논문 2016-11-25

컨볼루션 특징 맵의 상관관계를 이용한 영상물체추적

(Visual object tracking using inter-frame correlation of convolutional feature maps)

김 민 지, 김 성 찬*
(Min-Ji Kim, Sungchan Kim)

Abstract : Visual object tracking is one of the key tasks in computer vision. Robust trackers should address challenging issues such as fast motion, deformation, occlusion and so on. In this paper, we therefore propose a visual object tracking method that exploits inter-frame correlations of convolutional feature maps in Convolutional Neural Net (ConvNet). The proposed method predicts the location of a target by considering inter-frame spatial correlation between target location proposals in the present frame and its location in the previous frame. The experimental results show that the proposed algorithm outperforms the state-of-the-art work especially in hard-to-track sequences.

Keywords : Computer vision, tracking, Convolutional neural net, Feature map correlation

Ⅰ. 서 론

영상물체추적은 컴퓨터 비전 분야의 주요 분야 중 하나로, 행동기반 인식, 무인감시, 자동차 내비게이션, 교통정보 모니터링과 같은 다양한 분야에서 사용될 수 있다. 최근 임베디드 시스템의 연산 능력 향상으로 로봇이나 무인 자동차 등의 분야에서도 영상물체추적의 도입이 활발하게 이루어지고 있다 [1]. 영상물체추적 알고리즘들의 성능은 매우 빠르게 개선되고 있으며, 최근에는 Convolution Neural Network(ConvNet)이 널리 사용되고 있다 [2, 3].

ConvNet은 기계학습에서 딥러닝 기반의 지도학습 기법으로 분류된다. ConvNet을 이용한 연산은 컨볼루션을 이용한 특징추출 계층들과 완전 연결층으로 이루어진 다단계 신경망을 이용한다. ConvNet의 연산 과정은 대략 다음과 같다. 먼저 특징추출 계층에서 입력으로부터 컨볼루션을 이용하여 특징을 추출한 후, 일정 크기의 윈도우를 대표하는 값을 산출한다. 특징추출과 대푯값 산출과정을 반복하면 전체 이미지를 대표하는 추상적인 특징들이 추출된다. 결과적으로 컨볼루션 특징추출 계층의 후반부에서는 입력 이미지에 있는 분류 (또는 인식) 가능한 물체의 대략적인 형태가 이전 계층들에서 추출된 특징들의 조합으로 보이게 된다. 그 후 이를 완전 연결층의 입력으로 주어 해결하고자 하는 문제에 따라 분류나 인식을 수행한다 [4].

ConvNet을 이용한 물체인식 기법들은 컨볼루션을 이용한 특징추출 계층들을 거쳐 마지막 혹은 그 이전 계층에서의 특징 맵을 이용하여 물체의 위치를 예측한다. 하지만 이 경우 후반부의 컨볼루션 특징 맵은 이전 컨볼루션 계층들에서 반복되는 서브샘플링을 통해 저해상도의 매우 추상적인 특징만을 담고 있다. 따라서 물체의 공간정보가 상당부분 소실되어 정확한 위치 예측에 한계가 있다. 그 결과 ConvNet을 이용한 기존 연구들은 물체가 가려지고, 모양이 변하거나 빠르게 움직이는 물체에 대해서는 추적 성능의 한계를 보인다 [2].

본 논문에서는 ConvNet으로부터 추출된 특징 맵 중 컨볼루션 단계의 전반부에서 생성되는 입력 영상의 자세한 특징과 공간정보를 포함한 저수준 특징 맵을 이용한 추적 물체의 새로운 위치를 예측하는 기법을 제안한다. 제안하는 기법은 인접한 영상 프레임들 사이에서 추적 물체의 저수준 특징들이 공간적 위치에 대한 연관성을 갖는다는 점을 착안하여, ConvNet의 첫 번째 컨볼루션 단계를 거치고

