ISA기반 시·공간적 특징 학습을 통한 사람의 요리 동작 인식

이기석[1O], 김은솔[2], Karinne Ramirez Amaro[3], Michael Beetz[3], 장병탁[2]
[1]{kslee, eskim, btzhang}@bi.snu.ac.kr, [2]{ramirezk, beetz}@in.tum.de
[1]서울대학교 컴퓨터공학부, [2]뮌헨공과대학교 정보과학대학

Human cooking action recognition via spatio-temporal feature learning based on ISA

Kisuk Lee[1O], Eun-Sol Kim[1], Karinne Ramirez Amaro[2], Michael Beetz[2], Byoung-Tak Zhang[1]
[1]School of Computer Science and Engineering, Seoul National University
[2]Fakultät für Informatik, Technische Universität München

요 약

기계학습(machine learning) 기술을 이용해서 영상 데이터로부터 동작 패턴을 인식하는 연구에 있어서, 최근 들어 무감독학습(unsupervised learning)의 중요성이 부각되고 있다. 본 논문에서는 ISA 알고리즘에 기반한 최신 무감독학습 기법인 'Stacked Convolutional ISA' 알고리즘을 이용해서 샌드위치를 만드는 인간의 동작을 촬영한 영상 데이터를 분석, 동작 인식을 행하였다. 데이터로부터 직접 유용한 특징들을 학습하는 무감독학습 기법의 장점을 그대로 나타내어, 해당 알고리즘은 제한적인 학습 및 테스트 샘플 조건 하에서도 인상적인 성능을 나타냈다. 반면 요리동작에 있어서는 손동작 자체를 인식하는 것 이외에도 현재 손에 쥐어진 도구나 재료의 종류를 인식하는 것이 중요한데, 이러한 문맥 인식(context recognition)은 향후 추가적으로 연구해야 할 과제로 남아있다.

1. 서론

최근 들어 기계학습(machine learning) 기술을 이용해서 영상 데이터를 분석하여 각종 동작 패턴을 인식하고자 하는 연구가 활발히 진행되고 있다. 여기에서 핵심은, 영상 데이터로부터 유용한 시·공간적 특징(spatio-temporal features)들을 추출하여, 이러한 특징들의 분포 차이를 이용해서 각각의 동작 패턴들을 분류하는 것이다.

기존에는 영상 데이터로부터 유용한 시·공간적 특징들을 추출할 때 연구자가 직접 설계한 특징(hand-crafted features)들을 이용했다. 예를 들어, SIFT[1]는 이동, 확대·축소, 회전과 같은 국소적 공간 변형(local transformation)에 불변(invariant)하도록 설계된 특징들을 이용해서 장면 상의 물체를 인식한다.

하지만 이러한 기법들은 추출할 특징들을 연구자가 직접 정교하게 설계해야 하기 때문에 상대적으로 많은 시간과 노력이 필요하다. 또한 미리 설계한 고정된 특징 집합만을 이용하기 때문에 다양한 데이터에 유연하게 대처하기 어렵다. 최근에는 이러한 단점들을 보완하기 위해 무감독학습(unsupervised learning) 알고리즘을 이용해서 데이터로부터 직접 유용한 시·공간적 특징들을 학습하는 기법이 각광을 받고 있다.

본 논문에서는 최근에 발표된 [2]에서 제시한 무감독학습 알고리즘을 이용했는데, 이 알고리즘은 ISA (Independent Subspace Analysis) 알고리즘[3]을 확장시킨 것이다. 본 논문에서는 이러한 'Stacked Convolutional ISA' 알고리즘을 이용하여 사람이 요리하는 과정을 촬영한 영상 데이터로부터 직접 유용한 시·공간적 특징들을 학습하였고, 또한 [2]에서와 마찬가지로 이렇게 학습한 특징들을 최근의 동작 인식 분야에서 가장 널리 이용되고 있는 방법인 'bag-of-features SVM[4]' 기법에 접목시켜 동작 인식 및 분류를 행하였다.

2. 요리 동작 영상 데이터

본 논문에서 사용한 영상 데이터는 뮌헨공과대학(Technische Universität München, TUM)의 IAS(Intelligent Autonomous System) 그룹 Michael Beetz 교수 팀이 제작했다. 실험을 위하여 실제 사람이 빵, 오이, 치즈 등의 재료를 이용하여 샌드위치를 만드는 과정을 카메라를 이용해서 세 방향에서 촬영하였다 (그림 1). 동작 인식 및 분류를 위해서 사람이 샌드위치를 만드는 과정에서 특징적인 동작들을 총 9개의 범주로 분류하였고, 이를 정리한 것을 표 1에서 찾아볼 수 있다.

