다중 판별기를 이용한 비디오 행동 인식

*김세민 **노용만
*한국과학기술원, 정보통신공학과 **한국과학기술원, 전기및전자공학과
*resemin@kaist.ac.kr **ymro@ee.kaist.ac.kr

Human Action Recognition in Videos using Multi-classifiers

*Kim, Semin **Ro, Yong Man
*Dept. Information and Communications Engineering, KAIST
**Dept. Electrical Engineering, KAIST

요약

최근 다양한 방송 및 영상 분야에서 사람의 행동을 인식하여는 연구들이 많이 이루어지고 있다. 영상은 다양한 형태를 가질 수 있기 때문에 제약된 환경에서 유용한 템플릿 방법들보다 특징점에 기반한 연구들이 실제 사용자 환경에서 더욱 관심을 받고 있다. 특징점 기반의 연구들은 영상에서 움직임이 발생하는 지점들을 찾아내어 이를 3차원 패치들로 생성한다. 이를 이용하여 영상의 움직임을 히스토그램에 기반한 descriptor(서술자)로 표현하고 학습기반의 판별기(classifier)로 최종적으로 영상 내에 존재하는 행동들을 인식하였다.

그러나 단일 판별기를 이용한 다양한 영상 인식을 수용하기에는 힘들다. 최근에 이를 개선하기 위하여 다중 판별기를 활용한 연구들이 영상 판별 및 물체 검출 영역에서 사용되고 있다. 따라서 본 논문에서는 행동 인식을 위하여 support vector machine과 spare representation을 이용한 decision-level fusion 방법을 제안하고자 한다. 제안된 논문의 방법은 영상에서 특징점 기반의 descriptor를 추출하고 이를 각각의 판별기를 통하여 판별 결과들을 획득한다. 이 후 학습단계에서 획득된 가중치를 활용하여 각 결과들을 융합하여 최종 결과를 도출하였다. 본 논문에 실험에서 제안된 방법은 기존의 융합 방법보다 높은 행동 인식 성능을 보여 주었다.

1. 서론

현재 다양한 방송 컨텐츠에서 사람의 특정한 행동 또는 행위를 검출해내는 연구가 진행되고 있다. 또한 스마트 TV 등의 보급으로 시청자들의 움직임을 반영하려는 연구도 많이 이루어 졌다. 이러한 영상 컨텐츠에서 사람의 크기 및 위치가 다양하기 때문에 최근 특징 점에 기반한 행동 인식에 관한 연구가 많이 진행되고 있다.

특징점 기반의 연구는 특징점 추출, local descriptor 추출, bag-of-words 추출, 행동 판별 등의 총 4가지 단계로 이루어 진다. 먼저 영상에서 움직임이 발생하는 영역 검출하여 3차원 패치들로 추출해낸다. 특징점을 추출해내는 다양한 방법들이 제안되고 있는데 대표적으로 Cuboids [1], Harris3D [2], Hessian [3] 등의 방법들이 있다. 이후 local descriptor 들이 3차원 패치들로부터 추출되는데 HOG [4], HOF [5], HOG3D [6] 등의 방법들이 주로 사용된다. 추출된 local descriptor들은 'Bag-of-Words' [7] 라는 히스토그램으로 표현이 되고 판별 모델 등을 통하여 해당 영상이 어떠한 행동을 포함하고 있는지를 판별하였다.

현재 대부분의 행동인식 방법들은 다양한 local descriptor를 융합하는데 초점을 두며 하나의 판별 모델을 사용하였다. 그러나 최근 보행자 검출이나 [8] 영상 분류 [9]를 위하여 다수개의 판별 모델을 사용하여 영상 인식 성능을 향상시킨 연구가 많이 진행 되었다. 이러한 방법들의 특징은 다양한 특성의 판별 모델을 사용함으로써 서로 상호보완적 효과를 얻을 수 있기 때문에 판별 성능을 향상 시킬 수가 있었다.

본 논문에서는 support vector machine (SVM)과 sparse representation (SR)을 등을 활용하여 다중 판별 모델 기반한 행동 인식 방법을 제안하고자 한다. SVM과 SR 현재 대부분의 행동 인식 연구에서 많이 사용되어지고 있다. 또한 Liu와 Li [9]에 따르면 이 두가지 판별 모델을 사용하는 것이 영상 판별에 효율적인 성능을 가져다 주었다고 한다. 그러나 Liu의 방법은 주어진 테스트 셋(set)에서 성능을 극대화시키는 것에 초점을 주었기 때문에 실제 응용과는 거리가 있었다.