*Corresponding Author (s.kim@chonbuk.ac.kr)
Received: 29 June 2016, Revised: 12 July 2016, Accepted: 21 July 2016.
S. Kim, M.J. Kim: Chonbuk National University
ⓒ IEMEK J. Embed. Sys. Appl. 2016 Aug. 11(4) 219-225
ISSN : 1975-5066
http://dx.doi.org/10.14372/IEMEK.2016.11.4.219

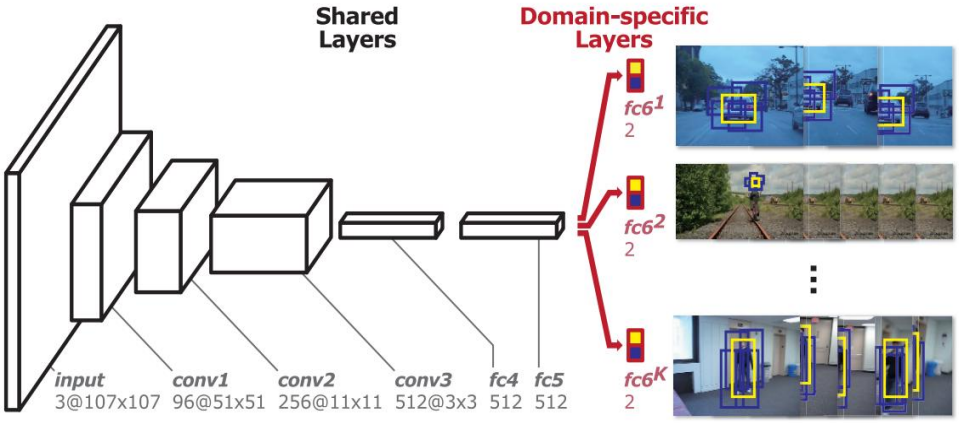


그림 1. MDNet에서 물체추적에서 사용되는 ConvNet의 구조[2]
Fig. 1 Structure of ConvNet for object tracking in MDNet [2]

난 저수준 특징 맵들 사이의 상관관계를 추적 후보 위치들의 ConvNet을 통한 적합도 산출에 반영해 추적 정확도를 향상시킨다.

본 논문의 구성은 다음과 같다. 2장에서는 제안하는 기법의 알고리즘을 설명한다. 3장에서는 실험을 통해 제안하는 기법의 성능을 입증하며 마지막으로 4장에서 결론을 맺는다.

II. 제안하는 기법

1. ConvNet을 이용한 기존의 추적 기법

먼저 ConvNet을 이용한 물체추적 기법의 기본 절차에 대해 설명한다. 본 논문에서 물체추적기법의 ConvNet 구조는 MDNet에 기반을 둔다 [2]. MDNet은 현재까지 가장 우수한 성능을 가지는 ConvNet기반의 단일 물체 추적 알고리즘 중의 하나로 알려져 있다. 그림 1은 MDNet 논문에서 사용된 ConvNet의 구조를 나타낸다. MDNet은 완전연결층 마지막에 입력된 영상에 특화된 층을 ConvNet에 추가한 구조를 가지고 있다. MDNet에서는 이런 특징을 가진 ConvNet을 ImageNet [5]을 이용하여 미리 학습시킨다.

영상물체추적이 시작되면, 각 위치후보들에 대해 컨볼루션 단계와 표본추출 단계를 거쳐 512개의 3×3 특징 맵이 추출된다. 추출된 특징 맵들은 완전연결층으로 입력되고, 물체의 위치가 될 수 있는 확률이 계산된다. 물체의 새로운 위치는 완전연결층에서 구한 확률이 가장 높은 5개의 위치후보들의 평균으로 계산한다. 또한, 변화하는 물체의 특징을 반영하기 위하여 주기적으로 또는 추적이 실패하는 경우 신경망의 완전연결층을 업데이트한다.

추적 물체를 나타내는 바운딩박스의 위치 및 크기는 첫 번째 프레임의 바운딩 박스 영역의 마지막, 즉 세 번째 컨볼루션 특징 맵(Conv3)들의 활성화 값들에 의한 회귀 모델을 이용해 계산한다. 두 번째 프레임부터는 확률이 높은 5개의 위치후보들의 확률의 평균이 50%보다 클 경우, 추적이 잘 되었다고 판단하고 새로운 위치의 크기를 5개의 위치후보들을 이용하여 조절한다.

