**졸업논문 제안서**

**URL과 이미지 비교를 이용한 웹 사이트**

**위•변조 탐지 및 알림 서비스**

**Ver. 1.2**

**2019. 04. 02**

**한국외국어대학교 정보통신공학과**

**201403586 허성윤**

**201602119 유진솔**

지 도 교 수 : 정성호

**문서 정보**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 구 분 | 소 속 | 성 명 | 날 짜 | 서 명 |
| 작성자 | 한국외국어대학교 | 유진솔 | 2019. 04. 02 |  |
| 한국외국어대학교 | 허성윤 | 2019. 04. 02 |  |
| 검토자 | 한국외국어대학교 | 유진솔 | 2019. 04. 02 |  |
| 한국외국어대학교 | 허성윤 | 2019. 04. 02 |  |
| 승인자 | 한국외국어대학교 | 정성호 | 2019. 04. 02 |  |

**머리말**

본 문서는 URL과 이미지 비교를 이용한 웹사이트 위•변조 탐지 및 알림 서비스 소개와 연구 목적, 배경을 설명하고 관련 연구 분석과 함께 본 시스템을 구축하기 위한 시나리오를 기술한다.

**개정 이력**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **버전** | **작성자** | **개정일자** | **개정 내역** | **승인자** | |
| 1.0 | 허성윤 | 2019. 03. 31 | 초안 작성 |  | |
| 유진솔 |
| 검토자 | 허성윤, 유진솔 | | | |
| 1.1 | 허성윤 | 2019. 04. 01 | 초안 수정 |  | |
| 유진솔 |
| 검토자 | 허성윤, 유진솔 | | | |
| 1.2 | 허성윤 | 2019. 04. 01 | 최종 수정 | |  |
| 유진솔 |
| 검토자 | 허성윤, 유진솔 | | | |

목 차

목차

[목 차 5](#_Toc4933517)

[1. 개요 7](#_Toc4933518)

[1.1 서비스 정의 7](#_Toc4933519)

[1.2 서비스 기획 배경 및 목적 7](#_Toc4933520)

[2. 관련 연구 조사 및 분석 8](#_Toc4933521)

[2.1 Feature Matching 기반 탐지 기법 8](#_Toc4933522)

[2.2 URL 기반 탐지 기법 8](#_Toc4933525)

[2.3 시각적 추출 기법(Visual Descriptor)과 SVM을 이용한 탐지 기법 9](#_Toc4933531)

[3. 주제 제안 9](#_Toc4933532)

[3.1 시나리오 9](#_Toc4933533)

[3.1.1 URL 일치 여부 확인 10](#_Toc4933535)

[3.1.2 Inception-v3를 이용한 이미지 유사도 검출 10](#_Toc4933536)

[3.1.3 dHash를 이용한 이미지 유사도 검출 10](#_Toc4933537)

[3.2 사용 기술 설명 10](#_Toc4933538)

[3.2.1 TensorFlow 10](#_Toc4933539)

[3.2.2 Transfer Learning 10](#_Toc4933540)

[3.2.3 dHash Algorithm 10](#_Toc4933541)

[4. 개발 일정 11](#_Toc4933543)

[5. 참고 문헌 12](#_Toc4933544)

1. 개요

1.1 서비스 정의

본 서비스는 URL과 이미지 비교를 이용하여 공공기관 및 금융기관 웹 사이트의 위•변조 여부를 탐지하여 사용자에게 알림까지 제공하도록 응용 프로그램을 제공한다.

1.2 서비스 기획 배경 및 목적

* 본 문서는 증강현실 기반 악기를 이용한 음악 창작 공유 플랫폼 개발 프로젝트 중 다양한 요구에 대해 유연하게 대응하기 위해 시스템 요구사항을 명세하고 있다.
* 본 문서는 사용자, 기획팀, 프로젝트 관리자를 대상으로 한다.
* 본 문서를 바탕으로 고객의 요구사항을 명확하게 도출하여 향후 개발 과정에서 이를 반영하는데 그 목적이 있다. 따라서 본 문서는 고객의 정확한 요구사항을 수집 하고 분석하여 명세 한다.
* 본 문서는 고객과 개발자간에 다음과 같은 역할을 한다.