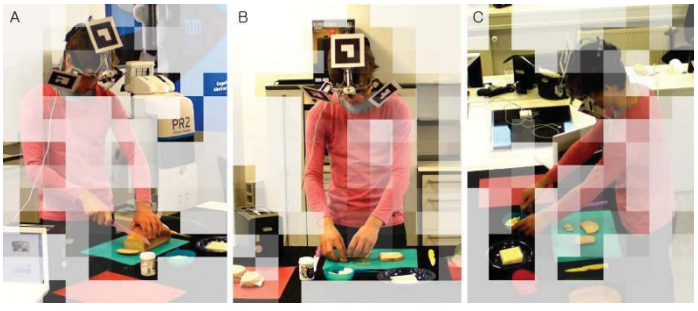

그림 1. 세 방향에서 촬영한 샌드위치 요리 과정. Norm-thresholding interest points detection 결과가 함께 표시되어 있다 (본문 참조).

3. ISA 기반 요리 동작 인식

3.1. Independent Subspace Analysis

ISA (Independent Subspace Analysis) 알고리즘[3]은 이미지 패치로부터 유용한 특징(features)들을 학습하는 무감독학습 알고리즘이다[2]. ISA 알고리즘은 구조적으로 ISA 네트워크라는 2계층 네트워크(two-layered network)로 나타낼 수 있다[5]. 이를 신경망(neural network) 구조로 나타낸 것이 그림 2이다.

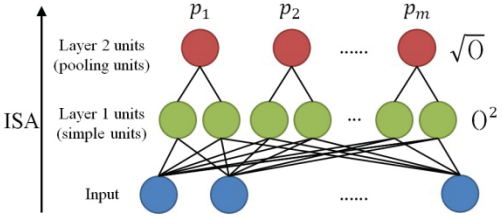

그림 2. ISA 네트워크의 신경망 구조 (adapted from Figure 1 in [2])

ISA 네트워크 첫 번째 계층의 구성 단위를 simple unit이라고 하며, 주어진 입력 패턴 x^t와 simple unit들은 학습 가능한 가중치 집합 $\boldsymbol{W}$로 연결되어 있다. ISA 네트워크 두 번째 계층의 구성 단위는 pooling unit이라고 하며, simple unit과 pooling unit은 일반적으로 미리 고정된 가중치 집합 $\boldsymbol{V}$로 연결되어 있다. 주어진 입력 패턴 x^t에 대해, pooling unit의 활성값(activation)은 다음과 같이 나타낼 수 있다.

$$p_i(x^t;\boldsymbol{W},\boldsymbol{V}) = \sqrt{\sum_{l=1}^{k}\boldsymbol{V}_{il}\left(\sum_{j=1}^{n}\boldsymbol{W}_{lj}x_j^t\right)^2} \qquad (1)$$

ISA 알고리즘은 ISA 네트워크 두 번째 계층의 sparse feature representation을 찾음으로써 첫 번째 계층의 학습 가능한 가중치 집합 W를 학습하는데, 이 때 다음 식을 이용한다.

$$\underset{\boldsymbol{W}}{\text{minimize}}\sum_{t=1}^{T}\sum_{i=1}^{m}p_i(x^t;\boldsymbol{W},\boldsymbol{V}), \qquad (2)$$
$$\text{subject to } \boldsymbol{W}\boldsymbol{W}^T = \boldsymbol{I}$$

여기에서 입력 패턴 $\{x^t\}_{t=1}^T$은 whitening 된 입력 예제들이다. n,k,m은 각각 입력 차원(input dimension), simple unit 개수, pooling unit 개수를 나타내며, 따라서 $\boldsymbol{W}\in\mathbb{R}^{k\times n}$, $\boldsymbol{V}\in\mathbb{R}^{m\times k}$ 이다. 식 (2)의 orthonormal constraint는 ISA 알고리즘에 의해 학습된 특징들의 다양성을 보장하는 조건으로서, 수학적으로 자세한 설명을 원할 경우 [5]을 참고하면 된다.

3.2. Stacked Convolutional ISA

앞에서 설명한 ISA 네트워크 구조는 작은 크기의 이미지 패치에 대해서는 실용적이지만, 입력 차원이 높아질수록 ISA 네트워크를 학습시키는데 소요되는 시간이 기하급수적으로 증가한다[2]. 따라서 ISA 알고리즘을 일반적인 크기의 이미지에 직접 적용시키는 것은 매우 비효율적이다.

이러한 문제점에 대한 돌파구를 [2]에서 심층 학습(deep learning) 기법을 통해 마련했다. 즉, 입력 데이터를 작은 차원으로 세분하여 ISA 네트워크를 적용시킨 뒤, 각각의 결과값을 취합하여(convolution) 이를 다시 새로운 ISA 네트워크의 입력 데이터로 사용하는 것이다. 이러한 방식을 반복하여(stacking) 계층적인 구조를 만들면 이른바 Stacked Convolutional ISA 네트워크가 만들어지고, 이를 통해 높은 차원의 입력 데이터를 효율적인 방식으로 다룰 수 있게 된다 (그림 3).