2. 첫 번째 계층의 컨볼루션 특징 맵 선택

ConvNet을 이용한 이전의 연구에서는 마지막 컨볼루션 단계의 출력인 3×3크기의 특징 맵을 사용한다. 이 경우 특징 맵은 매우 추상화된 대표 특징만을 담고 있기 때문에, 영상에서의 물체 위치정보가 소실되는 문제가 있다. 반면 컨볼루션 초기 단계에서 생성되는 특징 맵은 영상의 저수준 특징들을 담고 있기 때문에, 물체의 위치 정보의 소실이 적다.

본 논문에서는 첫 번째 컨볼루션 단계의 결과인 51×51 크기를 가진 96장의 특징 맵에 포함된 추적 물체의 저수준 특징들을 이용하여 물체의 위치에 대한 프레임간 상관관계를 고려한다. 제안하는 기법은 MDNet의 추출된 특징을 이용하여 위치 후보들의 점수를 계산하는 알고리즘을 개선하였으며, 특징을 추출할 특징 맵을 선택하는 알고리즘을 추가하였다.

Conv1 Feature map selection algorithm

1 **For each** feature map_i **in** feature map set
2 Calculate sum of all pixel values in feature map_i
sum_feat[i] ← sum(|all feature pixels in feature map_i|)
3 **end for**
4 sort(sum_feat)
5 select 5 most activated feature maps from sum_feat

그림 2. 특징 맵 선택 알고리즘
Fig. 2 Feature map selection algorithm

특징을 추출할 특징 맵을 선택하는 알고리즘은 계산 량을 줄이기 위해 사용된다. 96장의 특징 맵 중 추적하고자 하는 물체의 특징을 가장 잘 나타내는 특징 맵이 존재한다. 그 특징 맵은 다른 특징 맵에 비하여 활성화가 많이 되어 있기 때문에 모든 특징 맵을 살펴보지 않아도 활성도가 높은 특징 맵을 살펴본다면 계산 량을 줄이며 물체추적이 가능하다. 특징 맵의 활성도는 각 픽셀 값의 활성도를 이용하여 구할 수 있는데, 각 픽셀이 0이 아닌 값을 가지고 있다면 그 픽셀은 활성화 되었다고 생각한다. 각 픽셀은 음수 값을 가질 수도 있고, 양수 값을 가질 수도 있기 때문에, 정확한 활성도를 계산하기 위하여 각 픽셀의 절대값의 합을 이용하여 특징 맵의 활성도를 계산한다.

활성화된 특징 맵을 선택하는 알고리즘은 다음과 같다. 첫 번째 컨볼루션 단계(Conv1)를 거친 모든 특징 맵들의 집합 $\{feature\ map_i\}$에서, 각 특징 맵 $feature\ map_i$ 의 모든 픽셀의 절대값의 합을 sum_feat 배열에 저장한다. 이후, sum_feat을 정렬한 후, 가장 높은 값을 가지는, 즉 가장 많이 활성화된 5개의 특징 맵을 선택한다.

추적 물체에 대해 가장 많이 활성화된 특징 맵들은 상관관계를 구할 때 사용된다.

3. 물체 위치 추정

그림 2의 알고리즘에서 구한 5개의 특징 맵을 이용하여 위치후보들의 적합도를 계산하는 방법을 설명한다. 기본 원리는 완전연결층의 마지막에 상관관계를 구하여 점수에 더하는 층을 추가하여 완전연결층에서 앞선 단계의 결과 값을 이용하여 어떤 후보가 물체의 위치가 될 확률이 높은지 계산하도록 하는 것이다.

