(2012추계 우수발표논문)ISA 기반 시·공간적 학습을 통한 사람의 요리 동작 인식

(Human Cooking Action Recognition via Spatio-temporal Feature Learning based on ISA)

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' 알고리즘[2]을 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.

키워드 : 동작 인식, 무감독학습, Stacked Convolutional ISA

정보과학회논문지: 컴퓨팅의 실제 및 레터 제XX권 제X호(2013.XX)

Abstract In the research of action recognition from video data based on machine learning, unsupervised learning approach has recently been spotlighted. In this paper, we adopted 'stacked convolutional ISA' algorithm, a state-of-the-art unsupervised learning technique based on independent subspace analysis (ISA) algorithm that has recently been suggested in [2], to the human cooking action recognition from video data. The algorithm extracted useful spatio-temporal features directly from the video data, which can be regarded as the most significant advantage of unsupervised learning approach, resulting in impressive performance despite of the restricted number of training and test sets. In human cooking action recognition, it is imperative to recognize the identity of cooking utensils or food materials currently held in hands besides the hand action itself. This sort of context recognition remains open to the future study.

Key words : Action Recognition, Unsupervised Learning, Stacked Convolutional ISA

1. 서 론

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 (Scale-Invariant Feature Transform, [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 분석하고자 하는 데이터의 종류와 성격에 따라 효과적인 특징 집합이 천차만별로 달라지는데도 불구하고, 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.

본 논문은 총 6 절로 구성되어 있다. 2 절에서는 국내외 주요 동작인식 방법에 대하여 소개하고, 3 절에서는 본 논문에서 사용한 요리 동작 영상 데이터에 대하여 기술한다. 4 절에서는 핵심 알고리즘을 설명하고, 5 절에서는 제안된 방법에 대한 실험 결과를 기술하고 분석하여 타당성을 검증한다. 마지막으로 6 절에서는 결론을 맺고 향후 연구 과제에 대해서 검토한다.

2. 관련 연구

2.1 2차원 물체 인식 및 3차원 동작 인식

2 차원 이미지에서의 물체 인식(object recognition) 연구와 3 차원 영상에서의 동작 인식(action recognition) 연구는 서로 깊게 연관되어 있다. 2 차원 이미지에서의 물체 인식 분야에서는 연구자들이 직접 설계한 저수준 특징(low-level hand-crafted features)들이 최근까지도 활발하게 연구되고 있다. 이러한 설계된 특징들을 2 차원에서 3 차원으로 확장시키는 접근법이 영상 기반 동작 인식 연구의 주를 이루고 있다[2].

위에서 언급한 각종 설계된 특징들이 저수준 이미지 처리(low-level image processing) 분야의 이론적 지식을 활용했다면, 최근 들어서는 생물학 분야의 이론적 지식을 동원한 새로운 접근법들이 활발하게 제안되고 있다. 특히, 인간의 뇌에서 시·공간적 시각 정보를 처리하여 동작을 인식하는 과정에 대한 과학적 연구가 활발하게 진행되고 있으며[5], 이러한 신경과학적 지식을 바탕으로 한 새로운 동작 인식 시스템이 속속 등장하고 있다[6]. 하지만 이러한 시스템들 역시 뇌신경회로의 정보처리 과정을 모방해서 설계한 특징들을 이용하기 때문에, 서론에서 언급한 설계된 특징들의 단점을 그대로 나타낸다고 볼 수 있다.

2.2 심층학습 기법을 적용한 동작 인식

최근 들어 기계학습 분야에서 주목 받고 있는 심층학습(deep learning, [7],[8],[9]) 역시 신경과학적 원리에 그 바탕을 두고 있다고 할 수 있다. 이러한 심층 학습 알고리즘들 역시 동작 인식 연구 분야에서 활발하게 응용되고 있다. 이러한 심층 학습 기법에서는 심층 네트워크(deep networks)를 학습시키는 과정에서 자연스럽게 계층적이고 복합적인 특징들을 학습하게 되는 특성이 있다.