따라서 본 논문에서는 위의 SVM과 SR을 통한 다양한 행동인식 결과를 효율적으로 융합하는 것에 목적을 둔다. 이를 위하여 학습단계에서 각 판별 모델의 성능을 행동의 종류마다 측정하고, 두 판별 모델의 정확성에 따라 각 행동에 대한 가중치(weight)들을 계산한다. 따라서 각 판별 모델들은 행동 마다 정확도가 비교가 가능하고 가중치의 차이만큼 정확도가 높은 판별모델의 결과가 중요시 될 수 있게 된다. 이후 실제 테스트 영상이 입력되면 각 판별모델을 통하여 일차적인 결과를 획득한다. 이 후 각 결과들은 정규화(normalization) 과정을 거치고 학습단계에서 획득한 가중치를 이용하여 두 결과를 융합(fusion)하

게 된다. 마지막으로 융합된 결과 중 가장 높은 값을 보여주는 행동으로 테스트 영상을 판별하게 된다. 본 논문에서 제안한 방법을 사용할 경우 기존의 방법보다 약 2% 이상의 성능 향상을 얻을 수가 있었다.

본 논문의 구성은 다음과 같다. 2장에서는 영상으로부터 특징을 추출해내는 과정을 서술하며 3장에서는 본 논문에서 제안한 융합 방법을 설명한다. 4장에서는 제안된 융합 방법을 통하여 획득한 행동 인식 결과를 보여주며, 마지막 장에는 본 논문의 결론을 보여준다.

2. 행동 인식 특징 추출

본 논문에서는 행동 인식을 위한 영상 특징을 추출하기 위하여 cuboid [1] 와 histograms of oriented gradients (HOG) [4]를 활용하였다. Cuboid는 많은 기존 연구에서 다른 특징점 추출보다 행동 인식에서 우수한 성능을 보여 주었다 [7]. 먼저 Cuboid는 영상 내에서 특징점을 추출하기 위하여 움직임이 많이 발생하는 지점을 찾는다. 영상의 움직임 정보를 구하는 함수 R은 아래 식과 같이 정의 된다.

$$R=(I*g*h_{ev})^2+(I*g*h_{od})^2 , \quad (1)$$

여기서 I는 입력된 영상이며, g는 Gaussian smoothing 함수, $g(x,y;\sigma)$ 이며 σ는 필터의 scale이다. hev와 hod 는 1D-Gabor 필터의 quadrature pair 함수들이며 아래와 같이 정의 된다.

$$h_{ev}(t;\tau,\omega) = -\cos(2\pi t\omega)e^{-t^2/\tau^2}, \quad (2)$$

$$h_{od}(t;\tau,\omega) = -\sin(2\pi t\omega)e^{-t^2/\tau^2}, \quad (3)$$

여기서 $\omega = 4/\tau$ 이며, τ는 시간 축 scale이다. 위의 과정을 통하여 함수 R에서 움직임이 높게 나오는 부분에 대하여 Liu [9]의 방법에 따라 25 × 25 × 13의 크기를 가지는 3D 패치를 추출하였다.

추출된 3D 패치들에서 HOG descriptor를 추출하였다. HOG는 3D 패치의 한 단면마다 공간축 방향으로 미분하여 기울기의 방향(orientaion)과 크기(magnitude)를 구한다. 이렇게 추출된 기울기의 방향과 크기를 이용하여 3D패치를 히스토그램화 할 수 있다. 본 논문에서는 Laptev [5]의 방법에 따라 각 3D 패치를 3 × 3 × 2의 셀로 나누고 각 셀에서 기울기를 4개의 방향으로 나누어 히스토그램을 추출하였다. 즉, 하나의 3D패치는 72차원의 크기를 가지는 HOG로 추출 된다.