상관관계를 구하는 알고리즘을 그림 3에 제시하였다. $conv1_{prev}$은 이전 프레임에서 물체의 위치에

Correlation algorithm

1 **For each** $candidate_n$ **in** candidate set
2 **For each** feature map_i **in** feature map set
3 $corr_n \leftarrow corr_n + calc\_corr(conv1_{prev}, conv1_{cur})$
4 $corr_{gt_n} \leftarrow corr_{gt_n} + calc\_corr(conv1_{gt}, conv1_{cur})$
5 **end for**
6 $score_{cur} = score_{fc} \times (1 + 0.7 \times corr_n + 0.3 \times corr\_gt_n)$
7 **end for**
8 $locate_{cur} \leftarrow$ mean location of top 5 candidates in $score_{cur}$

그림 3. Correlation을 이용한 현재 프레임에서의 위치 예측
Fig. 3 Location prediction algorithm using correlation in the current image frame

대하여 그림 2에서 구한 상위 5개의 첫 번째 컨볼루션 단계(Conv1)를 거친 특징 맵이고, $conv1_{cur}$은 현재 물체위치 후보들 각각의 상위 5개 특징 맵들을 나타낸다. $conv1_{gt}$은 첫 번째 프레임에서 주어진 물체위치의 특징 맵들이다. $score_{fc}$은 완전연결층을 거쳐 계산된 후보들의 점수이고, $calc_corr()$는 특징 맵들의 상관관계를 구하는 함수이며 아래와 같이 정의된다.

$$calc\_corr(A,B) = \frac{\sum_m \sum_n (A_{mn} - \overline{A})(B_{mn} - \overline{B})}{\sqrt{(\sum_m \sum_n (A_{mn} - \overline{A})^2)(\sum_m \sum_n (B_{mn} - \overline{B})^2)}} \quad (1)$$

위의 수식에서 A와 B는 컨볼루션 단계를 거친 특징 맵이며 두 특징 맵의 크기는 $m \times n$으로 같다. $\overline{A}$와 $\overline{B}$는 각 특징 맵 속 픽셀의 평균값을 나타낸다. 위의 식을 이용하여 상관관계를 구하면 각 상관관계는 -1에서 1까지의 값을 가진다. 여기에서 상관관계가 1에 가까우면 양의 상관관계를 가지고 있으며, -1에 가까워지면 음의 상관관계를 가지고 있다. 만약 상관관계가 0이라면 두 특징 맵은 상관관계가 없다. 본 논문에서 제안하는 알고리즘에서는 활성화된 정도를 기준으로 특징 맵 사이의 상관관계를 계산하기 때문에 양의 상관관계를 가지며, 1에 가까운 값을 가지고 있을수록 상관관계가 높다고 판단한다.

상관관계를 이용하여 새로운 프레임에서 물체의 위치를 추측하는 알고리즘은 다음과 같다. 우선 각 후보 위치 $candidate_n$의 적합도 $score_{cur}$를 구하기