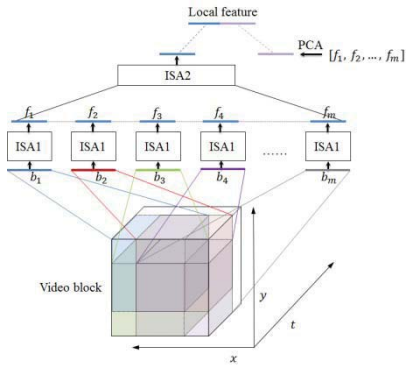

그림 3. Stacked Convolutional ISA 구조 (adapted from Figure 5 in [2])

Stacked Convolutional ISA 네트워크의 학습 과정은 심층 학습 분야의 각종 문헌[6]에서 제안한 greedy layer-wise training 기법을 사용한다[2].

4. 실험 및 결과

4.1. 시·공간적 특징 학습

본 논문에서 사용한 Stacked Convolutional ISA 네트워크의 파라미터는 [2]에서와 동일하게 설정했다. 우선 Stacked Convolutional ISA 네트워크는 두 계층으로 구성했다. 하위 계층에서 사용된 ISA 네트워크(**ISA1**)에 대한 입력 차원(또는 receptive field)은 $n = 16\times16\times10 = 2{,}560$ 으로 설정했고, $k = m = 300$ 으로 설정했다. 영상 데이터로부터 무작위로 $16\times16\times10$ 비디오 블록을 100,000 개 추출해서 **ISA1**을 학습시켰다.

표 1. 동작 범주 별 분류 정확도

(K는 K-fold cross-validation의 실행 회수를 의미)

	B 시점 (그림 1-B)			C 시점 (그림 1-C)		
Action	**K = 1**	**K = 2**	**K = 3**	**K = 1**	**K = 2**	**K = 3**
Crumpling	97.0% (33/34)	96.9% (32/33)	96.9% (32/33)	97.0% (33/34)	96.9% (32/33)	96.9% (32/33)
CuttingSomething	100% (34/34)	100% (33/33)	100% (33/33)	100% (34/34)	100% (33/33)	100% (33/33)
DisposeAnObject	100% (34/34)	100% (33/33)	96.9% (32/33)	100% (34/34)	100% (33/33)	100% (33/33)
Reaching	76.4% (26/34)	72.7% (24/33)	75.7% (25/33)	82.3% (28/34)	63.6% (21/33)	72.7% (24/33)
ReleaseGraspOfSomething	70.5% (24/34)	63.6% (21/33)	72.7% (24/33)	70.5% (24/34)	63.6% (21/33)	69.6% (23/33)
SpreadingOntoSurface	97.0% (33/34)	90.9% (30/33)	96.9% (32/33)	100% (34/34)	93.9% (31/33)	100% (33/33)
Sprinkle	100% (34/34)	100% (33/33)	100% (33/33)	100% (34/34)	100% (33/33)	100% (33/33)
TurningOnPowerDevice	97.0% (33/34)	96.9% (32/33)	96.9% (32/33)	97.0% (33/34)	96.9% (32/33)	96.9% (32/33)
UnWrappingSomething	100% (34/34)	100% (33/33)	100% (33/33)	100% (34/34)	100% (33/33)	100% (33/33)
Mean	93.1%	91.2%	92.9%	94.1%	90.5%	92.9%
K-folded Mean	92.4%			92.5%		

전체 Stacked Convolutional ISA 네트워크에 대한 입력 차원 n 은 $20 \times 20 \times 14 = 5{,}600$ 으로 설정했다. 상위 계층에 사용된 ISA 네트워크(**ISA2**)를 학습시키기 위해 역시 영상 데이터로부터 무작위로 $20 \times 20 \times 14$ 비디오 블록 100,000 개를 추출했다. **ISA2**에서는 $k = 200$, $m = 100$으로 설정했다.

최종적인 시·공간적 특징은 하위 계층의 중간 출력값 2,400 개를 PCA 차원 감소를 통해 100 개로 줄이고, 여기에 상위 계층 최종 출력값 100 개를 더해서 총 200 개의 값을 이용했다.

4.2. Norm-thresholding Interest Points Detection

[2]에서는 **ISA1**의 출력값의 총 합(activation norm)에 경계값(threshold value)을 적용해서 동작의 움직임이 통계적으로 유의미한 지점을 골라내는 이른바 'norm-thresholding interest points detecting' 기법을 선보였다.

본 실험에서 경계값을 30% 로 잡고 동일한 실험을 수행하여 움직임이 많은 지점을 시각화 한 것이 그림 1에 나타나 있다.