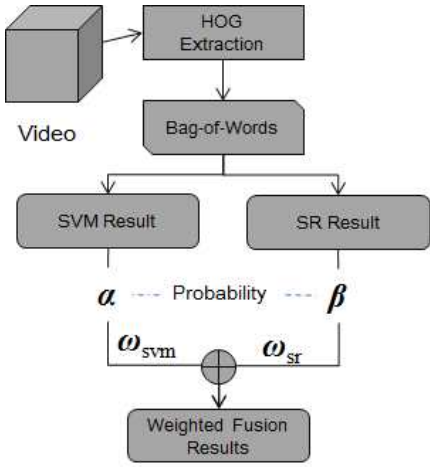
영상에서 HOG들이 추출되면 bag-of-word[7]를 활용하여 하나의 히스토그램으로 표현한다. 왜냐하면 영상마다 다양한 수의 HOG들이 추출되어 직접적인 매칭으로 비교가 힘들기 때문에, 이를 동일한 크기를 가지는 히스토그램으로 정규화(normalization)를 해야 되기 때문이다. Bag-of-word는 HOG들을 사전에 정의된 visual-word들로 할당한다. 이러한 visual-word들은 학습 영상으로부터 추출된 HOG들을 클러스터링을 통하여 획득된다. 본 논문에서는 k-means 알고리즘을 활용하여 visual-word들을 생성하였다. 따라서 테스트 영상이 입력되면 HOG들을 추출하여 Wang[7]의 방법을 참고하여 visual-word들로 할당하고 bag-of-word 히스토그램으로 영상을 표현하였다.

3. 행동 인식 다중 판별 및 융합

본 장에서는 본 논문에서 제안하는 다중 판별기를 활용한 행동 인식 방법에 대하여 다룬다. 1장에서 언급한 것처럼 SVM과 SR을 활용하여 행동 인식 결과를 융합(fusion)한다.

SVM은 기본적으로 두 가지 클래스를 다루는 이진(binary)판별기이다. 이는 두 가지 클래스를 구성하고 있는 샘플들을 특징 공간(feature space)로 투영하여 각 클래스를 가장 잘 구분하는 hyper-plane을 찾는 것에 초점을 두고 있다. 따라서 두 개 이상의 클래스를 구분하기 위하여 모든 클래스 간에 이진 모델을 생성하고 이를 활용하여 다중 판별이 가능한 SVM이 제안되었다 [10]. 본 논문에서는 histogram intersection kernel을 활용한 SVM이 사용되었다.

SR은 신호를 분해하여 이를 재구성할 때에 필요한 가장 대표적인 계수들만 선택한다. 이를 통하여 신호의 크기를 줄일 수 있고 또한 신호가 가장 많이 분포된 영역을 알 수 있기 때문에, 이를 활용한 판별이 가능하다. 현재 다양한 SR 방법이 제안되고 있는데 본 논문에서는 Liu[9]의 방법에 따라 orthogonal matching pursuit (OMP)[11]를 사용하였다.



[그림 1] 제안한 방법의 다중 판별 융합 순서도

그림 1은 본 논문에서 제안하는 다중 판별 융합 순서도를 보여주고 있다. 영상에서 HOG를 추출하여 이를 bag-of-words 히스토그램으로 표현하고 SVM과 SR 판별기를 통하여 각각 일차적인 판별 결과를 추출한다. 그리고 각 결과들을 $\boldsymbol{\alpha}$와 $\boldsymbol{\beta}$로 표현되는데 각각 SVM과 SR의 결과 값들을 0에서 1사이의 확률 값으로 표현한 것이며 아래와 같이 정의 된다.

$$\boldsymbol{\alpha} = \alpha_1, \alpha_2, \ldots, \alpha_J, \quad (4)$$

$$\boldsymbol{\beta} = \beta_1, \beta_2, \ldots, \beta_J, \quad (5)$$

여기서 J는 영상이 판별될 수 있는 행동의 가지수이다. 그리고 각 판별기에 대한 가중치는 ω_{svm}와 ω_{sr}은 학습 단계에서 각 판별 모델의 정확도로 계산된다. 학습 샘플을 S_{tr}이라 하고 J개의 액션을 가지고 있다면 이를 N-fold cross-validation을 SVM과 SR에 대하여 각각 실험을 한다. 이를 통하여 각 학습 모델에 대하여 precision(P)과 recall(R)을 다음과 같이 정의 할 수 있다.

$$P_{svm} = p_{svm,1}, p_{svm,2}, \ldots, p_{svm,J}, \quad (6)$$

$$P_{sr} = p_{sr,1}, p_{sr,2}, \ldots, p_{sr,J}. \quad (7)$$

$$R_{svm} = r_{svm,1}, r_{svm,2}, \ldots, r_{svm,J} \quad (8)$$

$$R_{sr} = r_{sr,1}, r_{sr,2}, \ldots, r_{sr,J}. \quad (9)$$

이렇게 구해진 P와 R을 이용하여 각 모델에 대한 F1-score[7]를 구하여 가중치를 다음과 같이 정의 할 수 있다.

