Performing Human Action Recognition with Optical-Flow, Back-of-Words and SVM

Alvaro Rojas Felipe Moreno Israel Chaparro

***Abstract*— Current years, Human activity recognition has an important impact research based on identify fights, violence, stole or any penable action trough cameras. Among the mapping algorithms of environments highlights the ORB-SLAM algorithm, which allows to obtain information from an unknown environment by generating a point cloud with a high precision in the location of information. How to use a SLAM algorithm allows both a map and a highly accurate location, using Homography-Matrix models fundamental for the initialization of a map that will be updated with the key frames. For the present project, the original version was modified and so that it can be executed in real time with any web camera, which implies a calibration process to readjust the location of the points in each image.**

I. INTRODUCTION

The algorithms of reconstruction of surroundings they emphasize 2 groups according to the type of result that they return. The first group consists of the so-called direct methods, which are those algorithms that return a massive amount of points called dense maps.

While the second is composed of algorithms based on feature detection, which return a cloud of points with less information but located more precisely.

Among this last group, the ORB-SLAM algorithm is usually the best known due to the accuracy of its results and the robustness of the algorithm.

A. PIPELINE:

- Extract Optical Flow.
- Make Dataset.
- Run K-means.
- Make Bag of Words.
- Train SVM.
- Test data on SVM.
- Get results.
- Vdeo obtenido con la

B. OPTICAL FLOW:

Given a current frame and its previous frame, we can compute its optical flow feature using the built-in dense optical flow Gunnar Farneback's algorithm of OpenCV. Thus, given a video with N frames, we can compute a set of N-1 optical flow feature descriptors.
As the videos' resolution are 160x120 and we want to save memory, we only sample the optical flow values on the rows and columns whose indices are multiples of 10 (i.e. row and columnn 0, 10, 20, ...). The optical flow descriptor for a frame will have size 16x12x2 = 384 (2 comes from the horizontal and vertical direction).

C. MAKE DATASET:

Split the computed optical flow figures using an identifier of the person in the record generating train and test sets. This generated train-keypoints.p (features).
TRAIN-PEOPLE-ID=[11,12,13,14,15,16,17,18,19,20,21,23,24,25,1,4] = 383 videos
TEST-PEOPLE-ID=[22, 2, 3, 5, 6, 7, 8, 9, 10] = 216 videos

D. RUNNING K-MEANS:

Run K-means on "train-keypoints.p" (features) with 200 as the number of clusters and produce the codebook.

II. BUILD BAG OF WORDS:

Make Bag-Of-Words for every video (not frame) in the training and test set, using the computed clusters (K-means) using Term frequency Inverse document frequency (TF-IDF) weighting scheme.
A BoW vector is like a histogram that counts the frequency of optical flow descriptors which appear in a video.
TF-IDF: Term frequency Inverse document frequency.

A. TRAIN SVM:

We then train a linear SVM classifier on BoW vectors of training set. As the number of videos in our training set is only 383, the training process takes less than a second, which is a lot faster than the method of considering each individual frame as an instance.
Parameters: C=1, kernel=linear

B. EVALUATE SVM:

Use computed SVM classifier to classify videos in train and test set and get the accuracy result.

	BOX	CLAP	WAVE	JOG	RUN	WALK
BOX	32	5	3	0	0	0
CLAP	2	31	4	0	0	1
WAVE	0	0	29	0	0	0
JOG	0	0	0	22	5	7
RUN	1	0	0	9	29	0
WALK	1	0	0	5	2	28
ACCU(%)	88.8	86.1	80.5	61.1	80.5	77.7

III. RESULTS:

171/216 Correct Videos in Test.
Accuracy = 0.79166

Confusion Matrix:

REFERENCES

[1] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, MonoSLAM: Real-time single camera SLAM, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 6, pp. 10521067, 2007.

[2] Raúl, Montiel, J. M. M. and Tardós, Juan D., ORB-SLAM: A Versatile and Accurate Monocular SLAM System, Mur-Artal, IEEE Transactions on Robotics, vol. 31, no. 5, pp. 1147-1163, 2015.

[3] J. Engel, T. Schops, and D. Cremers, LSD-SLAM: Large-scale direct monocular SLAM, in European Conference on Computer Vision (ECCV), Zurich, Switzerland, September 2014, pp. 834849.

[4] J. Engel, J. Sturm, D. Cremers, Semi-Dense Visual Odometry for a Monocular Camera, ICCV '13.

[5] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, ORB: an efficient alternative to SIFT or SURF, in IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, November 2011, pp. 2564 2571.

[6] Edward Rosten and Tom Drummond, Machine learning for high-speed corner detection, Department of Engineering, Cambridge University, UK.

[7] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua, BRIEF: Binary Robust Independent Elementary Features, CVLab, EPFL, Lausanne, Switzerland.

[8] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, Bundle adjustment a modern synthesis, in Vision algorithms: theory and practice, 2000, pp. 298372.

[9] Dorian Glvez-Lpez and Juan D. Tards, Bags of Binary Words for Fast Place Recognition in Image Sequences, IEEE Transactions on Robotics, vol. 28, no. 5, pp. 1188-1197, 2012.