1. 본 문서는 고객과 개발자간의 계약서와 동일한 효력을 갖는다.
2. 본 문서는 고객 요구사항을 구체적으로 명시한다.
3. 개발자는 본 문서에 명세된 고객의 요구사항에 따라 목적물을 개발한다.
4. 개발된 목적물은 본 문서에 명세된 모든 요구사항을 만족해야 한다.
5. 본 문서에 명세 된 모든 요구사항은 향후 개발 과정에서 발생하는 모든 변경 사항의 베이스 라인으로 설정한다.
6. 본 문서는 명세된 요구 사항을 바탕으로 설계하고, 테스트 케이스를 작성한다.
7. 관련 연구 조사 및 분석

2.1 Feature Matching 기반 탐지 기법



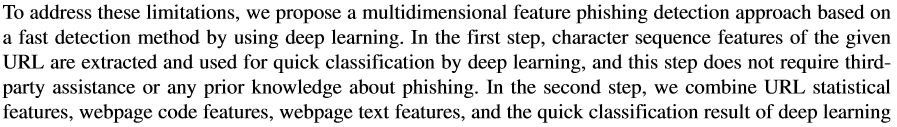
[6]

Feature Matching이란 수많은 Feature를 추출하여 다른 이미지의 Feature Set과 비교 후 유사성을 검사하는 OpenCV의 일종이다. 이미지에서 검출한 특성 포인트에 대해 색상을 지정해서 표시하고, 일치하는 특성 포인트를 하나의 선으로 연결한다. 특성 디스크립터의 개수를 줄이거나 늘임으로써 선의 개수를 조절할 수 있다. 하지만 선의 개수를 늘린다고 해도 이미지 전체에 대한 특성을 추출해내는 데에 한계가 있고, 그 특징을 추출하는 데에 시간이 상당히 소요되기 때문에 유사도 정확성과 효율이 떨어진다.

2.2 URL 기반 탐지 기법



[7]



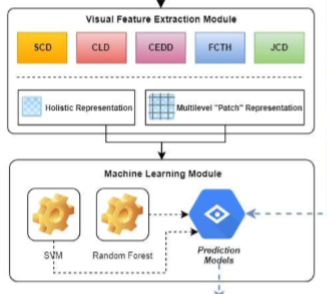
위 논문은 URL을 기반으로 위•변조 웹 사이트를 탐지한 연구이다. 이 연구에서는 URL 시퀀스

기능을 추출하여 Deep Learning을 통해 분류하는 로직으로, URL 통계 기능, 웹 페이지 코드 기능, 웹

페이지 텍스트 기능 및 Deep Learning 분류 결과를 이용한다. 그러나, 이는 단지 URL만 탐지에 사용하기 때문에 정상 도메인 주소를 가진 피싱 사이트 탐지에 부적합하다는 결점이 있다.

