Sparse Representation 기반의 인간행동인식에 대한 지역특징과 전역특징 비교

황정현 민현석 노용만
한국과학기술원
neojordan@kaist.ac.kr, hsmin@kaist.ac.kr, ymro@ee.kaist.ac.kr

Comparison of Local and Global Features for Sparse Representation-based Human Action Recognition

Hwang, Jung-Hyon Min, Hyun-seok Ro, Yong Man
Korea Advanced Institute of Science and Technology (KAIST)

요 약

인간행동의 자동인식 기술은 영상보안 및 인간-사물 상호작용 분야에 핵심적 기술이다. 그러나 실제 비디오 환경에서는 인간 행동의 다양성 및 잡음 등 많은 제한점들로 인해 효과적인 행동인식에 어려움이 있다. 최근 이러한 문제점을 해결하기 위하여 많은 영상 처리 및 인식 분야에서 연구되고 있는 sparse representation 기반의 방법들이 제시되고 있다. 이에 본 논문에서는 효과적으로 sparse representation을 행동인식에 적용하고, sparse representation 기반 인간행동인식을 위해 사용되는 지역특징 및 전역특징에 대하여 비교했다.

1. 서론

영상에서의 인간 행동 자동인식 기술은 영상 보안 및 인간-사물 상호작용 분야에 핵심적인 기술이다. 최근 들어 얼굴 인식 등 많은 영상 처리 및 인식 환경에서 다양한 잡음 및 복잡성 등의 문제를 극복하기 위하여 sparse representation (SR) 기반 인식 기법들이 많이 제안되어 왔다 [1]. Sparse representation은 입력 신호를 몇몇의 예제 특징들의 선형 조합으로 표현하는 방법으로써, 이런 예제 특징을 기반으로 영상을 인식하는 방법이다 [2]. 이런 예제 특징을 기반으로 하기에 잡음에 강하고, 다양성이 높은 인식 분야에서는 강인한 판단을 할 수 있는 장점을 가지고 있다. 이에 본 논문에서는 SR 기법을 행동인식에 적용하고 SR의 예제목록(dictionary)을 구성하는 특징벡터를 지역특징과 전역특징으로 분류하여 비교한다.

본 논문의 구성은 다음과 같다. 2 절에서는 일반적인 SR 기반 행동인식 시스템을 설명하고, 3 절에서는 실험을 통해 지역 및 전역특징에 따른 SR기반 행동인식 성능을 분석한다. 4 절은 본 논문의 결론으로 마무리 한다.

2. SR 기반 행동인식 시스템

SR은 학습된 dictionary 의 몇몇 예제 신호(atoms)들의 선형조합으로 입력신호를 표현한다. 이 선형조합은 원소들이 대부분 0의 값을 갖는 sparse coefficient 벡터로 나타난다. Sparse representation은 신호획득 및 압축에 있어 유용할 뿐 만 아니라, 신호 [2] 및 text 분류 [3]는 물론 얼굴인식 [1]과 같은 computer vision문제에도 효과적이다.

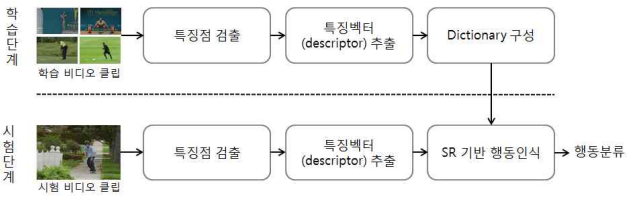

그림 1. SR기반 행동인식시스템

이제 본 논문에서 사용한 SR기반의 행동인식에 대해 간략히 소개하도록 하겠다. 그림 1에서 표현된 것과 같이 SR기반 행동인식 시스템은 두 단계로 구성 된다: 1) 학습단계; 2) 시험단계. 학습단계에서는 학습용 비디오 클립으로부터 특징벡터를 추출해 dictionary를 구성하고 시험단계에서는 구성된 dictionary를 이용하여 시험용 비디오 클립을 특정 행동으로 분류한다. SR을 적용함에 있어 시험용 비디오 클립을 표현하는 dictionary $\mathbf{D}$는 통상 다음과 같이 표현된다. [2][4]:

