SIFT 와 Particle 특징 궤적 기반 행동인식

유정민*, 양이화*, 전문구*
*광주과학기술원 정보통신공학부
e-mail : estevan119@gist.ac.kr

Action recognition by SIFT and particle feature trajectories

Jeong-Min Yu*, E-hwa Yang*, Moon-Gu Jeon*
* School of Information and Communications, Gwangju Institute of Science and Technology

요 약

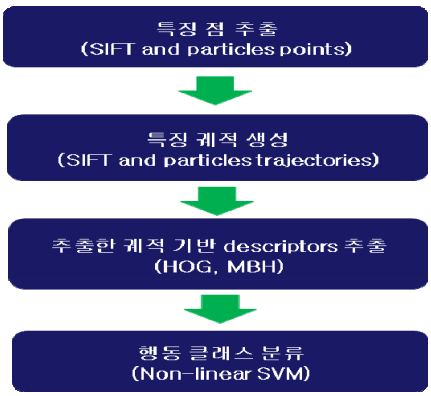
본 논문에서는 SIFT 와 particle 특징 궤적을 이용한 새로운 행동 인식 시스템을 제안한다. 먼저, 영상에서 중요한 지역적 특징 정보를 얻기 위하여 SIFT 특징 점들을 탐지하고, 탐지한 특징 점들을 SIFT descriptor matching 기법을 이용하여 그 궤적을 추출한다. 또한, SIFT 특징 궤적들의 수량이 적은 점과 영상내의 조명변화, 부분적 가려짐 등의 변화로 인해 SIFT 특징 궤적이 종종 없어지는 단점을 보완하기 위하여, SIFT 특징 궤적 주위에 particle 점들을 탐지하고, dense optical flow 기법을 기반으로 그 특징 궤적을 추출한다. 그리고 SIFT 와 particle 궤적의 중요도를 조절하기 위해 가중치를 부여한다. 제안한 행동 인식 시스템의 효율성을 범용 데이터 셋을 이용한 실험을 통해 증명하였다.

1. 서론

사람 행동 인식은 컴퓨터 비전 분야에서 활발한 연구가 진행되고 있는 주제로써 비디오 감시, 자동적 비디오 인덱싱, 사람-컴퓨터 상호작용 등의 다양한 응용분야에서 그 기술이 응용되고 있다. 하지만, 비디오 영상내의 큰 intra-class 변화, 부분적 가려짐, 저 해상도, 복잡한 배경 등의 환경적 제약 때문에 행동인식이 어렵다. 최근, 이러한 문제들을 극복하기 위해 영상에서 지역적 특징 정보를 기반한 지역적 시-공간 특징(local spatio-temporal feature) 행동인식 기법 [4, 9, 10] 들이 좋은 결과를 보여주고 있다.

지역적 시-공간 특징 정보는 보통 3D 특징 점 [9, 10] 과 특징 점들의 궤적 형태로 [3, 11] 추출된다. 먼저, 3D 특징 점은 사람의 지역적 모션 및 형태 특성을 잘 표현한다. 하지만, 이러한 3D 특징 점은 짧은 시간 내의 간단한 모션 정보만을 추출한다. 따라서, 복잡한 형태의 모션이나 긴 시간의 모션을 표현하기엔 취약한 단점이 있다. 최근 이러한 문제점을 극복하기 위해, 탐지된 특징 점을 추적하여 그 특징 궤적을 추출하는 기법들이 나왔다. 예를 들어, [3] 에서는 SIFT 특징 점을 탐지하고 그 궤적을 SIFT descriptor matching 기법을 이용하여 그 궤적을 추출하여 행동인식을 하였다. 또한, [11] 에서는 Harris3D 특징 점들을 탐지하고, 그 점들을 KLT tracker [6] 를 이용하여 그 궤적을 추출하였다. 이러한 기법들은 좋은 결과를 보여주었지만, 특징 궤적들의 수량이 적은 점과 영상내의 조명변화, 부분적 가려짐 등의 변화로 특징 궤적이 종종 없어지는 단점을 가지고 있다.