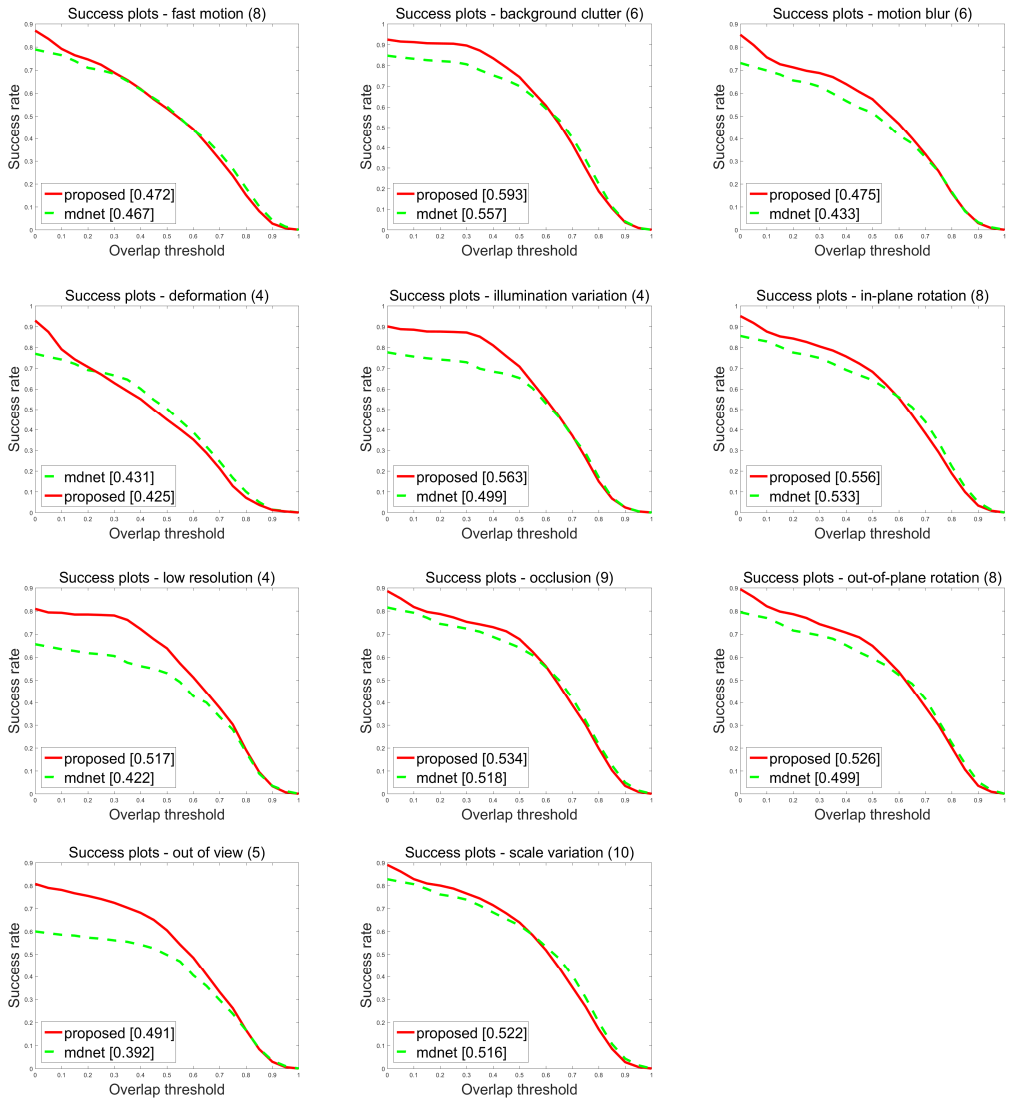


그림 4. 영상물체추적 성능이 저하되는 상황에서 제안하는 기법과 MDNet의 영상 특징별 성능비교
Fig. 4 Comparison of the proposed method to MDNet for hard-to-track sequences with respect to various sequence characteristics

위해서 이전 프레임에서 물체의 특징과 현재 후보 위치들의 특징 사이의 상관관계와 첫 번째 프레임에서 추적하고자 하는 물체의 특징과의 상관관계를 계산한다. 이때, 상위 5개의 특징 맵 각각에 대하여 상관관계를 구하여 더한 값이 $candidate_n$의 적합도가 된다. 영상에서 물체의 위치는 이전 프레임에서의 위치와 연관이 있고 추적하고자 하는 물체는 첫 번째 프레임에서의 물체와 동일한 물체이므로 특징이 지속된다고 가정한다. 위와 같은 과정을 거쳐, 현재 프레임에서 예측한 물체의 위치 $locate_{cur}$는 가장 높은 적합도를 가진 5개 위치후보들의 평균을 이용하여 계산한다.

Ⅲ. 실 험

제안하는 기법은 MATLAB의 MatConvNet toolbox를 이용해 구현하였다 [6]. 실험은 2개의 2.6GHz 6-코어 Xeon과 128 GB 메인메모리로 구성된 시스템에서 수행되는 리눅스 환경에서 이루어졌다. Object Tracking Benchmark(OTB)의 OTB50의 49개 영상을 사용하였다 [7].