$$\mathbf{D} = [\mathbf{z}_1^1,...,\mathbf{z}_{N_1}^1,...,\mathbf{z}_1^i,...,\mathbf{z}_{N_i}^i,...,\mathbf{z}_1^K,...,\mathbf{z}_{N_K}^K] \in \Re^{d \times N} \quad , (1)$$

위 식에서 Z_j는 i 번째 행동그룹의 j 번째 학습용 비디오 클립의 특징벡터들을 의미하며, N_i 와 K 는 각각 i 번째 행동의 학습용 비디오 클립 숫자와 전체 행동 그룹의 수를 의미한다. d 는 비디오 클립의 행동을 표현하는 특징 벡터의 차원을 나타낸다. 또한, N 은 총 학습용 비디오 클립의 수(즉, 모든 행동의 학습용 비디오클립 수의 합)를 뜻한다.

주어진 $\mathbf{D}$ 에 대하여, 시험용 비디오 클립 $\mathbf{V}$ 의 특징 벡터 $\mathbf{y}$는 다음과 같이 표현될 수 있다.

$$\mathbf{y} \approx \mathbf{D}\mathbf{x} \in \Re^d, \quad (2)$$

여기에서 $\mathbf{x} = [x_1^1, ..., x_{N_1}^1, ..., x_1^i, ..., x_{N_i}^i, ..., x_1^K, ..., x_{N_K}^K]$ 는 $\mathbf{V}$ 의 sparse 선형 표현이다. x_j^i 는 i 번째 행동의 j 번째 학습용 비디오 클립과 연관된 sparse coefficient 값을 나타낸다.

식(2)의 sparse 해인 $\mathbf{x}$ 를 구하면, 각 행동그룹에 대한 residual error를 다음 식을 이용해 계산할 수 있다.

$$r_i(\mathbf{y}) = \left\| \mathbf{y} - \mathbf{D}\delta_i(\mathbf{x}) \right\|_1 , \quad (3)$$

r_i 는 i번째 행동에 대한 residual error를, $\delta_i(\mathbf{x})$ 는 i번째 행동과 연관된 x의 요소들만 0이 아닌 새로운 벡터를 뜻한다. 따라서 식 (3)을 통해 가장 작은 residual error를 반환하는 i번째 행동으로 시험 비디오 $\mathbf{V}$를 분류할 수 있다.

3. 실험 및 분석

실험은 9분류의 스포츠 비디오 클립으로 구성된 UCF Sports Action Data Set [5]을 대상으로 실시했다. 9 종류의 스포츠는 다음과 같다: diving, golf swinging, kicking, lifting, riding, running, skating, swinging, walking.

특징벡터의 특성에 따른 SR기반 행동인식의 성능분석을 위해 최근 행동인식연구에 많이 사용되고 있는 지역특징(local key point) 기법과 전역특징 기법을 이용하여 특징벡터를 추출하여 dictionary를 구성하고 각각의 성능을 비교하였다. 전역특징은 비디오클립을 25*25*13 픽셀단위로 Dense sampling [6] 했으며, 지역특징은 σ=4, τ=2 파라미터를 적용한 Cuboid (25*25*13픽셀) [7] 로 검출했다. 이렇게 검출된 특징(cube)을 1440차원 gradient 1D histogram으로 표현했다. 그리고 다시 SR의 필요조건인 dictionary over-completeness 확보 및 빠른 연산을 위해 1440차원 특징벡터를 random projection [8]을 통해 144차원으로 축소시켜 dictionary를 구성했다.

그림 2 은 각 특징벡터로 구성된 dictionary를 사용한 SR 기반 행동인식 성능을 나타낸다. 실험 결과에서와 같이 Diving, Lift, Swing 등 행동특징 및 배경특징이 모두 강한 행동분류에서는 전역·지역특징 모두 좋은 성능을 보였다.

