폭력 행위 인식 기술 동향

곽 수 영*

1. 서 론

최근 우리나라는 살인, 강도 등 강력범죄 발생이 급증하는 등 사회적으로 불안감이 고조되고 있다. 특히, 성폭력, 학교폭력, 가정폭력 등 4대 사회악 관련 범죄가 심각한 사회문제로 대두되고 국민 생활의 불안 원인으로 지적되고 있다. 이러한 범죄로부터 안전한 생활이 가능하도록 하기 위해 정부에서는 실시간 상황인식이 가능해 위험의 적시 경보 및 대처가 가능한 CCTV 통합관제센터를 구축하고 있다[1]. 도심 또는 농어촌의 범죄가 많은 우범지역과 학교 앞 및 도심공원, 놀이터 등 어린이 보호구역에 성능이 좋은 방범용 CCTV 2만 9천여 대를 추가 설치하여 어린이 성폭력 및 각종 강력범죄 예방 기능을 강화한다는 것이다. 이와 함께 범죄자 행동 패턴 인식 기능 등, 카메라 자동 추적 기능 등 CCTV 최첨단 기술을 접목하여 범죄 및 사고현장을 과학적으로 감시한다는 방안도 제시되고 있다[2].

그 중에서도 학교폭력은 집단적이고 반복적으로 이루어지고 있으며, 그 가해 및 피해 연령이 점점 낮아지고 있다. 정부는 이러한 학교폭력의 최근 경향에 대응하기 위해서 CCTV를 설치하고 센터 지킴이들이 언제든지 화면을 모니터링 해 필요 시 즉각적인 조치를 취할 수 있는 환경을 만들고 있다. 학교뿐만 아니라 어린이집 폭행도 마찬가지다. 인천의 한 어린이집 CCTV에서 확인된 보육교사의 아동폭행 사건은 전국민의 공분과 함께 어린이집의 CCTV 의무 설치에 관한 사회적 논쟁을 불러왔다. 어린이집 폭행 재발방지 대책의 일환으로 CCTV를 강제 설치뿐만 아니라 실시간 모니터링이 가능한 시스템까지도 요구되고 있다[3].

이러한 폭행 사건으로 CCTV설치 지역이 많아지고 있지만 이를 장시간 모니터링하여 실시간으로 폭력 행동검출을 통한 예방에는 한계가 있다. 실제로 모니터 관제요원이 30분 이상 모니터링을 할 경우 발생하는 사건의 90%이상을 검출하지 못한다는 결과가 있다. 이런 문제점 해결을 위해 감시영상에서 폭력행위를 자동으로 검출하는 지능화된 기술에 대한 관심이 높아지고 있다. 따라서 본 기고에서는 영상 분석을 통해 폭력 행동을 인식하는 방법에 대한 동향을 소개하고자 한다. 기존의 개발된 폭력행위 검출 방법들은 크게 객체기반 방법과 학습기반 방법으로 나눌 수 있다.

※ 교신저자(Corresponding Author) : 곽수영, 주소 : (34158) 대전광역시 유성구 동서대로 125, 한밭대학교 전자·제어공학과 전화: 042)821-1167 FAX: 042)821-1164,
E-mail :sykwak@hanbat.ac.kr

2. 폭력행위 분석

초기의 폭력행위 검출에 대한 연구는 영화에서 폭력 장면을 검출하기 위해 시작되었다. 영화에서 촬영된 폭력 장면에는 주로 피를 흘리는 모습들이 자주 등장하기 때문에 움직임 정보와 색상 정보를 함께 사용하는 방법을 많이 사용하였다[4]. 하지만 실제 환경에서 촬영된 폭력행위의 경우 피가 묻은 장면은 자주 발생하지 않기 때문에 색상 정보를 이용한 폭력행위 검출에는 한계점이 존재한다. 본 기고에서는 실제 폭력행위가 발생하는 감시카메라와 같은 영상에서 분석하는 방법에 대해 초점을 맞추었다.

2.1 객체 기반의 폭력행위 분석

객체기반의 폭력행위 분석방법은 움직이는 사람을 검출하고, 이들로부터 분석된 특징을 이용하여 폭력행위인지 아닌지를 판단하는 방법이다. 초기에는 사람과 사람의 상호작용을 분석하여 긍정적인 상호작용인지 부정적인 상호작용인지를 분류하는 방법을 사용하였다. 상호작용 분석을 위해 먼저 사람의 실루엣을 추출하였고 추출된 실루엣으로부터 머리,팔,다리,몸통 등 사람 몸의 주요 파트를 분리하여 각 파트별 움직이는 방향, 궤적 등을 분석하여 폭력행위를 분석하도록 하였다.

