서포트 벡터 머신 기반 비디오 조각파일 분류

강현석¹, 이영석^{1*}
¹청운대학교 전자공학과

Support Vector Machines-based classification of video file fragments

Hyun-Suk Kang¹, Young-Seok Lee^{1*}

¹Department of Electronics, Chungwoon University

요 약 BitTorrent는 다수의 파일 공유자들로부터 조각파일을 전송 받아 하나의 완전한 파일을 완성할 수 있는 파일 공유 및 전송과 관련된 혁신적인 프로토콜이다. 그러나, 불법 또는 저작권과 관련된 비디오 데이터들이 임의로 배포되는 범죄행위가 발생하는 것이 현실이다. BitTorrent 상에 데이터에 대한 저작권 단속의 어려움은 데이터의 전송형태가 완전한 파일 형식이 아닌 조각 파일 형태로 전송된다는 점이다. 따라서, BitTorrent에서 얻어진 조각파일에서 디지털 콘텐츠를 복원하고, 저작권 위반 여부를 판단하기 위해서는 디지털 콘텐츠의 파일 형식에 대한 분류 과정이 선행 되어야 한다. 본 연구에서는 디지털 파일의 형식을 분류하기 위한 방법으로서 조각파일의 히스토그램 차분을 특징 벡터로 하는 SVM 분류기를 제안 하였다. 제안한 분류기는 3종류의 비디오 파일 형식에 적용하여 분류율로 성능을 평가하였다.

Abstract BitTorrent is an innovative protocol related to file-sharing and file-transferring, which allows users to receive pieces of files from multiple sharer on the Internet to make the pieces into complete files. In reality, however, free distribution of illegal or copyright related video data is counted for crime. Difficulty of regulation on the copyright of data on BitTorrent is caused by the fact that data is transferred with the pieces of files instead of the complete file formats. Therefore, the classification process of file formats of the digital contents should take precedence in order to restore digital contents from the pieces of files received from BitTorrent, and to check the violation of copyright. This study has suggested SVM classifier for the classification of digital files, which has the feature vector of histogram differential on the pieces of files. The suggested classifier has evaluated the performance with the division factor by applying the classifier to three different formats of video files.

Key Words: BitTorrent, histogram differential, SVM

1. 서론

웹하드, P2P사이트와 같은 파일, 정보공유 사이트가 급속도로 발전하고 늘어나면서 BitTorrent 같은 프로토콜이 등장하였다. 각 peer들 간에 효과적인 파일공유 및 전송을 위하여 제안된 BitTorrent 프로토콜은 인기 있는 파일 프로토콜 중 하나로서 URL에 의해 내용을 확인하고 웹과 통합될 수 있도록 설계 되었다. 그러나 BitTorrent의 순기능과 함께 저작권이 있는 영화, 음악,

소프트웨어 등이 불법으로 유통되고 확산되면서 사회적 으로 큰 손해를 주고 있는 것이 현실이다.

본 논문에서는 BitTorrent 환경에서 불법으로 유통되는 저작물을 탐색하고, 밝혀내는 기술적인 방법의 하나로서 BitTorrent 환경에서 얻어지는 조각파일들의 형식을 알아내기 위한 방법을 제안하였다. BitTorrent의 전송특성상 비디오 파일이 조각 형태로 존재하기 때문에 이조각파일에서 비디오 콘텐츠를 복구하기 위해서는 조각파일의 형식을 밝혀내는 것이 우선이다. 따라서 본 연구

이 연구는 문화체육관광부와 한국저작권위원회 2014년도 저작권 기술개발사업의 지원으로 이루어졌음.

*Corresponding Author : Youngseok Lee(Chungwoon University)

Tel: +82-32-770-8225 email: yslee@chungwoonac.kr

Received November 14, 2014 Revised (1st December 18, 2014, 2nd January 5, 2014) Accept

에서는 수집한 조각 파일 형식의 특징을 확인함으로써, 비디오 파일 조각의 형식을 밝혀내기 위한 기술을 제안 하였다. 비디오 파일의 종류를 인식하기 위하여 알 수 없 는 이진 비디오 파일들로부터 통계적인 특징을 추출하여 분류하는 방법이 포함되었으며 비디오 파일의 형식을 추 정할 수 있는 알고리즘을 제안하였다.