본 논문에서는, 이러한 단점들을 해결하기 위해 weighted SIFT and particle (WSAP) 궤적을 기반으로 한 행동 인식 시스템을 제안한다. 먼저, 비디오 영상에서 중요한 지역적 특징 정보를 유지하기 위해 SIFT 특징 점들을 탐지한다. 그리고 SIFT 특징 점의 주변으로 조밀하게 particle 특징 점들을 뿌려서 주변 문맥(context) 정보를 추출한다. 매 frame 에 탐지 된 SIFT 와 particle 특징 점들을 SIFT descriptor matching 과 dense optical flow 기법으로 궤적들을 각각 추출한다. 추출 된 SIFT 궤적의 수가 particle 궤적의 수보다 상대적으로 많이 적기 때문에 이들의 궤적의 중요도를 조절할 필요가 있다. 이 궤적들의 중요도를 조절하기 위해, 탐지된 SIFT 와 particle 점들의 정보를 기반으로 궤적들에 가중치를 주었다. 최종적으로, 제안한 WSAP 특징 궤적을 따라 가중치가 부여된 시-공간 volume descriptor 들을 생성하고, support vector machine (SVM) 분류기를 통하여 행동 클래스들을 분류한다. 그림 1 은 제안한 행동 인식 시스템의 개요이다.



(그림 1) 제안한 행동인식 시스템 개요

2. 제안하는 방법

행동인식에서 특징 점 탐지와 추적 기술은 성능에 큰 영향을 미치는 요소이다. 먼저, 영상 내에서 특징 점을 추출하기 위해서 SIFT [1] 기법을 이용한다. SIFT 특징 점들은 corner 와 같은 특징 점들에 비해 크기, 회전에 불변한 특성을 가지고 있어 환경적 제약이 있는 영상에서 강인한 성능을 보인다. 매 frame 마다 SIFT 점들을 탐지하고, 이를 SIFT descriptor matching 방법을 이용하여 SIFT 특징 궤적들을 추출한다. 자세히 말하자면, frame t 영상에서 하나의 SIFT point $p_t^r = (x_t^r, y_t^r)$ 가 다음 frame t+1 의 SIFT point $p_{t+1}^r = (x_{t+1}^r, y_{t+1}^r)$ 를 선택할 때 그들의 descriptors 간의 가장 작은 차이 값을 가지는 point 를 선택한다.

$$p_{t+1}^r = \arg\min_{\hat{p}_{t+1}^r} \left\| Des[p_t^r] - Des[\hat{p}_{t+1}^r] \right\|_1, \qquad (1)$$

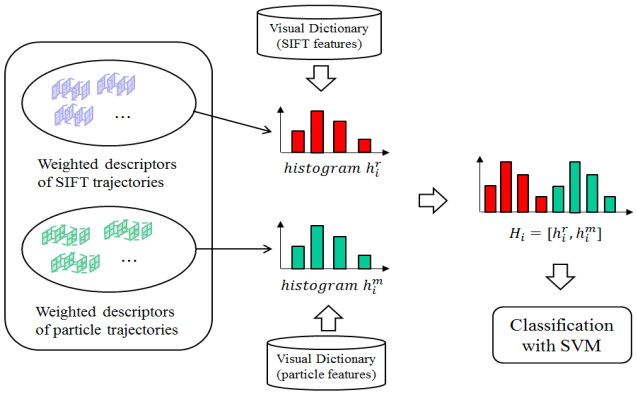
$Des[p_t^r]$ 는 SIFT point p_t^r 의 descriptor 를 의미하고, $Des[\hat{p}_{t+1}^r]$ 는 frame t+1 영상의 후보 SIFT point $\hat{p}_{t+1}^r$ 의 descriptor 를 의미한다.

추출된 SIFT 특징 궤적들이 행동인식에 중요한 지역의 정보를 가지고 있지만, 이 특징 궤적들의 수량이 적고 종종 끊어지는 단점이 있다. 이러한 문제점을 해결하기 위해 SIFT points 주변에 particle points 를 탐지하고, 이를 dense optical flow 기법을 이용하여 그 궤적들을 생성한다. Particle points 는 탐지된 SIFT points 주변 $Z \times Z$ spatial window 내에서 W pixels 들의 간격으로 샘플링 한다. Particle points 를 뽑는 기준은 Shi and Tomasi [2] 에 corner 특징 점들을 뽑는 기법을 사용하였다. 추출한 particle points 들은 dense optical flow 기법을 기반한 [3] 의 point matching 방법을 사용하였다. 이 방법은 기존의 KLT tracker [6] 특징 점 추적기술 보다 급격한 영상의 움직임에 강인하다. 자세하게, frame t 영상내의 particle point $p_t^m = (x_t^m, y_t^m)$ 는 dense optical flow $G = (u_t, v_t)$ 정보와 median filtering 기법을 이용하여 다음 frame 의 particle points 를 선택한다.