이들 중에서 특히 CNNs (Convolutional Neural Networks, [8])를 3 차원으로 확장시켜서 동작 인식에 적용시킨 연구가 주목할 만하다[10]. 하지만 [10]에서는 CNNs 의 가장 하위 단에 연구자들이 직접 설계한 특징들을 적용시켰기 때문에 앞에서 언급했던 설계된 특징들의 단점을 그대로 나타내고 있다. 또한 CNNs는 주로 감독학습(supervised learning) 기법으로 학습시키기 때문에, 다량의 분류된 데이터(labeled data)를 준비해야 한다는 단점이 있다. 분류된 데이터를 준비하기 위해선 사람이 직접 상당한 시간과 노력을 들여야 하기 때문에, 이는 감독학습 기법에 있어서 큰 단점으로 작용한다.

2.3 무감독학습 기법을 적용한 동작 인식

위에서 언급한 각종 동작 인식 기법들의 단점을 해결하기 위한 접근법으로서 무감독학습 기법이 최근 들어 주목 받고 있다. 무감독학습 기법에서는 실험 데이터로부터 유용한 시·공간적 특징들을 직접 학습하기 때문에 다량의 분류된 데이터를 준비할 필요가 없을뿐더러, 설계된 특징이 가지고 있는 단점들을 효과적으로 극복할 수 있는 것이다.

본 논문에서 사용한 핵심 알고리즘이 바로 이러한 무감독학습에 기반한 시·공간적 특징 학습 기법이다[2]. 특히 특징 학습을 위한 무감독학습 기법으로 택한 ISA [3] 알고리즘은 뇌신경회로에서 일어나는 정보처리 원리와도 깊은 관련이 있기 때문에, 앞서 언급했던 다양한 기법들의 장점을 취하면서 단점을 극복할 수 있는 효과적인 접근법이라 할 수 있겠다.

3. 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 9 개의 범주로 분류하였고, 이를 정리한 것을 표 1 에서 찾아볼 수 있다.

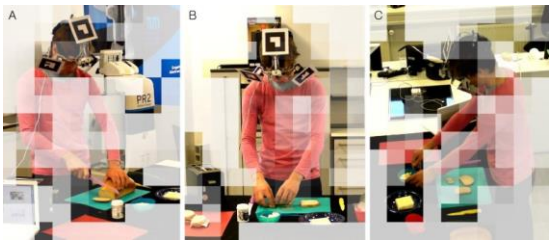

그림 1. 세 방향에서 촬영한 샌드위치 요리 과정. Norm-thresholding interest points detection 결과가 함께 표시되어 있다 (본문 참조).
Fig 1. Sandwich-making videos from three different angles. Norm-thresholding interest points detection results are also represented.

4. ISA 기반 요리 동작 인식

4.1 Independent Subspace Analysis

ISA 알고리즘[3]은 이미지 패치로부터 유용한 특징(features)들을 학습하는 무감독학습 알고리즘이다[2]. ISA 알고리즘은 구조적으로 ISA 네트워크라는 2-계층 네트워크(two-layered network)로 나타낼 수 있다[11]. 이를 신경망(neural network) 구조로 나타낸 것이 그림 2 이다.

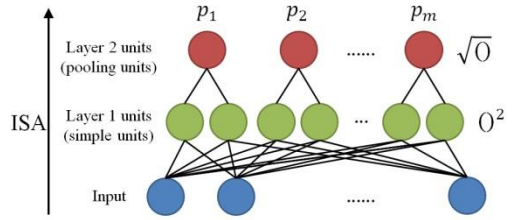

그림 2. ISA 네트워크의 신경망 구조 (참고문헌 [2]의 그림 1에서 차용함)
Fig 2. Architecture for ISA network (adapted from Figure 1 in [2])

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

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

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

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

여기에서 입력 패턴 $\{x^t\}_{t=1}^T$은 whitening 된 입력 예제들이다. n,k,m은 각각 입력 차원(input dimension), simple units 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 constraint 는 ISA 알고리즘에 의해 학습된 특징들의 다양성을 보장하는 조건으로서, 수학적으로 자세한 설명을 원할 경우 [11]을 참고하면 된다.

4.2 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 심층학습 기법[7]을 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).