2. 이론적 배경

2.1 BitTorrent상에서의 불법 다운로드

BitTorrent는 2001년 4월부터 설계되어, 2001년 7월 2 일 미국의 소프트웨어 개발자 Bram Cohen에 의해 발표 되었다. BitTorrent는 파일을 공유할 수 있는 프로토콜을 제공하고 이를 통하여 파일공유가 가능할 수 있도록 동 작한다. BitTorrent는 분산 해시 테이블을 활용하여 오픈 소스 저작권에서 배포하였으며, 이와 호환된 uTorrent 등 각종 클라이언트(client)들이 만들어져 널리 이용되고 있다. BitTorrent는 크게 클라이언트와 트랙커(tracker) 로 구성이 되며, 클라이언트에서 생성된 .torrent 파일을 웹페이지 또는 웹하드 등으로 공유하여 서비스가 시작된 다. 즉, 공유하려는 파일을 가진 자가 자신의 컴퓨터에 설 치된 BitTorrent 클라이언트를 통해 .torrent 파일을 생성 하고 이를 공유함으로써 시더(seeder) 역할을 하게 되며, 이를 다운로드 받은 사람들이 다시 .torrent 파일을 공유 함으로써 여러 사람이 하나의 파일을 여러 컴퓨터에서 다운로드 받을 수 있는 형태를 가지게 된다.

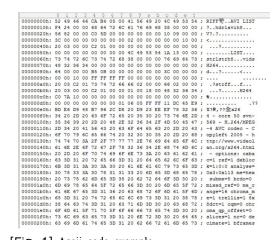
이러한 BitTorrent를 이용하면 인터넷 상에 저장되어 있는 파일을 다수의 접속을 통하여 동시에 여러 곳에서 여러 조각파일을 다운받을 수 있기 때문에 다운로드 속도가 향상되는 효과를 얻을 수 있다[1]. BitTorrent의 상용화로 인하여 음악이나 영화, 드라마, 소프트웨어 등과같은 파일들의 전송이 가능해 지면서 디지털화된 저작물들이 순식간에 엄청난 규모의 침해를 당하는 문제가발생하였다. BitTorrent에서는 조각파일 형식의 이진 데이터들이 다운로드 되기 때문에 조각파일로부터 비디오파일 형식을 인식 하고 이를 복구하여 불법복제 여부를확인하는 것은 매우 어려운 일이다.

따라서 본 연구에서 영상물 불법복제를 단속하기 위한 첫 번째 단계로서 BitTorrent상의 조각파일들로부터비디오 파일의 형식을 분류하는 것이 본 논문의 목적이다.

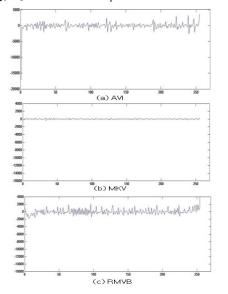
2.2 미디어 특징 추출

사용된 파일은 이진 데이터 형식을 취하고 있다. 본 연구에서는 파일 데이터를 Fig. 1과 같이 아스키 코드 형식으로 추출하였으며, 추출된 데이터는 히스토그램의 차분을 통한 특징을 추출하는데 사용하였다. Fig. 1은 비디오데이터로부터 추출한 아스키 코드의 일레이다.

각각의 비디오 형식의 분류를 위하여 BitTorrent 상에서 가장 많이 사용되고 유통되는 비디오 형식인 AVI, MKV, RMVB 비디오 형식이 적용되었다. Fig. 2는 각 비디오 파일 형식에 따라 추출한 특징 벡터들을 나타내고 있다. 동일한 비디오 콘텐츠에 대하여 구분 가능한 특징을 나타내고 있는 것을 알 수 있다.