$$\boldsymbol{\omega}_{\mathrm{svm}}(P_{\mathrm{svm}}, R_{\mathrm{svm}}) = \omega_{\mathrm{svm},1}, \omega_{\mathrm{svm},2}, \ldots, \omega_{\mathrm{svm},J}, \quad (12)$$

$$\boldsymbol{\omega}_{\mathrm{sr}}(P_{\mathrm{sr}}, R_{\mathrm{sr}}) = \omega_{\mathrm{sr},1}, \omega_{\mathrm{sr},2}, \ldots, \omega_{\mathrm{sr},J}, \quad (13)$$

$$\omega_i = \frac{2 \times p_i \times r_i}{p_i + r_i}. \quad (14)$$

따라서 일차 판별 결과인 α와 β와 위에서 구해진 가중치 $\boldsymbol{\omega}_{\mathrm{svm}}$와 $\boldsymbol{\omega}_{\mathrm{sr}}$를 통하여 최종 판별 결과 $\mathbf{x}$를 아래와 같은 공식으로 구할 수 있다.

$$\mathbf{x} = x_1, x_2, \ldots x_J, \quad (15)$$

$$x_i = \frac{\omega_{\mathrm{svm},i} \times \alpha_i}{\omega_{\mathrm{svm},i} + \omega_{\mathrm{sr},i}} + \frac{\omega_{\mathrm{sr},i} \times \beta_i}{\omega_{\mathrm{svm},i} + \omega_{\mathrm{sr},i}}. \quad (16)$$

따라서 영상에 최종 행동 인식 결과는 x_i 가 가장 높은 클래스로 판별되어 진다.

4. 실험 결과

본 장에서는 제안된 다중 판별 행동 인식의 방법을 UCF-sports[12]를 활용하여 분석하였다. UCF-sports는 총 10가지의 행동들이 존재하는데 그림 2와 같다. 각 클래스별로 6 ~ 22개의 영상이 존재하고 총 150개의 영상으로 구성되어 있다. 본 실험에서는 Wang [7]의 실험 방법에 따라 실험을 진행 하였다.

(1) (2) (3) (4) (5)

(6) (7) (8) (9) (10)

[그림 2] UCF-sports의 행동 구성. (1) diving, (2) golf, (3) kick, (4) lifting, (5) riding-horse, (6) run, (7) skateboarding, (8) swing-bench, (9) swing-side, (10) walk.

먼저 SVM과 SR을 각각 단독적으로 사용하였을 경우에 성능을 confusion table로 계산하였다. 표 1과 2는 각각 의 성능을 보여준다. 이때에 ID는 그림 2의 각 행동 들을 나타낸다. 두 표에서 보이는 것처럼 전체적으로 비슷한 성능을 보여주고 있으나 3번 kick 행동이나 4번 lifting과 같이 한쪽 판별기가 우수한 성능을 보이고 있는 것을 확인 할 수 있다. 즉 kick 행동을 판별할 때에 SVM 결과에 더 높은 가중치를 부여하고 lifting을 판별할 때에는 SR 결과에 더 높은 가중치를 부여한다면 최종적으로 향상된 판별 결과를 얻을 수 있다. 표 3은 본 논문에서 각 판별 모델을 단독으로 사용했을 때와 기존 융합 방법[9], 본 논문에서 제안한 방법을 비교하여 최종적인 행동인식결과를 보여주고 있다. 표에서 보이는 것처럼 제안된 방법이 가장 우수한 성능을 보여주고 있는 것을 확인 할 수 있다.

5. 결론

본 논문에서는 SVM과 SR 판별기를 이용하여 영상내에 행동을 인식하는 방법을 제안하였다. 제안된 방법은 두 개의 판별 결과를 융합하기 위하여 학습모델 단계에서 각 판별기에 대한 가중치를 계산하고, 이를 활용하여 기존의 융합 방법보다 우수한 성능을 보였다.

[표 1] SVM을 이용한 행동 인식 결과 (Confusion table).

ID	1	2	3	4	5	6	7	8	9	10
1	**1.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	**0.72**	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.17
3	0.00	0.20	**0.60**	0.00	0.05	0.10	0.05	0.00	0.00	0.00
4	0.00	0.00	0.00	**0.83**	0.00	0.00	0.00	0.00	0.00	0.17
5	0.08	0.00	0.17	0.00	**0.58**	0.17	0.00	0.00	0.00	0.00
6	0.08	0.15	0.23	0.00	0.08	**0.46**	0.00	0.00	0.00	0.00
7	0.00	0.08	0.00	0.00	0.00	0.08	**0.67**	0.00	0.00	0.17
8	0.00	0.00	0.05	0.00	0.00	0.00	0.00	**0.95**	0.00	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**1.00**	0.00
10	0.00	0.18	0.09	0.00	0.00	0.00	0.00	0.00	0.00	**0.73**

[표 2] SR을 이용한 행동 인식 결과 (Confusion table).