$$p_{t+1}^m = (x_{t+1}^m, y_{t+1}^m) = (x_t^m, y_t^m) + (M * G)\left|(\bar{x}_t^m, \bar{y}_t^m)\right. \qquad (2)$$

M 은 median filtering kernel 이고, $*$ 는 the convolution operator 를 의미하고, $(\bar{x}_t^m, \bar{y}_t^m)$ 는 (x_t^m, y_t^m) 의 인근(이웃) 점들을 의미한다.

추출된 SIFT 궤적들은 particle 궤적 보다 더 중요한 정보를 담고 있지만, particle 궤적의 수가 상대적으로 많기 때문에 그 성능에 큰 영향을 주지 못하는 단점이 있다. 이러한 문제점을 극복하기 위하여, 탐지된 SIFT 와 particle points 의 분포 정보를 기반한 새로운 가중치 방법을 제안한다. 우선, 각 frame t 마다 SIFT points n_t^r 와 particle points n_t^m 을 계산하다. 그리고 나서, 각 SIFT points 에 가중치 $w_t^r = 1/n_t^r$, particle points 에 $w_t^m = 1/n_t^m$ 의 가중치를 부여한다. 최종적으로, 궤적의 길이가 미리 정의된 L 이 되면, 각 추출된 궤적에 가중치 W^q 를 부여한다.



(그림 2)시-공간 volume descriptor 기반한 행동분류

$$W^q = \frac{1}{L}\sum_{t=1}^{L} w_t^q, \; where \quad q = \{r, m\}. \qquad (3)$$

가중치가 부여된 WSAP 특징 궤적 주변으로, 관심 대상의 모션 및 형태 정보를 표현하기 위해 시-공간 volume descriptor 를 사용한다. Volume 크기는 $N \times N \times L$ 이고, N 는 pixel 수이고, L 는 궤적 길이를 의미한다. 사용되는 visual descriptor 로는 histograms of oriented gradients (HOG) [5] 과 motion boundary histograms (MBH) [4] 이다. HOG descriptor 는 사람의 형태 정보를, MBH descriptor 는 사람의 모션 정보를 담아낸다.

계산한 volume descriptor 들을 바탕으로, 시-공간 bag-of-features (BOF) 를 생성한다. 우선, 각 descriptor (HOG, MBH) 의 visual dictionary 를 만들기 위해 k-means 를 사용한다. Visual word 의 개수는 SIFT 를 위해 k_1 을 particle 를 위해서는 k_2 를 지정한다. 최종적으로 SIFT 와 particle 의 histogram 계산하고 통합하여 video descriptor 를 생성한다 (그림 2).

본 논문에서는 행동 분류를 위해 비선형 SVM 분류기를 사용한다. 그리고 SVM 을 학습하기 위해 multi-channel kernel [3] 을 사용한다.

$$K(H_i, H_j) = \exp(-\sum_{c \in C} \frac{1}{A_c} D_c(H_i, H_j)), \qquad (4)$$

$H_i=\{h_{ib}\}$ 와 $H_j=\{h_{jb}\}$ 는 비디오 샘플 x_i 와 x_j 의 histogram 을 의미한다. 파라미터 A_c 는 channel c 번째의 모든 훈련 집합 데이터 사이의 평균 거리 값을 의미하고, 거리 $D_c(H_i, H_j)$ 는 다음과 같이 정의된다.

$$D_c(H_i, H_j) = \frac{1}{2}\sum_{b=1}^{V} \frac{(h_{ib} - h_{jb})^2}{h_{ib} + h_{jb}}, \qquad (5)$$

$V = k_1 + k_2$ 는 vocabulary 크기를 의미한다.