식 (1)의 상수들은 다음과 같이 설정했다. 물체의 특징은 프레임이 지나감에 따라 변할 수 있지만, 이전프레임에서의 물체의 위치는 급격하게 변하지 않기 때문에, $corr_n$과 $corr_{gt_n}$의 비율을 70%와 30%로 설정하였다. 그리고 추적하는 물체의 다양한 크기 변화에 대응할 수 있도록 $locate_{cur}$의 중심좌표를 이전 프레임에서 물체의 위치를 평균, 분산 값을 10으로 설정한 정규분포를 사용하여 결정하고, 가로와 세로의 크기는 이전 프레임에서 추적한 물체의 크기의 ±10% 사이의 범위에서 결정하였다.

그래프에서 가로축의 overlap threshold는 ground truth와 예측 영역이 겹치는 넓이를 두 넓이를 합한 전체 넓이로 나눈 정도를 나타내며, 세로축은 overlap threshold 이상인 프레임의 비율을 나타낸다. 물체추적에 실패할 경우가 존재하기 때문에, Overlap threshold가 0일 때 Success rate이 1이 되지 못한다. Mdnet과 proposed옆의 숫자는 전체 성능의 평균값을 나타낸다.

그림 4는 추적 난이도가 높은 12개의 영상 (Biker, Bird1, Box, Car Scale, Clifbar, Couple, Football, Freeman4, Human4, Ironman, Jump, Motor Rolling)을 영상의 특징으로 분류하여 각 특징에 따라 기존의 기법과 제안하는 기법의 성능을 비교한 것이다.

각 특징으로는 background clutter, fast motion, out of view, illumination variation, motion blur, in-plane rotation, low resolution, occlusion, out-of-plane rotation, scales variation, deformation이 있다. 영상의 특징은 다음 표 1과 같다 [7].

영상의 특징 중, Deformation의 경우에 제안하는 기법의 성능이 기존의 기법보다 떨어지는 것을 볼 수 있다. 그 이유는 다음과 같다. 제안하는 기법에서는 이전 프레임에서의 영상의 특징을 초기의 물체의 특징보다 가중치를 주어 이용하기 때문에 인접한 프레임 사이에서 급격하게 특징이 변할 경우 이전 프레임의 특징과 달라지므로, 물체 추적에 있어 성능이 저하될 수 있다.

표 1. 영상 특징 리스트

Table 1. List of the attributes of sequence

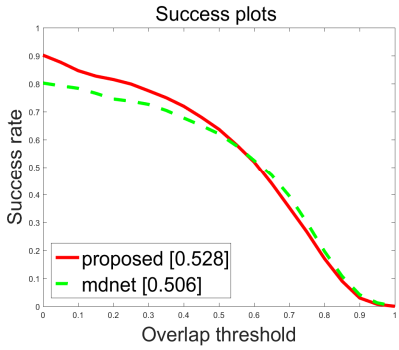
Feature	Description
Background clutter	The background near the target has the similar color or texture as the target.
Fast motion	The motion of the ground truth larger than 20 pixels.
Out of view	Some portion of the target leaves the view
Illumination variation	The illumination in the target region is significantly changed.
Motion blur	The target region is blurred due to the motion of target or camera.
Scale variation	The scale of the ground truth is changed.
Low resolution	The number of pixels inside the ground-truth is less than 400.
Occlusion	The target is partially or fully occluded.
Out-of-plane rotation	The target rotates out of the image plane.
In-plane rotation	The target rotates in the image plane.
deformation	Non-rigid object deformation.

그림 5의 (a)에서 Overlap threshold가 0.6 이상 0.9 이하일 때 성능이 떨어지는 것을 보이지만 그 차이는 미미하다. 반면 0.6 이하일 때는 두드러진 성능개선으로 전체적으로 MDNet보다 높은 추적 성능을 보인다. 그러나 그림 5(b)에서와 같이 49개 전체영상의 평균에서는 MDNet보다 성능이 떨어진다. 이는 성능을 비교하는 방법이 overlap threshold를 이용하여 평가하는 것에서 기인한다.

