Histogram of Oriented Depth Gradients for Action Recognition

Nachwa ABOU BAKR[1] James CROWLEY[2]

Univ. Grenoble Alpes, INRIA, Grenoble INP, CNRS, Laboratoire LIG
[1] nachwa.aboubakr@inria.fr
[2] james.crowley@inria.fr

Résumé

Cet article présente des expériences menées sur la perception de la profondeur de mouvements dans le cadre de la reconnaissance visuelle d'actions à partir de mesures locales extraites de séquences vidéo RGBD encodées selon la norme MPEG. Nous démontrons que ces mesures peuvent être combinées avec des descripteurs RGB spatio-temporels locaux pour l'élaboration d'une méthode de calcul efficace pour la reconnaissance d'actions. Des vecteurs de Fisher sont utilisés pour encoder et concaténer un descripteur de profondeur avec les descripteurs existants RGB locaux. Ces vecteurs sont ensuite utilisés pour la reconnaissance d'actions manuelles à l'aide d'un classifieur linéaire SVM. Nous comparons l'efficacité de ces mesures à l'état de l'art sur deux jeux de données récents pour la reconnaissance d'actions dans des environnements culinaires.

Mots Clef

Vision, Profondeur, Reconnaissance d'Actions, Codage MPEG.

Abstract

In this paper, we report on experiments with the use of local measures for depth motion for visual action recognition from MPEG encoded RGBD video sequences. We show that such measures can be combined with local space-time video descriptors for appearance to provide a computationally efficient method for recognition of actions. Fisher vectors are used for encoding and concatenating a depth descriptor with existing RGB local descriptors. We then employ a linear SVM for recognizing manipulation actions using such vectors. We evaluate the effectiveness of such measures by comparison to the state-of-the-art using two recent datasets for action recognition in kitchen environments.

Keywords

Vision, Depth, Action Recognition, MPEG encoding.

1 Introduction

Visual action recognition is the process of labelling image sequences with action labels. In this work we investigate the use of dense local descriptors for action recognition. Dense sampling has been shown to improve results when compared to sparse interest points for action recognition [3]. Requirements for reduced computational time have lead some authors to investigate methods that extract motion vectors directly from compressed video signals. For example, methods have been proposed to extract flow directly from MPEG encoded video [4]. The resulting method has been shown to provide useful description of Dense Trajectories (DT) [5], at the cost of some loss of recognition accuracy. This has led us to ask if this approach can be extended to provide fast computation of dense local descriptors from the depth channel of MPEG encoded RGBD image sequences.

In the following, we report the results on experiments in recognition of action labels using features extracted from depth channel of Kinect-like sensors using MPEG encoded motion flow. We describe a new descriptor based on oriented gradients of depth extracted from the motion channel of RGBD image sequences. This descriptor is used in addition to the MPEG Flow (MF) descriptors based on the RGB channels from the same sequence. Local video descriptors using Dense Trajectories [5] and MPEG Flow [4] are employed as baseline descriptors for comparison.

DT, as a video descriptor for action recognition, starts by sampling interest points densely at regular fixed grids. Detected interest points are then tracked over time and points for which no motion is detected are eliminated. This set of trajectories are then used as space-time interest points and descriptors are computed around those points. Four descriptors have been used in DT : Histogram of Oriented Gaussian (HOG), Histogram of Optical Flow (HOF), Motion Boundary Histogram in x,y directions (MBH_x, MBH_y). An analysis on computational requirements of DT shows that 61% of the running time is spent on the computation of optical flows. Reducing this computational burden is an interesting challenge.

The MF method exploits the estimated motion vectors used for video compression and uses them as interest points. This change provides a significant improvement in the time required for extraction of features at the cost of a small decrease in recognition accuracy [4].

Both DT and MF are video descriptors computed around motion trajectories. The primary differences between these two methods are that the interest points to be described in MF are at sparse positions of motion fields. In DT, points are sampled at the positions of a regular fixed grid, and they are tracked over time to build trajectories. Motion is encoded in DT using Gunnar-Farnebäck optical flows while MF uses estimated motion fields during video compression. Our goal is to determine if it is possible to improve MF results by exploiting data from depth channel of Kinect-like sensors.