[Fig. 1] Ascii code example



[Fig. 2] Feature Extraction

3. 파일 형식 분류를 위한 다중 SVM 사용의 제안

3.1 서포트 벡터 머신(SVM)

SVM(Support Vector Machine)은 패턴인식을 위한학습시스템의 한 분야로서 Vapnik에 의해 고안되었다. SVM은 초기에 주목 받지 못하다가 최근 생물정보학, 문자인식, 필기인식, 얼굴 및 물체인식 등 다양한 분야에서 사용되고 있으며 지도학습(supervised learning)에 의한패턴 분류에서 널리 사용된다[2]. SVM 분류기의 원리는 결정경계(decision boundary)와 훈련데이터 사이에 최대여유(margin)을 가지는 초평면(hyperplane)을 설계하는 것이 목적이다. 예를 들어 훈련데이터 집합 $(x_1,y_1),\cdots,(x_\iota,y_\iota)$, 입력변수 $x_\iota\!\in\!R^N$ 이고 클래스레이블(class label) $y_i\!\in\!-1,\!+1$ 일 때, 선형분리가 가능한 문제에서 결정함수의 초평면은 법선벡터 w와 중심에서 초평면까지의 거리 b와 같다. 따라서 초평면과 가장 가까운 점의 여유가 최대화 되면 식 1과 같이 이차 최적화 문제로 나타낼 수 있다[3].

$$\min_{w} \frac{1}{2} (w^T w) \tag{1}$$

조건은 식 2와 같다.

$$y_i((w.x) + b) \ge 1, i = 1,...,l$$
 (2)

선형적으로 분리가 가능하지 않은 경우, 커널함수를 사용하여 특징공간에서 선형분리를 수행한다. SVM은 최대의 여유을 가지면서도 비선형 데이터 까지도 커널함 수에 의해 구분할 수 있다는 장점이 있다.

식 2에서 (w,b)로 스케일링 하면 초평면에 가장 가까운 점의 거리는 $\frac{1}{\parallel w \parallel}$ 과 같다. 실질적인 상황에서 완벽하게 분리 가능한 초평면은 존재하지 않으므로 허용 오차를 고려하여 slack변수 ξ_i 를 두면, 식 3과 같은 조건일 때 식 4와 같은 조건식을 얻는다.

$$\xi \ge 1, \dots, l \tag{3}$$

$$y_i((w^T w) + b) \ge 1 - \xi_i, \quad i = 1, ..., l$$
 (4)

따라서 식 3과 식 4와 같은 조건에서 최적화 문제는 다음 식과 같다.

$$\min_{w, \xi} \frac{1}{2} (w^T w) + C \sum_{i=1}^{l} \xi_i$$
 (5)

위 식에서 C는 미결정 상수로 여유의 최대화와 분류에러 사이의 균형을 조정하는 정규화 변수라 부른다. 식 2와 같은 제한 상황(constraint condition)에서 최적해를 구하는 문제는 라그랑쥬 승수법(Lagrange multiplier method)을 이용하여 식 6과 같이 라그랑쥬 승수 α_i 의 변화에 따른 w의 값을 결정하는 방법으로 해결할 수 있다.

$$w = \sum_{i=1}^{l} y_i \alpha_i x_i \tag{6}$$

대부분의 경우 $\alpha_i \neq 0$ 이고, 해당하는 x_i 의 쌍들은 서포트 벡터들로 알려져 있으며 완벽한 결정경계로 정의된다. 모든 훈련 샘플에서 α_i 값이 0에 상응한다면 자동적으로 조건 식 2에 의해 슬랙변수 ξ_i 는 0이 되며, 또한벡터 x에 대한 초평면 결정함수는 식 7과 같이 나타난다.

$$f(x) = sgn\left[\sum_{i=1}^{l} y_i \alpha_i(w^T w_i) + b\right]$$
 (7)