2.3 시각적 추출 기법(Visual Descriptor)과 SVM을 이용한 탐지 기법

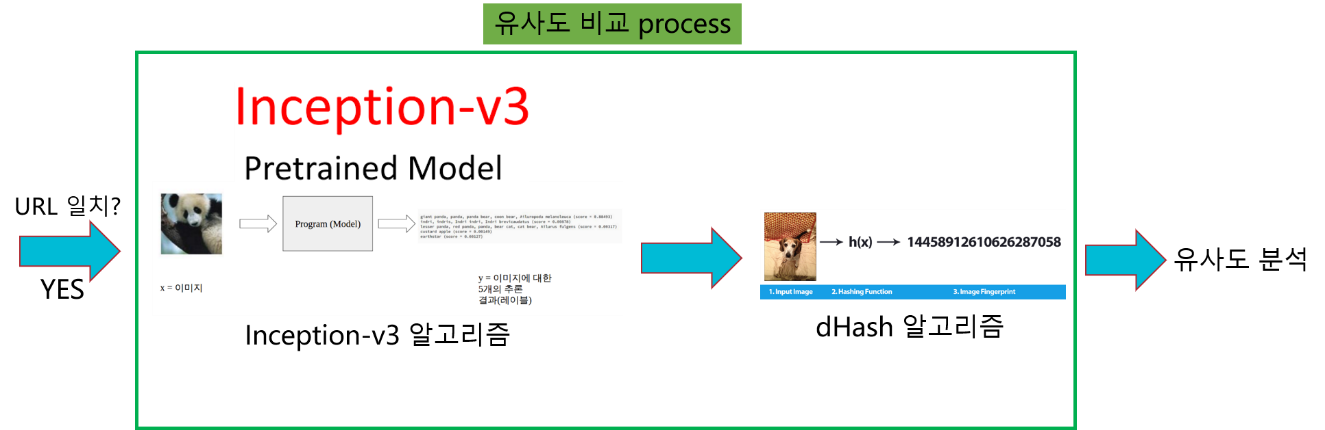


**[8]**

위 연구에서는 이미지의 특징을 추출하는 데에 Visual Descriptor를 사용해서 이미지의 색상이나 질감 등의 시각적 특징만을 추출해 낸다. 이후에 이미지를 분류하는 과정에 SVM을 사용한다. 하지만, Visual Descriptor는 이미 오래 전 기술이고 단순히 이미지의 시각적 특징만을 추출해 내기 때문에 다중 필터를 사용하는 Inception-v3의 성능이 더 우수하다고 판단하였다. 또한, 일반적으로 SVM은 본 연구에서 이용하는 Inception-v3에 비해 정확성이 떨어진다.

1. 주제 제안

3.1 시나리오



접속한 웹 페이지의 URL과 오리지널 URL의 일치 여부를 확인하고 일치 한다면, Inception-v3로 학습시켜 둔 모델에 비교 대상 이미지를 입력으로 넣어준다. 그렇게 해서 나온 결과 유사도 값과 dHash를 이용한 이미지 유사도를 분석하여 적당한 가중치를 준 뒤 이러한 과정을 적용하여 더 정확한 이미지 유사도를 검출하고자 한다.

3.1.1 URL 일치 여부 확인

3.1.2 Inception-v3를 이용한 이미지 유사도 검출

3.1.3 dHash를 이용한 이미지 유사도 검출

* 1. 오리지널 사이트의 이미지를 데이터베이스에 미리 저장해 둔다.