2 Histogram of Oriented Depth Gradients

Depth can capture the structural features of objects. In addition to encoding the position of objects in space. Thus it is

reasonable to ask if we can improve the results of MF video descriptor by including the depth channel to better describe the scene. To do so, we compute gradients of depth which encodes the structure of the depth image. As in MF, interest points are a sparse set of points extracted while compressing video. This ensures describing area around motion points.
Practically, a volume of $32 \times 32 \times 15$ is computed around interest points. This volume is divided by a grid of 2 which produces a grid cell of 16×5. For each grid cell, we extract the gradients by applying the Sobel filter $[1, 0, -1]$ in a 16×16 pixels. Then, an 8-orientation bins discretized histogram is built to encode the gradient orientations. Gradient magnitudes are used to weight the gradient orientations. For each temporal slice the gradient is normalized using $l2$-normalization.
After that descriptors are encoded by transforming the collection of local video descriptors $\{x_i, \ldots, x_N\}$ into a fixed-size vector representation. Following [4], we chose to use Fisher Vectors (FV) [6] for descriptor encoding. In particular, only first-order differences of FV are used. Action categories are then classified by a linear one-vs-rest SVM classifier.

3 Experiments

Experiments are performed on two publicly available datasets : 50 salads [1] and Actions of Cooking Eggs (ACE) [2]. For 50 salads dataset; we used provided annotations of *core* actions only as verbs without taking into account their corresponding objects; in total, we used 6 different action labels. ACE also provides 6 different annotated actions but only one level of granularity which includes pre- and post movements of performed actions. For the evaluation, we measured the mean of Average Precision (mAP).
We consider two image sequences; one provides an RGB view of the scene and the other encodes scene depth. To study the impact of motion depth image sequences in describing actions, we extracted a Histogram of Oriented Depth Gradients (HODG) from depth image sequences. We extracted as well RGB descriptors of DT [1] and MF [2]. Combining both descriptors improves the results of MF method on two datasets as shown in Table 1 where RGB+HODG is a concatenation of encoded descriptors by FV. We compare the results to DT since MF follows the same structure and set of descriptors.
DT approach achieved the best results in this experiment at the cost of the computation time. In the other hand, adding HODG to MF descriptors achieves better results than [4] while preserving the computational cost [3]. Table 2 shows the number of processed frames per second while computing both RGB-based and HODG using the implementation of DT and MF. From the tables, we can also see that HODG alone performs relatively good in action recognition.

4 Conclusion

In this work, we reported an improvement of the results of MF method by using the benefit of depth channel in Kinect-like sensors. Since depth channels provide separate images that can be treated in parallel with no additional cost or sequentially while preserving a close to real-time capability. The idea is to combine the gradients of depth image sequences (HODG) extracted at sparse motion fields which are estimated in video compression to the set of RGB descriptors of MF [4]. Both RGB and Depth descriptors are encoded using FV.
We found that trajectories of depth image sequences, described by HODG, can describe manipulation actions with an acceptable recognition precision. To this end, we evaluated our proposed method on two recent datasets; 50 salads, and ACE. We recorded the computational time during extraction of descriptors using both methods. Results show a significant improvement in recognition precision with almost no significant additional cost.

	50salads	ACE
DT(RGB)	**99.60**%	94.12%
MF(RGB)	90.56%	92.04%
MF(HODG)	86.29%	89.94%
MF(RGB+HODG)	94.73%	**99.15**%

TABLE 1 – The results of our experiments using mAP.

	DT (fps)	MF (fps)
RGB (HOG, HOF, MBH)	3.87	27.78
Depth (HODG)	6.41	111.13

TABLE 2 – The number of processed frames per second while descriptors extraction for 50 salads dataset.

1. https://lear.inrialpes.fr/people/wang/dense_trajectories
2. https://github.com/vadimkantorov/cvpr2014
3. Each channel of the sensor can be treated individually in the descriptor extraction process

Références

[1] Stein, S. and McKenna, S.J. "Combining embedded accelerometers with computer vision for recognizing food preparation activities." Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing. ACM, 2013.

[2] Shimada, A., Kondo, K., Deguchi, D., Morin, G. and Stern, H. "Kitchen scene context based gesture recognition : A contest in ICPR2012". Advances in depth image analysis and applications. Springer Berlin Heidelberg, 2013. 168-185.

[3] Wang, H., Ullah, M.M., Klaser, A., Laptev, I. and Schmid. "Evaluation of local spatio-temporal features for action recognition". BMVC 2009-British Machine Vision Conference. BMVA Press, 2009.

[4] Kantorov, V. and Laptev, I. "Efficient feature extraction, encoding and classification for action recognition". In Proc. Computer Vision and Pattern Recognition (CVPR), IEEE, 2014.

[5] Wang, H., Kläser, A., Schmid, C. and Liu, C.L. "Action recognition by dense trajectories". Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011.

[6] Sánchez, J., Perronnin, F., Mensink, T. and Verbeek, J. "Image classification with the fisher vector : Theory and practice". International journal of computer vision 105.3 (2013) : 222-245.

[7] Pedregosa, Fabian, et al. "Scikit-learn : Machine learning in Python." Journal of Machine Learning Research 12.Oct (2011) : 2825-2830.