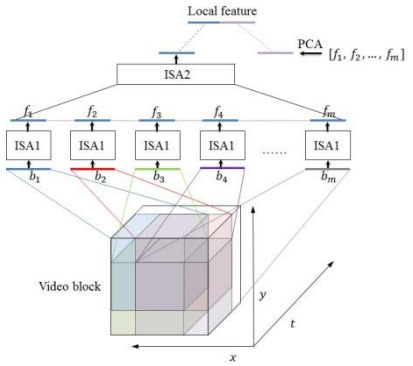

그림 3. Stacked Convolutional ISA 구조 (참고문헌 [2]의 그림 5에서 차용)
Fig 3. Architecture for stacked convolutional ISA (adapted from Figure 5 in [2])

[2]에서 제시한 방법을 간략히 설명하면 다음과 같다. 우선 ISA 네트워크를 작은 입력 데이터 패치들에 대해서 학습시킨다. 이렇게 학습된 ISA 네트워크를 좀 더 넓은 영역의 입력 데이터 패치에 분산 배치하고, 각각의 ISA 네트워크들의 결과값을 취합하여 사위 계층 ISA 네트워크의 입력값으로 사용한다. 그림 3 에서도 나타나 있듯이 하위 계층의 출력값을 상위 계층의 입력값으로 사용하는 과정에서 PCA (Principal Component Analysis) whitening 기법을 이용한 전처리 과정(preprocessing step)을 거친다. PCA whitening 기법은 이미지 처리 분야에서 표준적인 전처리 과정으로 사용되는데, 이에 관한 구체적인 설명은 [11]에 나와 있다.

Stacked Convolutional ISA 네트워크의 학습 과정은 심층 학습 분야의 각종 문헌[7,9]에서 제안한 greedy layer-wise training 기법을 사용한다[2]. 보다 구체적으로, 우선 첫 번째 계층의 ISA 네트워크를 수렴할 때까지 학습시킨 뒤, 이 네트워크를 하위계층에 분산 배치시켜서 Stacked Convolutional ISA 를 구성하고, 두 번째 계층의 ISA 네트워크를 마찬가지로 수렴할 때까지 학습시킨다. [2]에 의하면 이러한 기법을 통해 학습에 필요한 시간을 수 시간 정도로 대폭 감소시킬 수 있다.

5. 실험 결과 및 분석

5.1 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 16×16 픽셀 공간 차원과 10 프레임 시간 차원을 포괄해서 $n = 16 \times 16 \times 10 = 2{,}560$ 으로 설정했고, $k = m = 300$ 으로 설정했다. 영상 데이터로부터 무작위로 $16 \times 16 \times 10$ 비디오 블록을 100,000 개 추출해서 **ISA1** 을 학습시켰다 (그림 3).

앞서 언급했던 것처럼 2 계층으로 이루어진 Stacked Convolutional ISA 네트워크를 학습시키는 데에는 greedy layer-wise training [7,9] 기법을 사용했다. 이에 따라 먼저 하위 계층의 ISA 네트워크(**ISA1**)를 완전히 학습시킨 후 상위 계층의 ISA 네트워크를 학습시켰다. 전체 Stacked Convolutional ISA 네트워크에 대한 입력 차원 n은 $20 \times 20 \times 14 = 5{,}600$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. 앞서 미리 학습시킨 **ISA1** 을 $20 \times 20 \times 14$ 차원 비디오 블록의 각 모서리에 분산 배치하면 $2 \times 2 \times 2$ 총 8 개의 **ISA1** 이 하위 계층을 구성하게 된다 (그림 3). 이렇게 배치한 8 개의 ISA1 로부터 총 ㅋ개의 특징값(features)을 얻고, 이를 다시 상위 계층 ISA 네트워크(**ISA2**)의 입력값으로 사용했다. 따라서 **ISA2** 의 입력 차원은 2,400이 된다. **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 (그림 3).

5.2 Norm-thresholding Interest Points Detection

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

본 실험에서 경계값을 30% 로 잡고 동일한 실험을 수행하였고, 그 결과로 생성된 interest point 들을 시각화 한 것이 그림 1 에 나타나 있다.

5.3 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)을 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).

표 1. 샌드위치 요리 비디오의 동작 분류 정확도 (K는 K-fold cross-validation의 실행 회수를 의미)
Table 1. Classification accuracy for actions in sandwich making videos (K: K-fold cross-validation)

	Angle B (Fig 1-B)			Angle C (Fig 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%		

5.4 결과 및 분석

표 1 에 9 개 동작 범주 각각에 대한 binary SVM 분류 결과가 나타나 있다. 표 1 에서 확인할 수 있듯이 대부분의 범주에 대해서 높은 수준의 accuracy 를 보였으며, 총 9 개 범주의 성능 척도를 모두 평균해서 계산한 mean accuracy 는 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.