위 식에서 $sgn(\bullet)$ 은 시그넘(signum) 함수이다. 또한 벡터의 내적 (x^Tx_i) 을 kernel 함수 $K(x^Tx_i)$ 로 변경하면 입력데이터는 높은 차원의 공간으로 매핑된다. 이 의미는 낮은 차원에서 초평면에 의해 분리가 되지 않는다면 높은 고차원에서는 초평면에 의해 최대 여유를 가지는 데이터로 분리되는 것을 의미한다. 2-class만 분류하는 SVM의 경우에는 y_i 은 -1,1로 정의 되었지만 k class를 갖는 One-Again-One SVM의 경우에는 y_i 은 -1,k로 정의할 수 있으며 간단한 수학적 조작을 통해 2-class SVM 분류기로 확장할 수 있다!41.

본 연구에서는 비디오 조각 파일의 히스토그램의 차 분을 특징 벡터로 하는 multi-class SVM을 제안하였고 다음과 같은 실험과정을 거쳐 비디오 파일 형식을 분류 하고자 한다.

4. 실험방법 및 결과

4.1 학습 및 테스트 데이터 구성

취득한 전체 포맷 480개의 데이터 중에서 AVI, MKV, RMVB 포맷별로 각각 80개씩 240개의 데이터를 학습 데이터로 사용하였으며, 나머지 240개의 데이터를 테스트데이터로 사용하였다.

4.2 학습 및 테스트 결과

본 연구의 실험에서는 형식 별로 취득한 총 480개의 비디오 데이터를 학습 데이터와 테스트 데이터를 분류하고, 학습 후 multi-class SVM 분류 기법을 적용하여 비디오 파일 분류율을 분석하였다.

[Table 1] Formulation used for kernel functions

Kernel	K(x, y)
Linear	x·y
Polynomial	(x·y+1) ^d

Table 1은 본 실험에서 가장 좋은 성능을 보여주는 두 커널 함수들을 나타내고 있다[5]. 이 커널들외에 RBF(radial basis function) 커널의 경우에는 위의 두 커널에 비하여 낮은 성능을 나타내어 비교에서 제외하였다. Table 2와 Table 3은 각각 kernel 함수 사용별, 비디오 파일 길이별 각각의 분류율을 나타낸다.

[Table 2] Linear kernel used in classification rate

	3MB		avg	2MB		avg
	82%	84%		83%	68%	
	84%	80%		76%	68%	
RMVB	79%	85%	83%	72%	62%	75%
	82%	87%		88%	82%	
	84%	82%		79%	68%	
	65%	43%	66%	68%	68%	64%
	64%	74%		74%	74%	
AVI	66%	73%		71%	78%	
	53%	69%		81%	9%	
	73%	76%		64%	53%	
	88%	80%	84%	76%	93%	85%
MKV	83%	88%		90%	85%	
	85%	78%		95%	85%	
	85%	90%		78%	83%	
	80%	83%		78%	90%	

	1MB		avg	512KB		avg
RMVB	79%	70%	74%	55%	71%	68%
	77%	76%		66%	70%	
	72%	78%		66%	68%	
	72%	67%		67%	62%	
	78%	73%		74%	82%	
	61%	64%	62%	31%	50%	46%
	74%	51%		66%	35%	
AVI	70%	73%		61%	49%	
	58%	74%		55%	16%	
	43%	56%		74%	26%	
MKV	98%	80%	90%	98%	98%	95%
	83%	98%		95%	98%	
	93%	90%		90%	93%	
	98%	93%		85%	95%	
	88%	83%		98%	98%	