Datta가 제안한 방법은 배경 모델링을 통하여 사람을 배경으로부터 분리한 뒤, 분리된 사람을 그림 1과 같이 4가지 파트로 나누고, 팔과 다리의 움직임 방향성을 분석하는 방법을 제안하였다[5]. 사람과 사람이 싸우는 행위를 할 때 팔을 앞으로 뻗고 다리를 앞으로 차는 행위를 하기 때문에 팔과 다리의 움직임이 발생할 수밖에 없고 이들의 방향정보가 땅의 지면과 평행하게 나타날 가능성이 높다. 이 정의에 팔 다리 움직임의 방향성 정보 분석을 통해 폭력행위 검출 알고리즘도 제안하였다. Datta가 제안한 방법은 그림 2에서 보듯이 사람과 사람 즉 두명의 사람이 상호작용으로 나타나는 행동 중 폭력 행위를 유한오토마타로 정의 하였고 정의된 룰에 순차적으로 발생이 이루어질 때 그림 3과 같이 폭력행위가 검출된다. 이러한 방법은 카메라 방향이 고정되어 있는 경우에는 쉽고 빠르게 폭력 행위를 검출할 수 있지만 카메라가 촬영하는 방향에 따라 사람의 모습이 달리 보이고 방향도 다르기 카메라 뷰에 독립적이지 못하다는 단점이 있다.

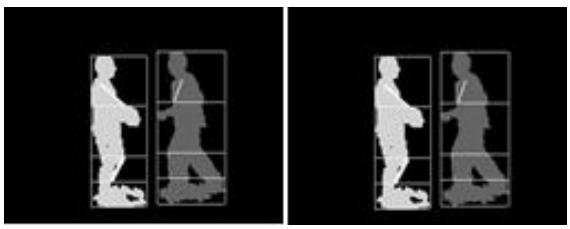


그림 1 배경모델링을 통한 사람 검출 및 4가지 파트 분류 결과

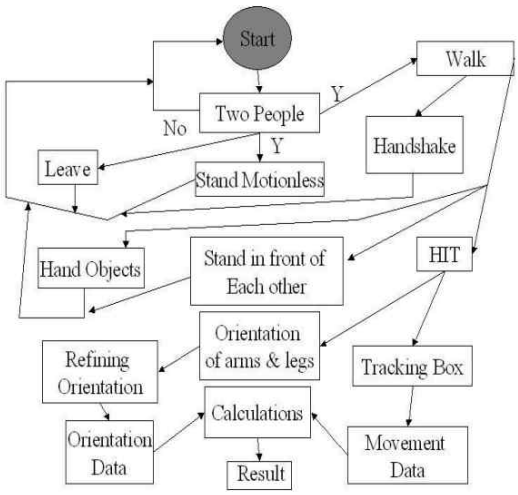


그림 2 Datta가 제안한 폭력행위 검출을 위한 유한 오토마타

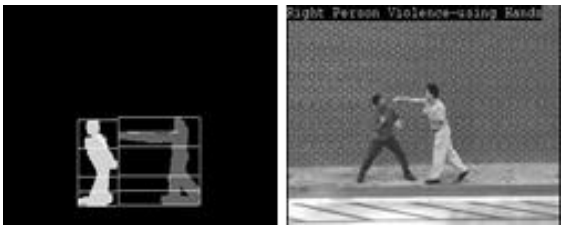


그림 3 팔과 다리의 방향 성분으로 폭력행위를 검출한 결과 영상

Ryoo는 그림 4와 같이 사람의 실루엣을 검출한 뒤 실루엣에서 상체와 하체 파트를 분리하여 각 파트별로 특징정보들을 분석하도록 하였다. 분석된 정보들은 미리 정의해 둔 폭력행위에 대한 Context-Free Grammar(CFG)의 표현법과 일치하는지를 판단하여 폭력 행위를 검출하는 방법을 제안하였다[6]. 그림 5는 폭력행위에 대한 CFG 예이다. 이러한 객체기반의 폭력행위 분석 방법은 영상분석의 첫 번째 단계인 사람 검출에서 실패하는 경우 분석이 제대로 이루어 질 수 없고, 카메라가 사람을 비추는 방향에 따라 특징 분석이 다르기 때문에 시점에 독립적이지 못하다는 단점이 있다. 또한, 사람을 검출하고 파트를 분리하는 단계에서 처리시간이 오래 소요될 수밖에 없기 때문에 실시간 처리에는 어려움이 있다.

