시스템은 동영상을 기반으로 서비스를 제공한다. 동영상 스트림은 연속된 이미지의 집합이며, 영상 데이터에 대한 감지 결과 또한 스트리밍으로 제공된다. 따라서 데이터를 받는 출력 단에서는, 이

결과 스트리밍을 통해 최종 결과를 판단한다. 모노리틱 환경과는 달리 마이크로 서비스의 마지막 출력 서비스는 결과를 각 단계와 독립적으로 제공하기 때문에 이에 대한 추가적인 판단 알고리즘이 필요하다. 이런 결정 방법으로 이동 평균(rolling average), 최댓값 추출 등이 사용될 수 있다. 본 연구는, 서비스 단에서 일정 시간 동안의 데이터들을 수집해 그 중 최댓값으로 최종 판단을 진행하도록 구현했다. 따라서 본 알고리즘에서는 약 1초 동안 받아진 데이터들의 집합에서 가장 정확도가 큰 데이터를 추출하여 이에 대한 결과들을 비교하여 최종 결과를 출력 한다. 결과들을 통해 그룹A인가 그룹B인가를 비교하는 과정은 [그림 1]과 같이 A, B중 한 가지가 True이면 그 결과를 출력하고, 둘 다 True이면 정확도가 더 높은 그룹의 결과를, 둘 다 False이거나 정확도가 낮은 경우는 둘 다 아닌 NA로 인식하여 본래의 메인 결과를 출력한다.



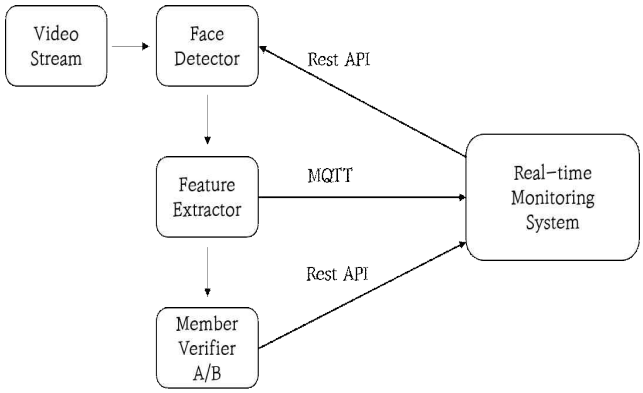
[그림 1] 동영상 스트림 결과 판단 과정

3. 구현 및 결과

서비스들을 한데 모아 사용자에게 출력을 제공하는 실시간 반응 시스템은(Realtime Monitoring System) 이들의 정보를 받아내어 최대한 시간 지연 없이 반응해야 한다. 본 연구는 이를 마이크로 서비스 구조로 설계 하였고, 각 서비스들은 하나의 컨테이너로 존재한다. 전체 설계는 [그림 2]와 같다. Face Detector에는 실시간 반응 시스템에서 단위당 추출되는 프레임 개수(FPS)를 조정하여 요청한다. 이 구간은 Rest API 구조로 HTTP 요청을 보내게 된다. 또한, Feature Extractor에서 받아진 사람 얼굴의 이미지는 MQTT protocol을 통해 경량의 메시지로 실시간 반응 시스템으로 보내져 사용자는 당시 추출된 사진의 모습을 실시간으로 보게 된다. 마지막으로, Member Verifier에서 결정된 결과가 Rest API 형식으로 말단에 보내지게 된다. 이런 컨테이너들의 통신으로 실시간 서비스를 생성하게 된다.

실시간 반응 시스템에서 가장 고려되어야 했던 부분은 일정 시간동안의 프레임 정보들을 어떻게 저장시켜 놓을지 이다. 동일 시스템 내에서 결과를 계산하는 것이 아니라 여러 마이크로 서비스에서 계산된 정보가 시스템에 전달되기 때문에 이를 최종 결과로 출력하기 까지 시간 지연 전달을 최소화 하여야 실시간 반응 서비스가 가능하다. 우리는 이 지연 시간을 줄이기 위해 시스템 자체 변수에 결과를 저장하는 방식을 적용하였다. 전단 마이크로 서비스의 결과를 데이터베이스에서 관리시 데이터 로드하는 도중 예상치 못한 시간 지연이 추가로 발생할 수 있기 때문이다. 해당 서비스 시나리오를 고려했을 때 짧은 시간 구간 동안 여러 마이크로서비스로부터 전달된 데이터들은 시스템의 내부에 충분히 저장 및 처리가 가능했다.

본 방식을 통해, 시스템 반응 시간이 평균 1ms 내였으며,, 데이터베이스에서 데이터를 읽어들이는 방식 보다 실시간 동기화가 가능했음을 확인하였다. 제안하는 웹 서비스의 구조는 사용자의 얼굴을 카메라에 비추면 그에 맞는 결과를 판단하여 적응적 컨텐츠를 제공함으로서 지능형 안내자 역할을 충실히 담당하는 것을 검증할 수 있었다.



[그림 2] micro-service architecture 구조

4. 결론

본 논문에서는 기존 모노리틱 방식이 아닌 마이크로 서비스 구조로 실시간 반응 서비스를 구현하는데 있어 새롭게 고려해야할 사항들을 살펴보았다. 마이크로서비스 기반 서비스를 설계하기 위해서는 각 서비스 간 정보 전달 과정으로 인해 생기는 시간 지연과 각 데이터에 대한 저장소의 역할을 반드시 고려해야 한다. 또한 보안에 강인하고, 반응이 빠른 프로토콜을 선정하고 그에 대한 반응 행동을 최적화하여 설계할 필요가 있다. 본 논문에서는 동영상 프레임들의 대한 결과를 일정시간 자체 변수에 저장시키는 방식으로 구현하여 동기화 가능한 서비스를 제공할 수 있음을 확인하였다. 높은 데이터 처리 성능을 요구하는 영상 처리 서비스를 제공함에 있어 엣지 디바이스(Edge device) 활용 효율성을 높인 컨텐츠 제공 서비스 구현이 가능하였다. 향후, 전단 서비스와 후단 서비스 간 총 시간 지연을 측정함으로써 판단의 정확도 및 시간 동기화 정도를 수치화 할 것이다.

[그림 3] 영상에 사람이 없을 때 출력 모습(왼쪽),
group A의 사람이 있을 때 출력 모습(오른쪽)

감사의 글

이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 1711075689, AI 어플리케이션을 지원하는 IoT 연동 분산 Edge 클라우드 기술 개발)

참고문헌

[1] Soumya Kanti Datta and Christian Bonnet, “Next-Generation, Data Centric and End-to-End IoT Architecture Based on Microservices”, 2018 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia).