범죄발생 요인 분석 기반 범죄예측 알고리즘 구현

박지호*, 차경현*, 김경호*, 이동창**, 손기준***, 김진영*

Implementation of Crime Prediction Algorithm based on Crime Influential Factors

Ji Ho Park*, Gyeong Hyeon Cha*, Kyung Ho Kim*, Dong Chang Lee**, Ki Jun Son***, and Jin Young Kim*

요 약

본 논문에서는 빅 데이터를 이용하여 범죄 발생 요인에 따른 범죄 예측 알고리즘을 구현했다. 제안된 알고리즘은 대검찰청에서 수집하여 공개한 범죄관련 빅 데이터를 사용하였으며, 통계분석을 통해 서울시의 2011-2013년 범죄발생 패턴을 분석했다. 범죄예측 알고리즘 구현을 위해 베이지안 네트워크를 적용하였으며, 범죄발생 요인으로서 공간적, 인구적, 사회적 특성 및 요일, 시간, 날씨와 같은 기타 요인으로 베이지안 네트워크의 노드를 구성하였다. 제안한 알고리즘의 구현 결과, 서울시의 각 구별로 범죄발생 패턴이 다르다는 것을 파악할 수 있었으며, 다양한 범죄발생 패턴을 분석하고, 범죄예측 알고리즘의 정확도를 확인할 수 있었다.

Key Words : Big Data, Crime Pattern, Crime Prediction, Bayesian Network, Bayesian Prediction.

ABSTRACT

In this paper, we proposed and implemented a crime prediction algorithm based upon crime influential factors. To collect the crime-related big data, we used a data which had been collected and was published in the supreme prosecutors'office. The algorithm analyzed various crime patterns in Seoul from 2011 to 2013 using the spatial statistics analysis. Also, for the crime prediction algorithm, we adopted a Bayesian network. The Bayesian network consist of various spatial, populational and social characteristics. In addition, for the more precise prediction, we also considered date, time, and weather factors. As the result of the proposed algorithm, we could figure out the different crime patterns in Seoul, and confirmed the prediction accuracy of the proposed algorithm.

I. 서 론

최근 폭발적으로 증가하는 디지털 정보에 따라 크기를 가늠할 수 없을 정도로 수많은 정보와 데이터가 생산되는 '빅 데이터(Big data)' 환경이 도래하고 있다. 빅 데이터란 기존 데이터를 수집, 저장, 관리, 분석할 수 있는 역량을 넘어 대량의 정형 또는 비정형 데이터 집합 및 이러한 데이터로부터 가치를 추출하고 결과를 분석하는 기술을 의미한다. 이러한 빅 데이터 환경에서 데이터는 그 거대한 양과 더불어 다양한 종류를 보여주는데, 이러한 데이터를 통해서 사람들의 행동은 물론 위치정보와 SNS를 통해 생각과 의견까지 분석하고 예측할 수 있게 되었다[1]. 이러한 빅 데이터의 중요성이 증가하면서 현재 정부는 내년부터 빅 데이터를 활용하여 실제 정책에 반영할 것을 추진하고 있다. 현재 빅 데이터를 접목할 수 있는 분야를 발굴하는 작업을 진행 중이며, 이미 미국, 유럽 각 국, 일본 등은 활발한 빅 데이터 도입을 통해 정책의 방향을 잡고 있다[2]. 한편, 본 논문에서 제안하는 범죄 예측의 경우, 빅 데이터를 활용하여 언제 어디서든 어떤 범죄가 발생하는지를 쉽게 예측 할 수 있다. 단순히 인적 정보만을 수집하고 분석하는 것이 아니라, 공간, 상황, 시간 정보를 수집하고 분석함으로써 언제 어느 지역에서 어떤 범죄가 발생할 것인지 통계적으로 예측 가능하다. 실제 이러한 빅 데이터 기반 범죄 예측 기술로, 미국 LA 경찰청은 절도 발생률을 33% 줄였으며, 미국 국세청에서는 빅 데이터를 기반으로 하여 탈세 및 사기 범죄 예방 시스템을 마련하였다. 빅 데이터 분석을 통한 이상 징후 발견과 과거의 행동 정보

※이 논문은 2014년 미래창조과학부의 재원으로 SW융합기술고도화 사업의 지원을 받아 수행된 연구임(S0170-15-1081).
*광운대학교 전자융합공학과 유비쿼터스 통신 연구실 (jihopark@kw.ac.kr, chagyeonghyeon@kw.ac.kr, gentle@kw.ac.kr, jinyoung@kw.ac.kr)
**(주)위니텍 (goldie64@naver.com)
***(주)더아이엠씨 (kjson@theimc.co.kr)
접수일자 : 2015년 5월 6일, 수정완료일자 : 2015년 5월 26일, 최종게재확정일자 : 2015년 5월 28일

분석을 통한 예측 모델링으로 사기 패턴과 유사한 행동을 검출하고, 소셜 네트워크 분석을 통해 계좌, 전화번호 등의 연관 관계 분석을 실시하여 범죄 네트워크 발굴 및 감시 시스템을 마련하였다. 또한, 2013년 미국 보스톤 마라톤 테러 사건에서도 미국은 범인 검거를 위해 현장에 있는 CCTV와 대회 참가자들이 찍은 영상과 같은 비정형 데이터를 수집하고 방대한 양의 빅 데이터 분석을 통해 테러 용의자를 파악하고 검거하였다[3-4]. 위와 같은 사례 들을 통해 빅 데이터 기술을 범인 검거 및 나아가 범죄 예측까지 활용할 수 있다는 것을 알 수 있다. 범죄 예측의 경우 현재 해외에서는 도입이 적극적으로 시도 되고 있으며, 국내 또한 사회 안전망 구축을 위한 빅 데이터 활용 기술이 끊임없이 연구 중에 있다.

이에 따라 본 논문에서는 관련된 선행연구를 바탕으로 대검찰청에서 수집하여 공공데이터 포털 사이트에 공개한 범죄 관련 빅 데이터를 이용하여 데이터 간의 유사 특성을 찾고, 베이지안 네트워크의 추론 특성을 적용하여 이를 어느 지역에서 해당 범죄가 일어났는지 조기에 파악하고, 해당 지역의 범죄 예측에 활용할 수 있는 범죄예측 알고리즘을 구현했다. 2장에서는 본 논문에서 제안하는 알고리즘에서 범죄발생 요인 분서을 위해 사용한 범죄 관련 빅 데이터에 대해 정의하고, 3장에서는 범죄발생 요인 분석 기반 범죄예측 알고리즘에 사용된 베이지안 네트워크를 다루며, 4장에서는 제안한 알고리즘 내용, 5장에서는 구현 결과를 보이고 6장의 결론으로 끝을 맺는다.

Ⅱ. 범죄 관련 빅 데이터의 정의

본 논문은 대검찰청에서 2011년부터 2013년까지 3년간 수집한 범죄 관련 빅 데이터를 이용하여 범죄발생 요인을 분석했다. 대검찰청에서 제공한 범죄 관련 빅 데이터에서 분석할 수 있는 범죄발생요인으로는 지역, 시간, 요일, 장소이며, 본 논문에서는 서울특별시를 분석 지역으로 선정해 범죄발생 요인을 분석했다. 전체 데이터 구성은 표 1과 같다.

서울특별시의 범죄발생지역 빅 데이터는 각 구별로 절도, 살인, 강도, 방화, 강간, 폭행, 상해, 공갈, 약취와 유인, 체포와 감금, 폭력행위 등 처벌에 관한 법률 위반, 도박과 복표, 과실치사상, 업무상 과실치사상 등 위 상기 범죄 유형에 따라 빈도수로 분류되어 있다. 범죄발생지역에서는 빈도수가 2012년이 171,796건으로 가장 높았다. 표 2는 5대 범죄(살인, 강도, 강간, 절도, 폭행) 발생 빈도수가 높은 5개의 구의 데이터를 나타낸다.

제공된 범죄발생시간 데이터는 범죄별 발생 빈도수를 새벽, 아침, 오전, 오후, 저녁, 밤의 시간대 별로 분류하였으며, 5대 범죄발생 빈도수는 표 2와 같다.

표 1. 범죄 관련 빅 데이터의 전체 구성.

데이터 제공	대검찰청
기간	2011 ~ 2013
수집지역	서울특별시

(단위 : 건)

	2011	2012	2013
범죄발생지역	161,097	171,796	169,473
범죄발생시간	811,116	851,722	842,504
범죄발생요일	811,116	851,722	842,504
범죄발생장소	811,116	851,722	842,504

표 2. 5대 범죄 발생 빈도수. **(단위 : 건)**

	살인	강도	강간	절도	폭행
강남구	44	158	1,738	13,323	6,736
송파구	23	91	931	11,933	4,993
관악구	37	137	1,224	9,727	6,307
서대문구	36	83	1,134	10,019	4,661
영등포구	51	102	934	8,950	5,214

표 2에 나타난 범죄별 발생 빈도수를 분석한 결과, 절도 범죄의 빈도수가 가장 많았고, 토요일 밤 노상에서의 범죄 발생 빈도수가 가장 높게 나타났다.

Ⅲ. 베이지안 네트워크

베이지안 네트워크(Bayesian Network)는 방향성 비순환 그래프 (Directed Acyclic Graph, DAG)로 여러 변수들의 조건부 확률을 각 노드와 호를 이용하여 그래픽 기반 모형으로 표현하는 확률적 모델이다. 노드는 확률 변수를 나타내며, 호는 노드 간의 의존성을 나타낸다. 베이지안 네트워크에서의 의존성은, 각 노드 간 원인에서 결과로 이어지는 우연적 관계를 나타내며, 이렇게 표현된 네트워크에서의 각 노드는 베이즈 추론을 기반으로 각 변수의 의존 관계를 통해 조건부 확률표를 갖는다. 베이지안 네트워크는 부분적인 증거만으로도 다른 노드에 대한 추론이 가능하기 때문에 불확실한 지식을 추론하는데 사용 가능하다[5].

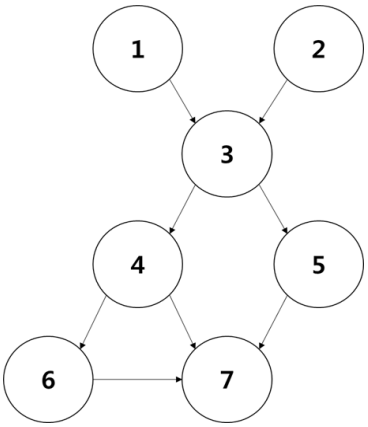


그림 1. 베이지안 네트워크 예시

한편, 이러한 방향성 그래프에서 상위 노드들은 하위 노드들이 발생 한 후 일어나게 되는 조건부 사건들로 정의할 수 있다. 조건부 확률은 사건 X가 발생했다는 가정 하에 사건 Y가 일어날 확률을 의미하며 이는 사건 X와 Y가 동시에 발생하였는데 X가 먼저 발생한 후, Y가 발생했다는 것을 의미한다. 이 때, 사건 X와 Y가 동시에 발생할 결합 확률 함수는 다음 식 1과 같이 표현 된다.

$$P(X, Y) = P(Y|X)P(X) \quad (1)$$

식 1을 사건 X가 발생한 후 Y가 발생할 조건부 확률을 구하는 것으로 정리하면 베이즈 규칙으로 잘 알려진 아래의 식 2가 된다.

$$P(Y|X) = \frac{P(X, Y)}{P(X)} . \quad (2)$$

방향성 그래프로 표현되는 베이지안 네트워크는 원인이 되는 부모노드와 결과노드인 자식 노드들에 대한 조건부 사건들로 정의되며, 그래프상의 모든 노드들에 대한 조건부 확률 관계를 연쇄법칙(Chain Rule)을 이용하고, 이를 일반화한 아래의 결합 확률밀도 함수로 표현 할 수 있다. 베이지안 네트워크를 구성하는 각 노드를 독립 변수로 표현하면 n개의 노드를 가진 베이지안 네트워크의 결합 확률 분포는 아래의 식 3과 같다.

$$P(x_1, x_2, \ldots, x_n) = \prod_{i}^{n} P(x_i | Parents(x_i)) , \quad (3)$$

여기서, n은 전체 노드의 수를 의미하며, i는 부모노드의 상태 수를 의미한다. 범죄 예측을 위한 베이지안 네트워크를 설계하기 위해서는 먼저 미래시간에 각 범죄에 영향을 미치는 것이 무엇인지 고려하여 범죄발생 요인 노드들을 설정해야 한다. 범죄발생 요인으로는 크게 공간적 특성(옥내 주차장, 풍속업소, 학교용지), 인구적 특성(인구밀도, 고령인구, 외국인), 사회적 특성(경찰력, 재산세) 및 시간, 요일, 기상변화요인 등이 있다[6].

Ⅳ. 베이지안 네트워크 적용 알고리즘

앞서 소개한 베이지안 네트워크를 실제 알고리즘에 적용하기 위해 그림 2를 통해 범죄발생 요인 기반 베이지안 네트워크를 보였다. 앞에서 다루었듯이, 범죄발생 요인은 크게 공간적 특성, 인구적 특성, 사회적 특성 및 시간, 요일, 날씨와 같은 기타 요인과 밀접한 관계를 지닌다. 이러한 요인들을 베이지안 네트워크의 원인 노드로 설정하였고, 그에 따른 결합 확률 분포는 식 3으로 표현 할 수 있다. 그림 3은 베이지안 네트워크 기반 범죄예측 알고리즘을 나타낸다.

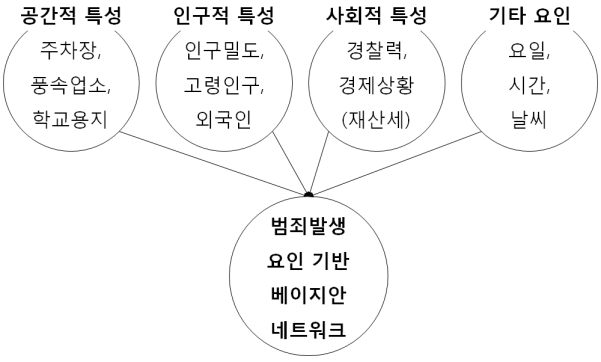


그림 2. 범죄발생 요인 기반 베이지안 네트워크.

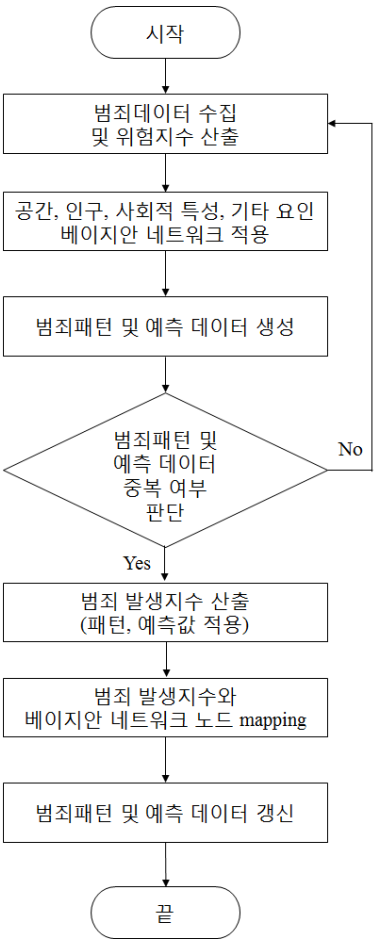


그림 3. 베이지안 네트워크 기반 범죄예측 알고리즘.

위 알고리즘은 크게 베이지안 네트워크, 범죄발생지수, 범죄패턴으로 설명된다. 시스템에 입력된 범죄데이터들의 분석을 통해 데이터를 트리 형태로 패턴을 형성하며, 형성된 패턴을 통해 데이터 간의 유사성을 비교하여, 범죄 예측 지표로 사용된다. 범죄데이터 간 패턴 생성은 본 논문의 선행연구로 본 연구실에서 진행한 참고문헌 [7]의 패턴 생성 알고리즘을 따른다. 한번 생성 된 패턴은 다시 생성되지 않으며, 다른 경우의 수로 패턴을 생성한다. 범죄 발생지수의 경우, 범죄 발생 빈도를 비교할 수 있는 척도로서 그 값이 클수록 범죄가 일어날 가능성이 크다는 것을 의미한다. 본 논문의 선행연구로 이뤄진 [7]의 논문에서는 범죄 발생지수가 1.8 이상이면 범죄발생빈도가 높은 것으로서 범죄가 일어날 가능성이 크고, 1.5 이하인 값들은 범죄 발생 가능성이 적다고 간주해 패턴에서 제외된다. 구해진 범죄 발생지수는 베이지안 네트워크에서 고려된 각 노드와 단계별로 지도에 원으로 맵핑된다. 이러한 베이지안 네트워크의 노드를 통해 범죄 패턴 및 예측 데이터를 시스템에 갱신하며, 예측 값을 향상시키기 위해 베이지안 네트워크의 노드는 범죄발생에 큰 영향을 미치는 요인들을 추가하게 된다.

Ⅴ. 알고리즘 구현 결과

제공된 범죄관련 빅 데이터 중 5대 범죄(절도, 살인, 강간, 폭행, 강도)를 공간통계분석을 통한 범죄발생 요인을 분석한 결과는 그림 4와 같다. 범죄 발생 중심점은 유동인구가 많은 강남구, 송파구, 종로구 중심으로 3개가 나타났으며, 각 타원체마다 방향성을 띄고 있었다. 총 범죄의 경우에는 강남구, 송파구, 서초구 일대가 가장 범죄 밀도가 높았고, 그 다음으로 영등포구, 동작구, 양천구 일대가 높게 나타났으며, 종로구, 성북구 순으로 범죄 밀도가 높았다. 그리고 폭행, 절도 범죄는 노상에서 많은 패턴 분포가 일어났고, 살인 범죄는 단독주택, 도박 범죄는 사무실에 특히 강하게 분포되는 것을 확인할 수 있었다.

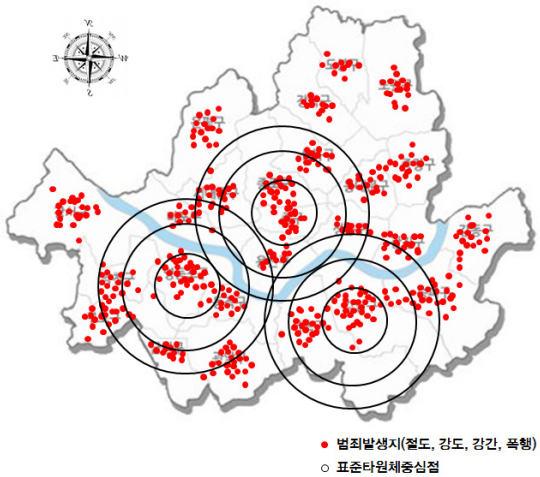


그림 4. 범죄발생지 및 타원체 중심.

표 3. 토요일 밤 노상에서 절도 범죄발생지수 및 예측 확률

	P_{RTDS} (기존)	P_{RTDS} (예측)
강남구	2.3	2.53
송파구	2.21	1.88
은평구	1.87	1.54
도봉구	1.72	1.61
금천구	1.7	1.45

표 4. 중랑구 범죄발생지수 및 예측 발생지수

범죄	발생요일	발생장소	P_{RTDS} (기존)	P_{RTDS} (예측)
절도	토요일	노상	1.82	1.98
살인	토요일	단독주택	2.46	2.34
폭행	토요일	노상	2.21	1.92
도박	금요일	사무실	2.18	1.66

표 5. 5대 범죄 발생 빈도수 및 예측 **(단위 : 건)**

	살인	강도	강간	절도	폭행
강남구	44	158	1,738	13,323	6,736
(예측)	39	172	1,694	13,127	6,724
송파구	23	91	931	11,933	4,993
(예측)	25	88	924	11,868	4,890
관악구	37	137	1,224	9,727	6,307
(예측)	32	133	1,227	9,730	6,228
서대문구	36	83	1,134	10,019	4,661
(예측)	30	79	1,138	9,987	4,357
영등포구	51	102	934	8,950	5,214
(예측)	55	121	886	8,769	5,198