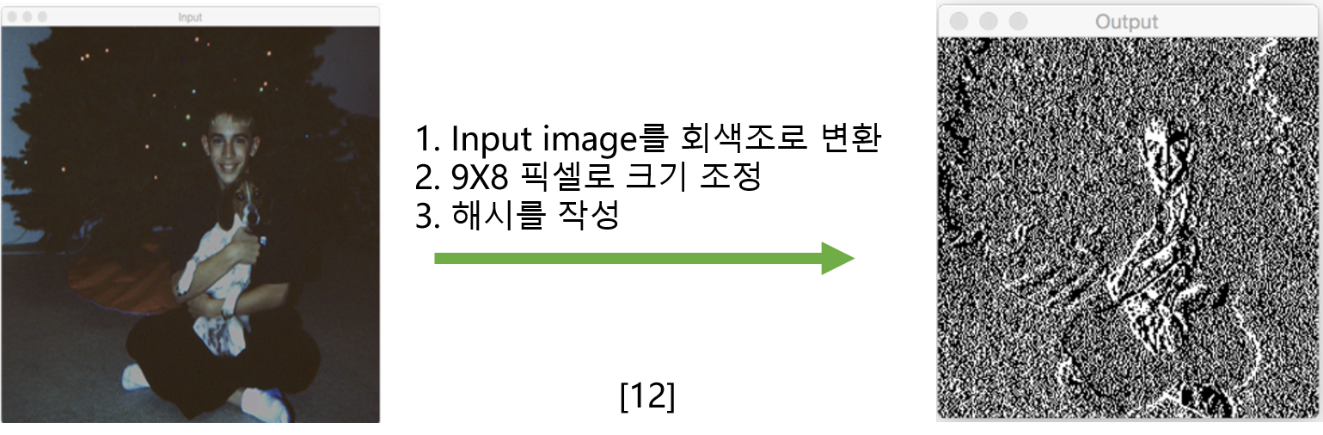
1. 웹 사이트에 접속하여 처음 보이는 홈 화면을 web application test framework의 일종인 selenium을 이용하여 캡쳐 후 로컬 파일에 자동 저장한다. (추가적인 비교를 위해 URL도 함께 저장한다.)
2. 웹 페이지의 유동적인 부분들로 인해 몇 분 간격으로 바뀌는 것까지 최대한 ‘같은 페이지의 많은 이미지’를 저장하기 위해 일정 시간 간격으로 캡쳐 하도록 한다.
   1. 오리지널 사이트의 이미지와 사용자가 접속한 사이트의 이미지를 비교하여 유사도를 검출한다.
3. Python에서 제공하는 libarary인 ImageHash에서 처리 속도가 빠르고 이미지 비교에 적합한 dHash 알고리즘을 사용한다.
4. 결과값으로 두 이미지의 차이 정도를 나타내는 해밍 거리 값을 얻는다.
   1. 검출된 유사도 정보로 사용자가 접속하고있는 사이트에 대한 유사도 정보를 제공한다.
5. 결과 해밍 거리 값을 가지고 오리지날과 유사한 정도를 백분율로 환산하여 100%라면 같은 이미지, 95%~99%라면 유사한 이미지, 그 이하라면 다른 이미지라고 판단한다.
6. 이 정보를 가지고 이 웹사이트의 위,변조 가능성을 경고하는 알림을 띄워준다.