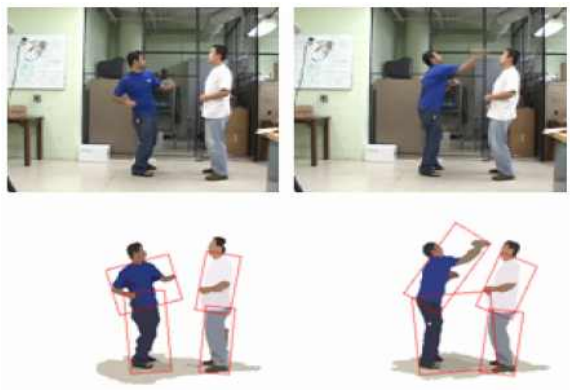
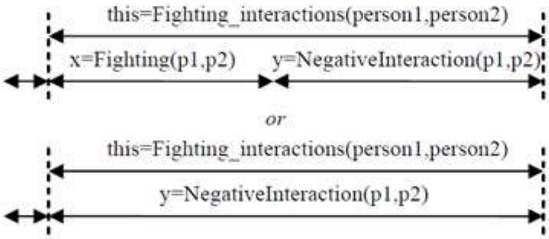


그림 4 실루엣 검출 결과 및 폭력행위 분석 결과




```
NegativeInteraction(i, j) = (
    list( def('x', PunchingInteraction(i, j)),
          list( def('y', KickingInteraction(i, j),
                def('z', PushingInteraction(i, j)) )),
    or(   equals('this', 'x'),
          or( equals('this', 'y'), equals('this', 'x') ))
);
FightingInteraction(i, j) = (
    list( def('x', FightingInteraction(i, j)),
          def('y', NegativeInteraction(i, j)) ),
    or( equals('y', 'this'),
        and(meets('x', 'y'),
            and(starts('x', 'this'), finishes('y', 'this'))))
);
```

그림 5 폭력행위에 대한 Context Free Grammar

2.2. 학습기반의 폭력행위 분석

최근에는 폭력행동에서만 나타나는 특징정보들을 서술자로 기술하고, 이를 학습시켜 폭력과 비폭력 동영상을 분류하는 방법에 대한 연구가 많이 진행 중에 있다. 폭력행위가 있는 장면들은 YouTube, 영화, 감시카메라 등으로 촬영된 비디오를 수집하고, 폭력행위가 아닌 비디오는 운동을 하거나 일상생활을 하는 장면을 수집하여 학습에 사용하여 폭력에 적합한 서술자 개발에 주안점을 두고 있다. 이러한 방법은 처리시간이 객체기반의 방법보다 빠르다는 장점이 있어 실시간 처리가 주요 이슈인 감시 시스템 등과 같은 응용분야에 많이 사용되는 방법이다.

2.2.1 MoSIFT(Motion SIFT)

Bermejo은 폭력행위 검출을 위해 MoSIFT 서술자를 제안하였다[7]. MoSIFT는 Motion SIFT(Scale-Invariant Feature Transform)의 약자로 움직임이 강한 부분에서 SIFT알고리즘으로 주요 특징점을 추출하여 추출된 워드에 대해 HOG(Histogram of Oriented Gradients)와 HOF(Histogram of Oriented optical Flow)를 결합한 서술자이다. 이는 폭력행위에서 나타나는 외형 정보와 움직임 정보의 크기 및 방향 정보를 함께 서술자로 표현한 방법이라고 설명할 수 있다. 그림 6은 MoSIFT 서술자를 생성하는 전체 흐름도를 나타낸 것이다. 먼저 특징점 검출을 위해 영상을 다양한 크기로 변환하여 SIFT알고리즘으로 주요한 특징점을 검출하고 추출된 주요 특징점 마다 HOG와 HOF 서술자를 결합하여 MoSIFT 서술자를 생성하게 된다. 이렇게 생성된 서술자는 SVM(Support Vector Machine) 이진 분류기의 입력값으로 적용된다.