표 3은 토요일 밤 노상에서 절도가 일어날 범죄발생지수를 5개의 지역구별로 나타낸 것이다. 기존 범죄발생지수는 본 논문의 선행연구 논문인 [7]에 제시된 알고리즘으로 산출하였으며, 예측 범죄발생지수는 본 논문에서 제안한 베이지안 네트워크 기반 범죄발생 요인 적용에 따른 알고리즘으로 산출하였다. 지역별 특징의 경우, 본 논문의 베이지안 알고리즘에서의 기타 요인 노드(요일, 시간, 날씨)는 예측값에 영향을 미치지 않았고, 공간적, 인구적 특성이 범죄발생지수와 밀접한 연관을 보였다. 특히, 인구적 특성의 경우 토요일 밤 많은 밀집 특성을 보이는 강남이 절도 범죄발생지수에서 기존 값대비 높은 예측 값을 보인다. 표 4는 중랑구에서 절도, 살인, 폭행, 도박 범죄의 범죄발생지수를 해당 범죄들이 일어날 수 있는 최대 확률을 고려하여 장소를 선택하고 산출하였다. 요일은 토요일로 동일한 조건하에 데이터를 얻었으며, 표 3과 마찬가지로 기존 논문에서 보인 범죄발생지수와 본 논문에서 제시한 알고리즘에 따른 예측 범죄발생지수로 비교하였다. 분석을 통한 결과, 기존 발생지수와 마찬가지로 중랑구는 토

요일 노상에서 절도와 폭행 범죄가 일어날 확률이 높고, 살인 범죄는 토요일 단독주택에서 발생지수가 높게 나타났다.

표 5는 5대 범죄 발생 빈도수가 높은 5개 구의 데이터를 나타낸다. 해당 자료는 2012년도 대검찰청 자료를 지역별 범죄 유형에 따라 표로 나타낸 것이며 예측 값은 2011년도 통계 자료를 본 논문에서 제안한 범죄 예측 알고리즘을 통해 산출하였다. 실제 데이터 값과 예측 알고리즘의 결과로 비교한 결과, 약 71%의 예측 정확도를 나타내었다.

Ⅵ. 결론

본 논문에서는 대검찰청에서 제공한 범죄관련 빅 데이터를 이용하여 범죄발생 요인을 분석했다. 수집한 범죄관련 데이터를 바탕으로 패턴을 형성하여 데이터 간의 유사성을 비교하였고, 이를 범죄 예측지표로 활용하였다. 또한, 범죄 예측 확률을 높이기 위해 부분적인 증거로도 추론이 가능하여 불확실한 지식 및 사건을 추론하는데 사용되는 베이지안 네트워크를 적용하였다. 베이지안 네트워크의 노드로는 공간적, 인구적, 사회적 특성과 시간, 요일, 날씨의 기타 요인을 적용하여 예측 알고리즘을 구성하였다. 제안된 알고리즘을 통한 분석으로 범죄 발생은 각 구별 인구 및 도시생활 등에 따라 다른 범죄 분포 패턴을 보이는 것을 알 수 있었고, 범죄 발생지수를 통해 수치화함으로써 그 값을 단계별로 나눌 수 있는 지표를 제시하였다. 이러한 범죄 예측과 더불어 범죄 데이터간의 다양한 패턴을 분석함으로서 여러 범죄 데이터 분석이 가능하다.

본 논문에서 제안한 알고리즘을 통해 해당 요일, 시간 및 지역에 따라 효율적인 범죄 예방 시스템을 갖춤으로써 범죄 발생률 감소 및 치안 유지에 필수적으로 활용될 수 있을 것으로 기대된다. 추후 연구로는 더욱 정확한 범죄 예측으로 활용되기 위해 각 구별 데이터가 아닌 각 구의 지역별 데이터를 통해 보다 상세하고 다양한 범죄관련 빅 데이터를 수집하고, GIS를 통한 핫스팟 분석이 필요하며, 더욱 다양한 범죄발생 요인 분석과 범죄를 다각도로 예측 할 수 있는 베이지안 네트워크 설계 연구가 필요할 것으로 판단된다.