3.2 사용 기술 설명

3.2.1 TensorFlow

3.2.2 Transfer Learning

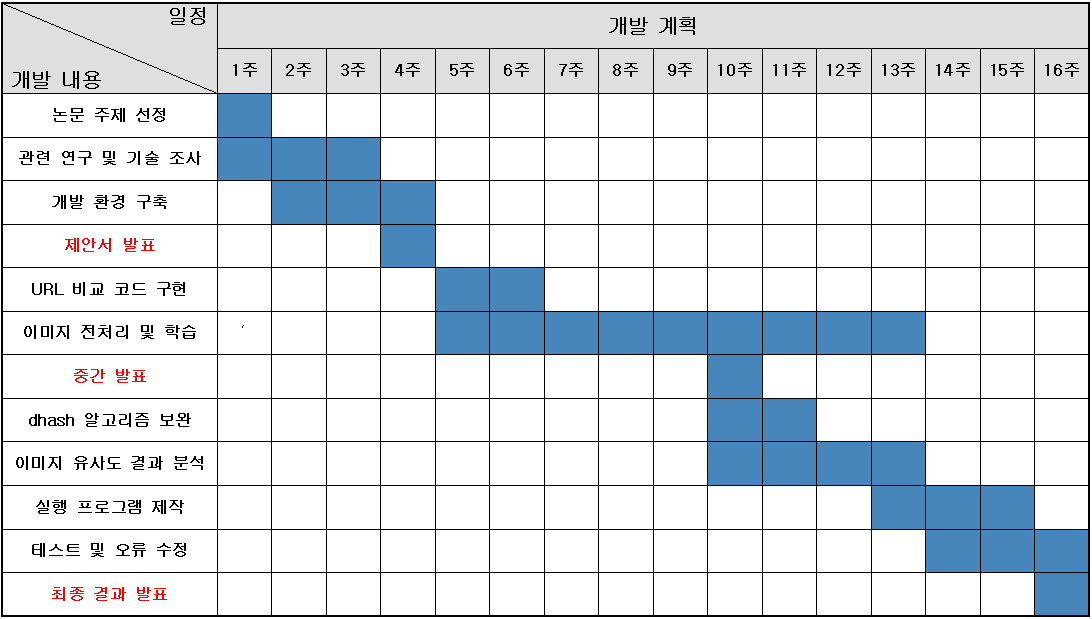
3.2.3 dHash Algorithm



dHash 알고리즘은 이미지의 인접 픽셀 간의 차이를 계산하여 만들어진 64비트 해시를 이용하여 해밍 거리를 구하고, 해밍 거리 값으로 이미지 유사도를 결정하는 방법이다.

* 1. 입력이미지를 회색조로 변환한다.
  2. 이미지 크기 조정을 9 x 8 픽셀 나노로 크기 조정을 하여, 결과 이미지 해시가 초기 공간 크기에 관계없이 비슷한 사진과 크기가 일치 하도록 한다.
  3. 인접한 픽셀의 차이를 계산하면 8가지 차이가 생겨서 8x8의 64비트 해시가 되고, 이 비트를 할당하여 해시를 작성한다.
  4. 서로 다른 두 해시의 비트를 비교하여 해밍거리를 구한다.
  5. 해밍거리가 0이면 두 해시가 동일하다, 즉 두 이미지가 동일하다는 것을 의미한다.

1. 개발 일정



1. 참고 문헌

[1]-<http://biz.chosun.com/site/data/html_dir/2018/10/04/2018100400109.html>

[2]-http://www.zdnet.co.kr/view/?no=20181002145050

[3]-<https://www.boannews.com/media/view.asp?idx=73846>

[4]-[https://m.blog.naver.com/PostView.nhn?blogId=zenmode&logNo=150175674967&proxyReferer=http%3A%2F%2Fwww.google.com%2Furl%3Fsa%3Dt%26rct%3Dj%26q%3D%26esrc%3Ds%26source%3Dweb%26cd%3D7%26ved%3D2ahUKEwjuhr2xmpXhAhWNvpQKHVKHD6MQFjAGegQIBRAB%26url%3Dhttp%253A%252F%252Fm.blog.naver.com%252Fzenmode%252F150175674967%26usg%3DAOvVaw28bK8NVDYDD3G0F3OT5HFg](https://m.blog.naver.com/PostView.nhn?blogId=zenmode&logNo=150175674967&proxyReferer=http://www.google.com/url?sa%3Dt%26rct%3Dj%26q%3D%26esrc%3Ds%26source%3Dweb%26cd%3D7%26ved%3D2ahUKEwjuhr2xmpXhAhWNvpQKHVKHD6MQFjAGegQIBRAB%26url%3Dhttp://m.blog.naver.com/zenmode/150175674967%26usg%3DAOvVaw28bK8NVDYDD3G0F3OT5HFg)

[5]-<https://m.post.naver.com/viewer/postView.nhn?volumeNo=16585222&memberNo=25598567>

[6]-https://www.earticle.net/Article/A263848 – 이미지를 이용한 웹사이트 위•변조 탐지

[7]-PENG YANG, GUANGZHEN ZHAO , AND PENG ZENG, “Phishing Website Detection Based on Multidimensional Features Driven by Deep Learning”, **Published in:**[IEEE Access](https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=6287639) ( Volume: 7 ), 11 January 2019

[8]-F.C. Dalgic1, A.S. Bozkir 2\* and M. Aydos 3, “Phish-IRIS: A New Approach for Vision Based Brand Prediction of Phishing Web Pages via Compact Visual Descriptors”, **Published in:**[2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)](https://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=8543123), 19-21 Oct. 2018

[9]-http://solarisailab.com/archives/2351

[10]-<https://yamerong.tistory.com/40>

[11]-https://jsideas.net/Inception\_v3\_transfer\_learning/

[12]-https://www.pyimagesearch.com/2017/11/27/image-hashing-opencv-python/

**GitHub URL-https://github.com/yjinsol/grad\_project.git**