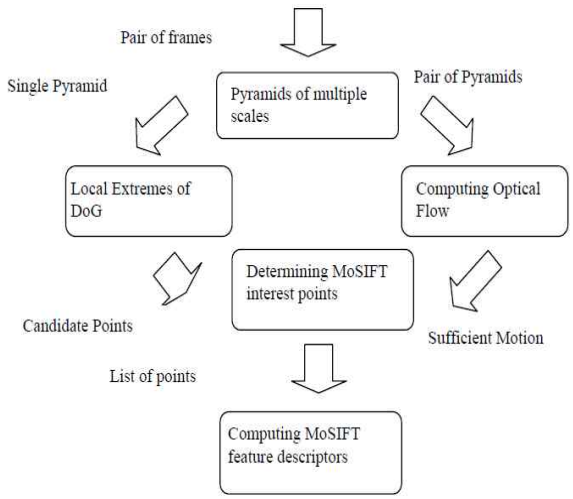


그림 6 MoSIFT(Scale-invariant feature transform) 서술자 생성의 흐름도

그림 7 폭력 검출에 사용한 학습 데이터

그림 7은 Bermejo가 폭력행위 검출을 위해 학습할 때 사용한 데이터를 나타낸 것이다. 이는 아이스하키 리그를 촬영한 비디오로 몸싸움 하는 부분에서 폭력행위 서술자를 추출하여 학습에 사용하였다.

2.2.2 ViF(Violence Flows)

Hassner은 그림 8에 나타낸 것처럼 훌리건 등과 같이 많은 사람들의 폭행 검출에 초점을 맞추었다. Hassner는 군중의 폭력행위를 검출하기 위해 ViF(Violent Flows) 서술자를 제안하였다[8]. ViF 서술자는 모든 픽셀의 움직임의 크기값($m_{x,y,t}$)을 수식 (1)과 같이 계산한 다음, 수식(2)와 같이 이전 프레임의 움직임의 크기와 다음 프레임의 움직임의 크기정보를 비교하여 움직임 크기의 변화가 큰 부분($b_{x,y,t}$)을 검출한다. 이후 수식(3)과 같이 검출된 부분에서만 움직임의 크기변화 발생하는 부분만 T프레임 누적($\overline{b_{x,y}}$)하여 이를 ViF 서술자로 나타내었다. 서술자 생성과정에서 보듯이 움직임의 크기변화가 큰 부분을 누적하여 사용하는 방법이기 때문에 카메라의 시점이 변하지 않는 영상에서만 적용이 가능하다.

$$m_{x,y,t} = \sqrt{(u_{x,y,t}^2 + v_{x,y,t}^2)} \quad (1)$$

$$b_{x,y,t} = \begin{cases} 1 & \text{if } |m_{x,y,t} - m_{x,y,t-1}| \geq \theta \\ 0 & otherwise \end{cases} \quad (2)$$

$$\overline{b_{x,y}} = \frac{1}{T} \Sigma_t b_{x,y,t} \quad (3)$$

ViF 서술자는 쉽고 빠르게 구성이 가능하기 때문에 실시간 검출에 유용하고, 훌리건과 같이 비디오 안에 등장하는 많은 사람들의 폭력행위 검출에는 적합하지만, 카메라 앵글 변화가 심하고, 소수의 사람이 싸우는 행위 검출에는 실패할 확률이 높다. 그림 9은 ViF 서술자를 이용하여 다수의 폭력행위 검출에 사용한 폭력 데이터와 비폭력 데이터를 나타낸 것이다.

그림 8 훌리건 등과 같은 군중의 폭력행위 검출의 예

그림 9 군중의 폭력 행위 검출에 사용된 학습 데이터

2.2.3 SFA(Slow Feature Analysis)

Wang은 비디오에서 움직임이 있는 영역의 궤적을 추적하여 3차원 궤적 서술자를 제안하였다[9]. 그림 10는 Wang에 제안한 폭력 검출에 대한 흐름도를 나타낸 것이다. 흐름도에서 보듯이 먼저 입력된 영상에서 픽셀별로 궤적을 추출한 뒤, 이를 일정시간 이상 누적하여 3차원 큐브 형태인 Cuboid를 생성하게 된다. 이를 SFA(Slow Feature Analysis) 학습의 입력값으로 적용하게 되면 SFA는 느리게 변화하는 특징을 추출하게 된다. 즉, 시간적 변화가 느림 순서대로 특징을 추출하게 되며 이를 폭력행위의 특징정보를 사용하여 SVM학습을 통해 폭력 행위와 비폭력 행위로 분류하는 방법을 제안하였다. 그림 11은 Wang이 폭력행위 검출을 위해 사용한 학습데이터로 비폭력 행위 동영상의 경우 일상생활에서 주로 나타날 수 있는 행동의 영상을 사용하였다.