4.3. 동작 인식 및 분류

동작 인식 및 분류 역시 [2]에서와 동일한 'bag-of-features SVM' 기법[4]을 사용했다. 앞에서 학습한 Stacked Convolutional ISA 네트워크를 영상 데이터에 적용시켜서 국소적 특징(local features)를 계산한 뒤, 이를 K-means clustering 기법으로 vector quantization 시킨다. 총 9개의 동작 범주 각각에 대한 X^2 -kernel binary SVM (Support Vector Machine)을 학습시키고 동작 인식 및 분류를 행한다. A, B, C 세 개의 시점 중 학습 및 테스트 샘플이 준비된 B, C 시점에 대해서만 분류를 행하였으며, 각각은 3-fold cross-validation을 통해 신뢰성을 높였다 (표 1).

4.4. 결과 및 분석

표 2에 9개 동작 범주 각각에 대한 binary SVM 분류 결과가 나타나 있다. 표 1에서 확인할 수 있듯이 대부분의 범주에 대해서 높은 수준의 accuracy를 보였으며, 총 9개 범주의 성능 척도를 모두 평균해서 계산한 mean accuracy는 대략 90% 초반을 나타냈다.

수치상으로만 보면 평균적으로 90% 이상의 정확도를 보였기 때문에 사용한 알고리즘이 상당히 좋은 성능을 보인다고 생각할 수 있다. 하지만 동작 범주 별 샘플 분포가 고르지 못하고, 또한 상당수의 동작 범주에서 샘플의 절대적 개수가 모자랐다. 이러한 악조건으로 인해 multi-class SVM 분류를 시도하지 못했고, 오직 binary SVM 분류만 시행되었다. 또한 정확도가 높은 경우에도 학습 및 테스트 샘플 수가 적은 경우 그러한 높은 정확도가 주로 negative example에 의해 성취되었다.

이러한 불리한 조건에도 불구하고 몇몇 동작 범주는 학습 및 테스트 샘플 수도 어느 정도 갖추고 있고 분류 결과도 상당히 좋은 경우가 존재했다. 따라서 사용한 알고리즘의 성능을 제한적으로 확인해 볼 수 있었다.

5. 결론 및 향후 연구

본 논문에서는 [2]에서 제시한 핵심 알고리즘, 즉 Stacked Convolutional ISA를 이용한 시·공간적 특징의 무감독학습 기법을 인간의 요리 동작을 촬영한 영상 데이터에 적용시켜 보았다. 본 논문에서 사용한 알고리즘은 무감독학습 측면에서는 분류되어 있지 않은 데이터(unlabeled data)에 유용하게 적용될 수 있다는 장점이 있고, 또한 ISA 알고리즘의 측면에서는 생물학적으로 타당한(biologically plausible) 특징들을 학습할 수 있다는 장점이 있다[2,5,7]. 특히 요리동작은 다양한 요리법의 종류만큼이나 다양하기 때문에, 데이터로부터 직접 유용한 특징들을 학습하는 것이 유리할 수 있다.

본 연구에서는 동작 인식에만 초점을 맞추고 있는데, 순수한 동작 이외에 동작에 수반되는 다른 객체들을 함께 인식하는 등의 문맥 인식(context recognition) 역시 중요한 연구 대상이다. 예를 들어, 요리동작에 있어서는 손동작 인식 이외에도 현재 손에 쥐어진 도구와 재료를 인식하는 것이 중요하다. 이러한 문맥 인식에 있어서는 특히 주의집중(attention)이 필수적인 요소로 고려돼야 한다. 이에 관한 인지과학적 협력 연구가 앞으로 활발히 진행될 것으로 기대된다. 이러한 다학제적인(multi-disciplinary) 연구를 통해 인공지능은 점점 더 인간 수준의 지능(human-level intelligence)에 다가갈 수 있을 것이다.

참고문헌

[1] David G. Lowe, Object recognition from local scale-invariant features, In *ICCV*, 1999.

[2] Quoc V. Le, Will Y. Zou, Serena Y. Yeung, Andrew Y. Ng, Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis, In *CVPR*, 2011.

[3] Apo Hyvärinen, Patrik Hoyer, Emergence of phase- and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural Comput.*, 12(7):1705-1720, 2000.

[4] Heng Wang *et al.*, Evaluation of local spatio-temporal features for action recognition, In *BMVC 2009 – British Machine Vision Conference*, 2009.

[5] A. Hyvärinen, P. Hoyer, Natural Image Statistics, Springer, 2009.

[6] Yoshua Bengio, Learning deep architectures for AI, *Foundation and Trends in Machine Learning*, 2(1):1-127, 2009.

[7] David Heeger, Normalization of cell responses in cat striate cortex, *Visual Computation*, 6, pp.559-601, 1992.