Table 2의 linear kernel 함수를 사용 시 RMVB 파일 의 분류율은 10번의 실험결과의 평균 분류율이 3MB일 때 83%, 2MB의 평균 분류율이 75%이고, 1M의 평균 분 류율이 74%, 512KB 파일 길이의 분류율은 68%의 평균 분류율을 보였다. AVI 파일의 분류율은 10번의 실험결과 의 평균 분류율이 3MB일 때 66%, 2MB의 평균 분류율이 64%이고, 1M의 평균 분류율이 62%, 512KB 파일 길이의 분류율은 46%의 평균 분류율을 보였다. MKV 파일의 분 류율은 10번의 실험결과의 평균 분류율이 3MB일 때 84%, 2MB의 평균 분류율이 85%이고, 1M의 평균 분류 율이 90%, 512KB 파일 길이의 분류율은 95%의 평균 분 류율을 보였다. Linear kernel 함수를 사용했을 때 AVI는 다소 떨어지는 분류율을 보이는 반면에 RMVB와 MKV 의 분류율은 상당한 분류율을 보이고 있다. Table 3에 나 타낸 바와 같이 polynomial kernel 함수를 사용시 RMVB 파일의 분류율은 10번의 실험결과의 평균 분류율이 3MB 일 때 92%, 2MB의 평균 분류율이 93%이고, 1M의 평균 분류율이 90%, 512KB 파일 길이의 분류율은 89%의 평 균 분류율을 보였다. AVI 파일의 분류율은 10번의 실험 결과의 평균 분류율이 3MB일 때 54%, 2MB의 평균 분류 율이 62%이고, 1M의 평균 분류율이 48%, 512KB 파일 길이의 분류율은 47%의 평균 분류율을 보였다. MKV 파 일의 분류율은 10번의 실험결과의 평균 분류율이 3MB일 때 87%, 2MB의 평균 분류율이 82%이고, 1M의 평균 분 류율이 84%, 512KB 파일 길이의 분류율은 85%의 평균 분류율을 보였다.

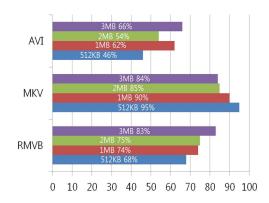
[Table 3]	Polynomial	kernel	used	in	classification	rate
TIADIE OF	roiviioiiiiai	KCHICI	uscu	ш	Ciassification	Taic

	ЗМВ		avg	2MB		avg
RMVB	93%	89%		95%	96%	
	90%	91%		89%	95%	
	90%	90%	92%	90%	94%	93%
	93%	93%		93%	96%	
	95%	93%		90%	87%	
	71%	50%	54%	60%	64%	62%
AVI	65%	60%		61%	55%	
	36%	41%		61%	65%	
	48%	49%		54%	58%	
	69%	49%		68%	70%	
MKV	85%	63%	81%	90%	90%	
	80%	85%		73%	83%	
	90%	85%		93%	93%	82%
	85%	80%		73%	71%	
	83%	78%		83%	71%	

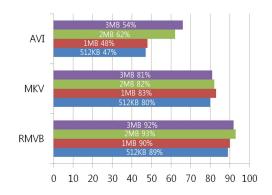
	1MB		avg	512KB		avg
RMVB	89%	91%		89%	88%	89%
	93%	90%		87%	89%	
	91%	93%	90%	91%	93%	
	87%	84%		87%	90%	
	89%	94%		87%	88%	
	51%	34%	48%	48%	24%	47%
	63%	6%		50%	64%	
AVI	64%	56%		46%	63%	
	69%	45%		59%	54%	
	80%	10%		56%	5%	
MKV	95%	80%	83%	88%	63%	
	82%	88%		85%	95%	
	85%	80%		85%	80%	80%
	76%	88%		83%	76%	
	85%	70%		68%	73%	

Polynomial kernel 함수를 사용했을 때는 linear kernel 함수를 사용했을 때와 마찬가지로 AVI의 분류율은 약간 떨어지는 반면에 RMVB와 AVI의 경우에는 높은 분류율 을 보이고 있다.

Fig. 3과 Fig. 4는 각각의 파일의 커널별, 파일 길이 별 분류율의 평균을 도수분포표로 나타낸 것이다.