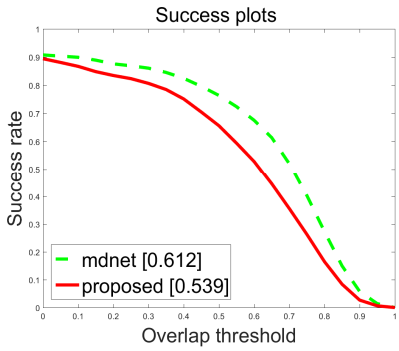
MDNet은 바운딩 박스의 크기가 거의 변하지 않는데 비해, 제안하는 방법에서는 바운딩 박스 크기가 가변적이다. 이로 인해 추적이 정확해도 바운딩 박스의 크기가 작을 경우에는 추적 성능은 낮게 평가된다. 이에 대한 개선은 향후 연구에서 이루어질 것이다.

Ⅳ. 결 론

본 논문에서는 기존 영상물체추적 알고리즘의 성능이 여전히 물체가 다른 물체에 가려지거나 모양이 변하고 빠르게 움직이는 경우 저하되는 문제점



(a)



(b)

그림 5. 제안기법과 MDNet 결과비교
(a) 추적 난이도가 높은 영상, (b) OTB50 전체 영상
Fig 5. Comparisons of the proposed method with MDNet for
(a) hard-to-track sequences and
(b) the entire sequences of the OTB50 benchmarks

들을 개선하기 위하여 추출된 저수준 특징을 이용한 추적하는 물체의 새로운 위치결정을 위한 기법을 소개하였다. 물체의 저수준 특징은 인접한 영상 프레임들 사이에서 공간적 위치에 대한 연관성을 갖는다는 점을 이용하여 제안하는 기법은 ConvNet의 완전연결층 마지막에 이전 프레임에서의 물체위치와 위치후보들의 상관관계를 구하는 층을 추가하여 위치후보의 점수를 계산한다. 이때, 여러 개의 저수준 특징 맵 중 계산 량을 줄이기 위해 추적 물체에 대해 가장 많이 활성화된 5개의 특징 맵을 이용한다. 실험을 통해 제안하는 기법이 기존연구가 추적의 난이도가 높은 영상에서 더 좋은 추적성능을 보이는 것을 입증하였다.

References

[1] A. Yilmaz, O. Javed, M. Shah, “Object Tracking: A Survey,” Preceeding of ACM Computing Surveys, Vol. 38, No. 4, pp. 1-45, 2006.

[1] H. Nam, B. Han, “Learning multi-domain convolutional neural networks for visual tracking,” CVPR, 2016.

[2] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, L. Cehovin, G. Fernandez, T. Vojir, G. Hager, G. Nebehay, R. Pflugfelder, “The Visual Object Tracking VOT2015 Challenge Results,” Preceeding of International Conference on Computer Vision Workshops, pp. 1-23, 2015.

[3] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, “Gradient-Based Learning Applied to Document Recognition,” Proceedings of the IEEE, Vol. 86, No. 11, pp. 2278-2324, 1998.

[4] http://www.image-net.org

[5] A. Vedaldi, K. Lenc. “Matconvnet-convolutional neural networks for matlab,” Proceedings of ACM international conference on Multimedia, pp. 689-692, 2015.

[6] Y. Wu, J. Lim, M.H. Yang, “Online object tracking: A benchmark,” Preceeding of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2411-2418, 2013.

Min-Ji Kim (김 민 지)

Minji Kim received the B.S. degree in computer science engineering from Chonbuk National University, Korea in 2015. She is currently a M.S. student in Chonbuk National University. Her research interests include embedded systems, machine learning and computer vision.

Email: kmkmj927@chonbuk.ac.kr

Sungchan Kim (김 성 찬)

Sungchan Kim received the B.S. degree in material science and engineering, the M.S. degree in computer engineering, and the Ph.D. degree in electrical engineering and computer science from Seoul National University, Korea, in 1998, 2000, and 2005, respectively. He is currently an Associate professor at Chonbuk National University, Korea. His research interests include embedded systems, cyber-physical systems, computer vision and machine learning.

Email: s.kim@chonbuk.ac.kr