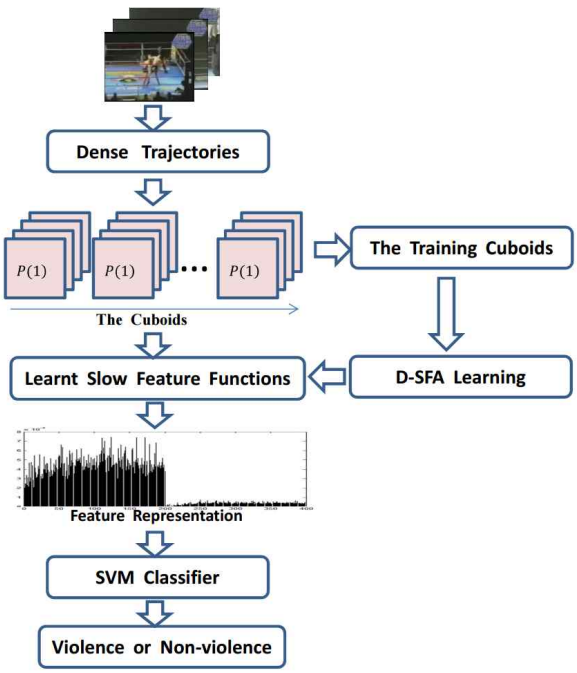


그림 10 Wang에 제안한 폭력행위 검출에 대한 흐름도

그림 11 Wang이 폭력행위검출에 사용한 학습 데이터 (상위 두줄의 영상: 폭력행위 학습 데이터, 하위 두줄의 영상: 비폭력행위 학습 데이터)

2.2.4 Interpersonal space

Rehg은 사람과 사람의 상호작용 중 폭력행위를 검출하기 위하여 두명의 사람이 상호작용할 때 사람과 사람의 공간 정보(Interpersonal space)를 사용하는 방법을 제안하였다[10]. 그림 12에서 보듯이 두명의 사람이 서서 이야기를 나누는 등의 상호작용을 보일 때에는 사람과 사람의 공간이 좁으면서 변화가 크게 나타나지 않은 상태를 유지하지만 싸움을 하거나 폭력행위가 발생할 때에는 일정 간격 이상 유지되면서 Interpersonal space의 간격 변화 발생할 가능성이 높다는 것을 특징정보로 사용하였다. 그림 14는 Rehg가 제안한 폭력행위 검출 시스템의 흐름도를 나타낸 것이다. 기본적으로 Bag of Words 기법을 사용하였지만 생성된 기본 코드북 정보에 제안하는 Interpersonal space정보를 추가적으로 사용하여 특징벡터를 구성하였고, 이를 학습의 입력데이터로 사용하였다. 실제 폭력행위 검출 데이터에 적용해 보기 위해 그림 13와 같이 2-3명 정도의 소수의 사람이 싸움을 하는 영상들을 사용한 것을 볼 수 있다.

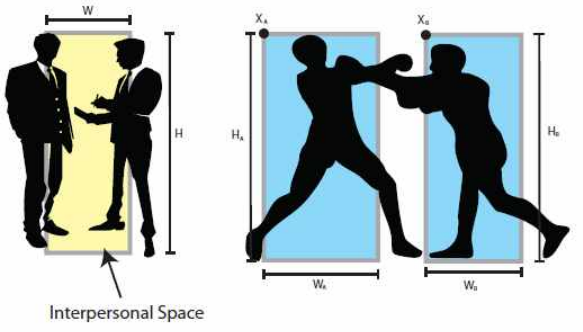


그림 12 폭력행위가 발생하지 않을 때와 발생할 때의 Interpersonal Space의 특징

그림 13 실제 폭력행위 검출에 사용된 데이터

3. 결 론

본 기고에서는 영상에서 폭력행위를 검출하는 방법들에 대해서 기술하였다. 초창기에는 영화 등과 같은 미디어 파일에서 폭력 동영상을 검출하기 위한 방법들이 많이 연구되었는데, 이러한 영상들은 연출된 영상이기 때문에 싸우는 소리도 크게 입력되어 있고, 시각적으로도 피를 많