3. 실험 결과

제안한 행동 인식 시스템의 성능을 평가하기 위해 TV human interaction [8] 데이터 셋을 사용하였다. 먼저,

(a) High five (b) Hug

(c) Kiss (d) Hand shake

(그림 3) TV human interaction의 샘플 영상

이 데이터 셋에서 성능을 평가하기 위해 사용된 파라미터 값을 기술한다. WSAP 특징 궤적을 추출하기 위해 Z=64, W=5, L=15 그리고 N=32 [4] 를 사용한다. 또한, visual dictionary 를 구성하기 위해 k_1=300 과 k_2=4000 을 사용하였다.

TV human interaction 는 네 가지 사람행동 클래스를 가지고 있다: hand shake (악수), high five (손 마주침), hug (껴안음) 그리고 kiss (입맞춤) 이다 (그림 3 참조). 이 데이터 셋의 각 행동 클래스는 TV 시트콤 영상에서 캡쳐한 짧은 영상 50 개들을 저장하고 있다. 추출된 영상들은 부분적 가려짐, 급격한 카메라 시점 변화, 카메라 각도, 복잡한 배경 등의 환경적 제약사항이 많은 것이 특징이다. 이 데이터 셋에서 제안한 행동 인식 시스템을 평가하기 위해 [8] 에서 사용한 실험 protocol 을 사용하였고, 성능 평가의 척도로는 average precision (AP) 를 사용하였다.

표 1 에서는 제안된 알고리즘과 최근 행동인식 기법들과의 성능을 비교한 것을 보여준다. 제안한 방법이 최근 다른 기법들보다 우수함을 확인할 수 있었다. 특히, 부분적으로 Hand shake, High five, Hug 클래스들이 기존 기법 보다 좋은 인식 결과를 보여주었다.

<표 1> 최근 기술들과 성능 비교 평가

	Patron-Perez(ID)[8]	Patron-Perez(SL)[8]	Cho [7]	제안한 방법
Hand shake	0.48	0.44	0.36	**0.485**
High five	0.32	0.33	0.52	**0.607**
Hug	0.42	0.44	0.56	**0.619**
Kiss	0.32	0.36	0.68	0.50
Mean AP	0.38	0.39	0.53	**0.553**

4. 결론

본 논문에서는 SIFT 와 particle 특징 궤적들을 기반한 새로운 행동 인식 시스템을 제안하였다. SIFT 특징 궤적 주변으로 particle 궤적 특징을 추출함으로 SIFT 궤적의 적은 수와 그 궤적이 종종 소실되는 단점을 보완하였다. 제안한 특징 궤적들의 주위로 모션 및 형태 정보를 추출하고, 이 정보를 바탕으로 SVM 분류기를 통해 행동 클래스들을 분류하였다. 본 논문의 실험을 통해 제안한 행동인식 시스템의 효율성을 증명하였다.

참고문헌

[1] D. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” Int. J. Comput. Vis., vol. 60, no. 2, pp. 91-110, 2004.

[2] J. Shi and C. Tomasi, “Good features to track,” in Proc. CVPR, pp.593-600, 1994.

[3] J. Sun, X. Wu, S. Yan, L. Cheong, T. Chua, and J. Li, “Hierarchical spatio-temporal context modeling for action recognition,” in Proc. CVPR, pp. 2004-2011, 2009

[4] H. Wang, A. Kläser, C. Schmid, and L. Cheng-Lin, “Action recognition by dense trajectories,” in Proc. CVPR, pp. 3169-3176, 2011.

[5] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in Proc. CVPR, pp. 886-893, 2005.

[6] B. D. Lucas and T. Kanade, “An iterative image registration technique with an application to stereo vision,” in Proc. IJCAI, 1981.

[7] S. Cho and H. Byun, “Human activity recognition using overlapping multi-feature descriptor,” Electronics Letters, vol. 47, no. 23, pp. 1275-1276, 2011.

[8] Patron-Perez A., Marszalek M., Zisserman A., and Reid I, “High five: recognising human interactions in TV shows” in Proc. BMVC, pp.1-11, 2010.

[9] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, “Learning realistic human actions from movies,” in Proc. CVPR, pp. 1-8, 2008.

[10] I. Laptev, “On Space-Time Interest Points,” Int. J. Comput. Vis., vol. 64, no. 2, pp. 107-123, 2005.

[11] Messing, R., Pal, C., Kautz, H.: Activity recognition using the velocity histories of tracked keypoints. ICCV (2009).