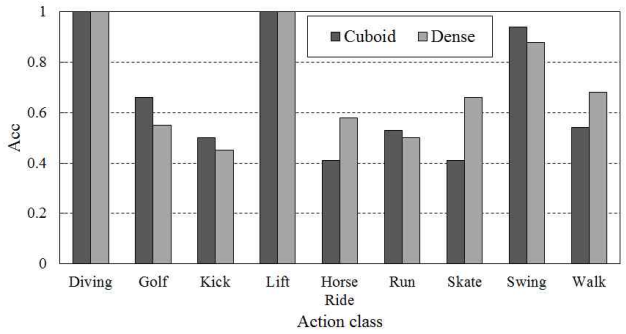

그림 2. 전역특징 및 지역특징에 따른 SR기반 행동인식 성능

한편, golf, kick 등 배경특징이 상대적으로 약한 행동은 지역특징으로 구성한 dictionary를 이용한 행동인식이 보다 좋은 성능을 보인반면, horse riding, stating, walk 등 배경의 특성이 높은 행동분류에서는 전역특징으로 구성된 dictionary를 이용한 인식결과가 상대적으로 높았다.

즉, 통상적으로 SR기반 행동인식은 지역특징으로 구성된 dictionary를 사용하는데 반해, 본 실험은 배경정보를 이용하는 특징으로 dictionary를 구성할 경우 성능향상을 보이는 행동분류(class)가 존재함을 보였다.

4. 결론

본 논문에서는 SR 기반 행동인식 시스템에서 전역·지역특징에 따른 성능분석 결과를 제시하였다. SR기반 행동인식 시스템은 학습용 비디오에서 특징벡터를 추출해 dictionary를 구성하고, 시험용 비디오 특징벡터의 Sparse 해를 구하여 가장 작은 residual error를 반환하는 class로 분류하게 된다. 이때 dictionary를 전역특징 및 지역특징으로 각각 구성하여 실험한 결과 배경의 특성이 큰 일부 행동에서는 전역특징을 이용한 성능이 우수함을 관찰하였다.

향후, 실험결과를 바탕으로 행동분류에 따라 배경정보를 선택적으로 이용할 수 있는 SR기반 행동인식 연구를 진행할 계획이다.

Acknowledgement

이 논문은 2013년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2011-0011383)

참고문헌

[1] J Wright A Yang A Ganesh S Sastry and Y Ma "Robust Face Recognition via Sparse Representation" *IEEE Trans Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210 - 227, 2009.

[2] K. Huang and S. Aviyente, "Sparse Representation for Signal Classification," *In Adv. NIPS*, 2006.

[3] M Zhaoa S Lia and J Kwokb "Text Detection in Images using Sparse Representation with Discriminative Dictionaries" *Image and Vision Computing*, vol. 28, no. 12, pp. 1590 - 1599, 2010.

[4] T Guha and R K Ward "Learning Sparse Representations for Human Action Recognition" *IEEE Trans Pattern Anal. Mach. Intell.*, vol. 34, no. 8, August 2012.

[5] Mikel D. Rodriguez, Javed Ahmed, and Mubarak Shah Action MACH: A Spatio-temporal Maximum Average Correlation Height Filter for Action Recognition.

[6] Konrad Schindler and Luc van Gool "Combining Densely Sampled Form and Motion for Human Aciton Recognition" Patter Recognition Lecture Notes in Computer Science Volume 5096, pp 122-131, 2008.

[7] P Dollar V Rabaud G Cottrell and S Belongie "Behavior recognition via Sparse Spatio-Temporal Features" *IEEE Joint Int'l Workshop Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp. 65 - 72, 2005.

[8] R Baraniuk and M Wakin "Random Projections of Smooth Manifolds" *Foundations of Computational Math.*, vol. 9, pp. 51 - 77, 2009.