ID	1	2	3	4	5	6	7	8	9	10
1	**1.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	**0.72**	0.06	0.00	0.00	0.00	0.00	0.00	0.06	0.17
3	0.00	0.20	**0.30**	0.05	0.05	0.30	0.05	0.00	0.00	0.05
4	0.00	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	0.00	0.00
5	0.08	0.00	0.08	0.00	**0.58**	0.25	0.00	0.00	0.00	0.00
6	0.08	0.23	0.15	0.00	0.15	**0.38**	0.00	0.00	0.00	0.00
7	0.00	0.08	0.00	0.00	0.00	0.00	**0.75**	0.00	0.00	0.17
8	0.00	0.00	0.00	0.00	0.00	0.05	0.00	**0.95**	0.00	0.00
9	0.00	0.00	0.08	0.00	0.00	0.00	0.00	0.00	**0.92**	0.00
10	0.00	0.14	0.00	0.00	0.00	0.00	0.05	0.00	0.00	**0.82**

[표 3] 최종 행동 인식 결과

Actions	SVM	SR	Liu [9]	Proposed
Diving	1.000	1.000	1.000	1.000
Golf	0.722	0.722	0.667	0.667
Kick	0.600	0.300	0.600	0.550
Lifting	0.833	1.000	0.833	1.000
Horse	0.583	0.583	0.667	0.667
Run	0.462	0.385	0.462	0.462
Skate	0.667	0.750	0.750	0.750
Bench	0.950	0.950	0.950	0.950
Side	1.000	0.923	1.000	1.000
Walk	0.727	0.818	0.727	0.818
Average	**0.754**	**0.743**	**0.766**	**0.786**

Acknowledgement

본 연구는 미래부가 지원한 2013년 정보통신·방송(ICT) 연구개발 사업의 연구결과로 수행되었음

참고문헌

[1] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, Behavior recognition via sparse spatio-temporal features, in Proc, IEEE Int. Work. Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005, pp. 65-72.
[2] I. Laptev, and T. Lindeberg, Space-time interest points, in Proc, IEEE Int. Conf. Computer Vision, 2003, pp. 432-439.
[3] G. Willems, T. Tuytelaars, and L. Gool, An Efficient Dense and

Scale-Invariant Spatio-Temporal Interest Point Detector, in Proc, Euro. Conf. Computer Vision, 2008, pp. 650-663.
[4] N. Dalal, and B. Triggs, Histograms of oriented gradients for human detection, in Proc, IEEE Int. Conf. Computer Vision and Pattern Recognition, 2005, pp. 886-893.
[5] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, Learning realistic human actions from movies, in Proc, IEEE Int. Conf. Computer Vision and Pattern Recognition, 2008, pp. 1-8.
[6] A. Klaser, M. Marzalek, and C. Schmid, A spatio-temporal descriptor based on 3D-gradients, in Proc. British Machine Vision Conf., 2008, pp. 995-1004.
[7] H. Wang, M.M. Ullah, A. Klaser, I. Laptev, and C. Schmid, Evaluation of local spatio-temporal features for action recognition, in Proc. British Machine Vision Conf., 2009.
[8] O.L. Junior, D. Delgado, V. Goncalves, and U. Nunes, Trainable Classifier-Fusion Schemes: an Application to Pedestrian Detection, in Proc. IEEE conf. Intelligent Transportation Systems, 2009, pp. 432-437.
[9] H. Liu, and S. Li, Decision fusion of sparse representation and support vector machine for SAR image target recognition, Neurocomputing Vol. 113, 2013, pp. 97-104.
[10] C.C. Chang, and C.J. Lin, http://www.csie.ntu.edu.tw /~cjlin/libsvm/
[11] A.Y. Yang, S.S. Sastry, A. Ganesh, and YiMa, Fast ℓ 1-minimization algorithms and an application in robust face recognition: A review, in Proc. IEEE Int. Conf. Image Processing, 2010, pp.1849-1852.
[12] UCF Sports, http://crcv.ucf.edu/data/UCF_Sports_Action.php