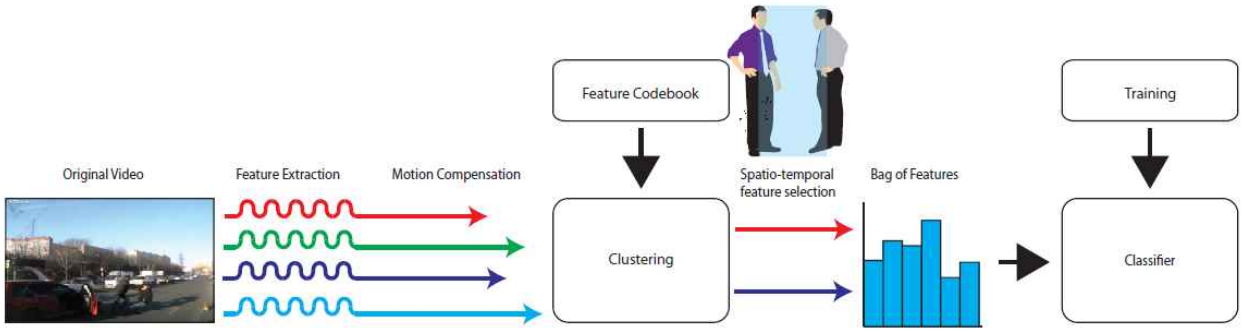


그림 14 Rehg가 제안한 폭력행위 검출의 시스템 흐름도

이 흘리는 장면이 등장한다. 초기에는 이런 음성 특징 정보와 피를 검출하기 위한 칼라 정보를 이용하여 폭력 영상을 구분하는 방법을 많이 사용하였으나 이는 실제 폭력 상황을 촬영한 영상 데이터는 적용하기 어려운 방법이라고 볼 수 있다. 폭력행위 검출을 위한 다양한 방법들 중 객체기반의 방법은 카메라 시점에 독립적이지 못하고, 배경과 사람을 분리해 내는 첫 번째 단계에서 오류가 발생하게 되면 분석 자체가 어렵다는 문제점 때문에 최근에는 학습기반의 방법을 많이 사용하고 있음을 알 수 있다. 학습기반의 방법의 경우 크게 사람과 사람의 인터랙션 즉, 상호작용을 기반으로 접근하는 방법이 있는 반면, 영상 전체에서 발생하는 움직임 정보를 다양한 서술자로 개발하여 폭력행위 검출하는 방법들을 제안하고 있다. 최근 연구 방법들은 연출되지 않은 실제 데이터에 적용이 가능한 방법들이 많이 제안되고 있지만, 아직까지 인식 성능에 대한 문제가 여전히 남아 있다고 볼 수 있다.

참 고 문 헌

[1] 범죄예방 대응체계 마련을 위한 ICT 활용방안, 한국정보화진흥원, 2013.12

[2] 전국은 지금 CCTV 통합관제센터 공사중, 월간 시큐리티월드 통권 제173호. 2011. 6

[3] 어린이집 CCTV 설치 의무화는 필요한가, 디베이팅데이, 2015. 1

[4] W. Zajdel, J.D. Krijnders, T. Andringa, and D.M. Gavrila, "CASSANDRA: audio-video sensor fusion for aggression detection", IEEE Int Conf. on Advanced Video and Signal Based Surveillance, pp. 200-205, 2007.

[5] A. Datta, M. Shah, N.D.V. Lobo, "Person-on-Person Violence Detection in Video Data", IEEE Int Conf. on Pattern Recognition, vol. 1, pp. 433- 438, 2002

[6] T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violence flows: real-time detection of violent crowd behavior", IEEE Int Conf. on Computer Vision and Pattern Recognition, pp. 1-6, 2012.

[7] E. Bermejo, O. Deniz, G. Bueno, R. Sukthankar. "Violence Detection in Video using Computer Vision Techniques", Proceedings of Computer Analysis of Images and Patterns, 2011.

[8] T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violence flows: real-time detection of violent crowd behavior", IEEE Int Conf. on Computer Vision and Pattern Recognition, pp. 1-6, 2012.

[9] K. Wang, Z. Zhang, L. Wang, "Violence video

detection by discriminative slow feature analysis“, The Chinese Conf. on Pattern Recognition pp. 137-144, 2012

[10] P.Rota, N.Conci, N.Sebe, J.M.Rehg, “Real-life violent social interaction detection“, IEEE Int Conf. on Image Processing, 2015

곽 수 영

- 2010년 2월 연세대학교 컴퓨터과학과 공학박사
- 2011년 1월 삼성전자 영상디스플레이사업부 책임연구원
- 2011년~현재 한밭대학교 전자·제어공학과 부교수
- 관심분야: 영상처리, 컴퓨터비전, 지능형시스템