참 고 문 헌

[1] S. Yin and O. Kaynak, “Big data for modern industry: Challenges and trends,”in Proc. of the IEEE. pp. 143-146, vol. 103, Feb. 2015.

[2] S. Rahim and T. Sun, "ICTs based crime control model: An application based study of Gilgit-Baltistan, Pakistan," in Proc. of 2011 10th International Conf. on Electronic Measurement & Instruments (ICEMI), pp. 1-6, Chengdu, China, Aug. 2011.

[3] A. C. Alegria, H. Sahli, and E. Zimanyi "Application of density analysis for landmine risk mapping," in Proc. of IEEE International Conf. on Spatial Data Mining and Geographical Knowledge Services (ICSDM), pp. 223-228, Fuzhou, China, June 2011.

[4] M. Saravanan and R. Thilagaraj, “Cyber crime spatial analysis,”Journal of Applied and Engineering Research of Integrated Publishing Association, vol. 23, no. 1, Aug. 2013.

[5] K. A. C. Baumgartner, S. Ferrari, and C. G. Salfati, “Bayesian network modeling of offender behavior for criminal profiling,”in Proc. of IEEE International Conf. on Decision and Control, Seville, Spain, 2005, pp. 2702-2709, Dec. 2005

[6] T. K. Lee, “A study on the causes of crime occurrence,” M.S. thesis, Dept. Administration, Kyunghee Univ., Seoul, Korea, 2010.

[7] K. H. Cha, K. H. Kim, S. K. Jun, S. J. Kim, D. C. Lee, and Jin Young Kim, “Analysis of Relation Between Criminal Types and Spatial Characteristics in Urban Areas,” Journal of Korea Society of Space Technology (J. KOSST), vol. 10, no. 1, pp. 6-11, Mar. 2015.

저자

박 지 호(Ji Ho Park) **학생회원**

・2014년 2월 : 광운대학교 전자융합공학과 졸업

・2014년 3월 ~ 현재 : 광운대학교 전파공학과 석사과정

<관심분야> : 위치공학, 데이터마이닝, 재난통신, 협력통신

차 경 현(Gyeong Hyeon Cha) **학생회원**

・2014년 7월 : 광운대학교 전자융합공학과 졸업

・2014년 8월 ~ 현재 : 광운대학교 전파공학과 석박사통합과정

<관심분야> : 데이터마이닝, 디지털통신, 5G 이동통신

김 경 호(Kyung Ho Kim) **학생회원**

・2013년 2월 : 광운대학교 전파공학과 졸업

・2013년 3월 ~ 현재 : 광운대학교 전파공학과 석박사통합과정

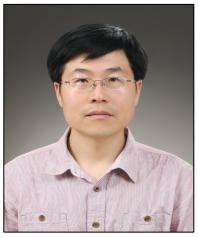
<관심분야> : 디지털통신, 스마트그리드, 데이터마이닝, 5G 이동통신

이 동 창(Dong Chang Lee)

・1990년 2월 : 경북대학교 전자공학과 학사
・2006년 11월 : 파마닉스/아트시스템 상무
・2013년 2월 ~ 현재 : 위니텍 해외사업본부 부장

<관심분야> : USN, Machine Vision, 영상처리, Multi-Vision System, 패턴 인식

손 기 준(Ki Jun Son)



・2005년 2월 : 경북대학교 컴퓨터공학과 공학박사 수료
・2011년 10월 : 에이투텍 플라톤 개발그룹 차장
・2013년 8월 ~ 현재 : 더아이엠씨 빅데이터팀 부장

<관심분야> : 자연어처리, 빅데이터 수집 및 분석

김 진 영(Jin Young Kim) **종신회원**

・1998년 2월 : 서울대학교 전자공학과 공학박사
・2001년 2월 : SK텔레콤 네트웍 연구소 책임연구원
・2001년 3월 ~ 현재 : 광운대학교 전자융합공학과 교수

<관심분야> : 디지털통신, 가시광통신, UWB, 부호화, 인지무선통신, 4G 이동통신