[Fig. 3] Classification ratio for 3, 2, 1, and 0.5 Mbyte video file fragments (linear kernel case)



[Fig. 4] Classification ratio for 3, 2, 1, and 0.5 Mbyte video file fragments (polynomial kernel case)

4. 결론

본 연구에서는 이진 분류에 적합한 분류성능을 가지고 있는 SVM 알고리즘을 조합한 multi-class SVM을 이용하여 서로 다른 형식의 비디오 조각 파일들의 분류성능을 향상시키는 새로운 방법을 연구했다. 비디오 파일형식인 RMVB, AVI, MKV의 정확한 특징을 추출하기위하여 각각 포맷의 아스키 코드 값을 추출하고 이로부터 히스토그램의 차분의 특징을 추출하였다. 형식별로취득한 총 480개의 학습 데이터와 테스트 데이터를 각각 multi-class SVM 분류 기법을 적용하여 linear kernel과 polynomial kernel을 사용하였을 때 AVI는 linear kernel에서의 평균 분류율이 56.50%, polynomial kernel 평균 분류율이 50.28%로 만족스러운 결과를 얻지 못했지만

RMVB 파일 형식은 커널별로 89.92%, 71.02% MKV 파일 형식은 커널별로 81%, 90.60%의 높은 분류율을 얻었다.

본 논문은 BitTorrent 환경에서 불규칙하게 공유되는 조각파일 만을 가지고 비디오 파일의 형식을 분류하는 방법을 처음으로 시도하였고 비디오 조각 파일만으로도 비디오 형식을 분류할 수 있다는 잠재적인 가능성을 보 였다.

따라서 본 연구는 불규칙한 조각 형태로 존재하는 비디오 파일이 분포된 BitTorrent 공간에서나 P2P사이트 공간 등에서 파일 식별을 위한 분류를 가능하게 함으로서 네트워크상에서 영화, 음악, 소프트웨어 등이 불법으로 유통되고 확산 되는 문제를 탐지하고 예방하는데 크게 기여할 수 있을 것으로 기대된다.

References

- [1] Kim. J. W., "State of the Art of the Countermeasure Technology for Copyright Infringement in the BitTorrent", Intellectual Property Right, Sangmyung Uvi, vol. 11, pp. 43–50, 2013.
- [2] P. Konar and P. Chattopadhyay, "Bearing fault detection of induction motor using wavelet and support vector machines (SVMs)," Applied Soft Computing, vol. 11, pp. 4203–4211, 2011.

DOI: http://dx.doi.org/10.1016/j.asoc.2011.03.014

- [3] I. S. Lee, J. H. Cho, H. M. Seo, and Y. S. Nam, "Fault detection and isolation of nonlinear system based on SVM and artificial neural network," 2012 27th ICROS Annual Conference (in Korean), pp. 408–409, 2012.
- [4] Yang, B. S., Hwang, W. W., Kim, D. 1. and Tan, A. C., 2005, "Condition Classification of Small Reciprocating Compressor for Refrigerators using Artificial Neural Networks and Support Vector Machines," Mechanical Systems and Signal Processing, Vol. 19, PP.371–390, 2013. DOI: http://dx.doi.org/10.1016/j.ymssp.2004.06.002
- [5] U. Kreße1. "Pairwise Classification and Support Vector Machines, in Advances in Kernel Methods-Support Vector Learning," B. Scholkopf, C. 1. C. Burges, A. J. Smola. Eds. MIT Press, Cambridge. PP. 255-268, 1999.

강 현 석(Hyunsuk Kang)

[준회원]



- 2006년 3월 ~ 2013년 2월 : 청운 대학교 전자공학과 (공학사)
- 2013년 3월 ~ 2015년 2월 : 청운 대학교 정보산업대학원 전자공학과 (공학석사)

<관심분야> 신호처리, 임베디드 시스템

이 영 석(Youngseok Lee)

[정회원]



- 1995년 2월 : 서울시립대학교 대학 원 전자공학과 (공학석사)
- 1998년 2월 : 서울시립대학교 대학 원 전자공학과 (공박석사)
- 1998년 3월 ~ 현재 : 청운대학교 전자공학과 교수

<관심분야> 신호처리, 임베디드 시스템