표 2. Hollywood2 비디오의 동작 분류 정확도
Table 2. Classification accuracy for actions in Hollywood2 video dataset,

	Accuracy	
Action	Angle B (Fig 1-B)	Angle C (Fig 1-C)
Crumpling	97.9% (46/47)	97.9% (46/47)
CuttingSomething	95.7% (45/47)	95.7% (45/47)
DisposeAnObject	95.7% (45/47)	97.9% (46/47)
Reaching	70.2% (33/47)	78.7% (37/47)
ReleaseGraspOfSomething	66.0% (31/47)	66.0% (31/47)
SpreadingOntoSurface	97.9% (46/47)	95.7% (45/47)
Sprinkle	100% (47/47)	100% (47/47)
TurningOnPowerDevice	97.9% (46/47)	97.9% (46/47)
UnWrappingSomething	100% (47/47)	97.9% (46/47)
Mean	91.3%	92.0%

5.5 Self-Taught Learning Paradigm

표 2 에서는 Stacked Convolutional ISA 네트워크를 실험 데이터와 전혀 별개인 Hollywood2 데이터[12]를 이용해서 학습시킨 후, 학습된 네트워크를 이용해서 실험 데이터의 동작을 분류한 결과로서, 표 1 과 비교해 봤을 때 큰 차이가 없는 것을 확인할 수 있다. 이러한 접근은 [2]에서도 언급된 'self-taught learning paradigm'[13]의 측면에서 생각해 볼 수 있다. 'Self-taught learning paradigm'이란 무감독학습 단계에서 실험 데이터와 전혀 별개의 새로운 데이터를 이용해서 유용한 시·공간적 특징들을 학습하는 것을 일컫는다.

ISA 알고리즘을 정적 자연 이미지 패치(static natural image patches)에 적용시키면 일반적으로 Gabor 필터와 같이 이미지 처리에 가장 기본이 되는 선형 필터들을 학습하듯이[11], Stacked Convolutional ISA 네트워크를 비디오 블록에 적용시켜 학습시키면 일반적으로 방위(orientation)와 속도(velocity)에 선택적(selective)인 움직이는 모서리 탐지기(moving edge detectors)를 학습하게 된다[2]. 따라서 동작 인식 및 분류를 행하고자 하는 데이터와 전혀 별개의 영상 데이터를 이용해서 Stacked Convolutional ISA 네트워크를 학습시키더라도 유용한 시·공간적 특징들을 학습할 수 있으며, 이러한 특징들이 가지는 일반성으로 말미암아 전혀 관련이 없는 데이터에서도 상당한 성능을 발휘할 수 있게 되는 것이다.

6. 결론 및 향후 연구

본 논문에서는 [2]에서 제시한 핵심 알고리즘, 즉 Stacked Convolutional ISA 를 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 장점

이 있고, 또한 ISA 알고리즘의 측면에서는 생물학적으로 타당한(biologically plausible) 특징들을 학습할 수 있다는 장점이 있다[2,11,14]. 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.

참 고 문 헌

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

[2] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis, In *CVPR*, 2011.

[3] A. Hyvärinen and P. 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] H. Wang *et al.*, Evaluation of local spatio-temporal features for action recognition, In *BMVC 2009 – British Machine Vision Conference*, 2009.

[5] M. A. Giese and T. Poggio, Neural mechanisms for the recognition of biological movements, *Nat. Rev. Neurosci.*, 4(3):179-192, 2003.

[6] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, A biologically inspired system for action recognition, In *ICCV*, pp.1-8, 2007.

[7] Y. Bengio, Learning deep architecture for AI, *Foundation and Trends in Machine Learning*, 2(1):1-127, 2009.

[8] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition, In *Proceedings of the IEEE*, 86, pp.2278-2324, 1998.

[9] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, Greedy layer-wise training of deep networks, In *NIPS*, 2006.

[10] S. Ji, W. Xu, M. Yang, and K. Yu, 3d convolutional neural networks for human action recognition, In *ICML*, 2010.

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

[12] M. Marszalek, I. Laptev, and C. Schmid, Actions in context, In *CVPR*, 2009.

[13] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, Self-taught learning: transfer learning from unlabeled data, In *ICML*, pp.759-766 2007.

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