Multi-view action recognition using bag-of-words and various fusion schemes

Gertjan Burghouts, Pieter T. Eendebak, Henri Bouma, Johan-Martijn ten Hove
TNO
Oude Waalsdorperweg 63, 2597 AK The Hague, The Netherlands
gertjan.burghouts@tno.nl

Abstract

In this paper, we summarize how the action recognition can be improved when multiple views are available. The novelty is that we explore various combination schemes within the robust and simple bag-of-words (BoW) framework, from early fusion of features to late fusion of multiple classifiers. In new experiments on the publicly available IXMAS dataset, we learn that action recognition can be improved significantly already by only adding one viewpoint. We demonstrate that the state-of-the-art on this dataset can be improved by 5% – achieving 96.4% accuracy – when multiple views are combined.

1. Introduction

Recognizing human actions is a critical aspect of many types of surveillance. Human actions can be indicative of a wide range of unwanted situations such as aggression, vandalism, theft, somebody falling, and becoming unwell. An obvious way to improve the action recognition accuracy is to increase the number of viewpoints on the action. This paper compares the action recognition accuracy when multiple 2D views are combined at different abstraction levels. It is a summary of our IEEE AVSS paper [2].

The bag-of-words (BoW) model is our basic pipeline [1, 3]. In this framework, fusion of multiple 2D views can be performed on the low (feature), intermediate (representation) or high (classifier) level. We evaluate the performance of these fusion schemes and compare them against the state-of-the-art on the multi-view IXMAS dataset of 12 human actions recorded by 5 cameras (4 side and 1 top view) [4]. Previous experiments on this dataset have shown the merit of adding additional cameras [6, 7]. In these works, there was no particular rationale for selecting the subsets of cameras. For surveillance systems, a key design issue is to properly select the number of cameras and place them such that the distinctive details about humans and their actions are visible. In our experiments, we assess systematically the merit of each camera by analyzing the performance of pairs of cameras under angles of 45, 90 and 135 degrees. Our experiments lead to insights on appropriate camera setups for human action recognition.

The paper is organized as follows. In Section 2 we propose the seven fusion schemes to combine the multiple 2D views. Section 3 is about the experiments and results, where we establish the performance of the combiners, compare to state-of-the-art, and assess the merit of multiple cameras and their viewing angles. Finally, Section 4 concludes with the main findings.

2. Multi-view fusion schemes

We used a bag-of-word (BoW) pipeline consisting of STIP features to capture motion, a random forest (RF) to quantize the features into histograms, and an SVM classifier that serves as action detector [2]. In the BoW framework, the 2D views can be combined in several ways. The STIP features can be collected from all cameras, and transformed into a single histogram, i.e. early fusion (feature level). Intermediate fusion combines RF-histograms and it can be performed in two ways. The first is to collect the histograms to obtain more samples and the second is to concatenate the histograms and obtaining longer samples. We distinguish four types of late fusion, which combines the posterior output of SVMs. The first type selects the action with the highest posterior after averaging over the views. The second selects the action that has the maximum posterior without averaging. The third learns the optimal view per action on the train set. And the fourth is an adaptation from [1], where a second-stage SVM is introduced that takes all posteriors as input values and trains a mapping from those values.

3. Experiments and results

The experiments are performed on the four side-view cameras of the IXMAS dataset with leave-one-actor-out cross validation. This dataset consists of 12 actions with each action executed three times by 12 subjects [4].

3.1. Performance of fusion schemes

The best performing multi-view combination schemes are the posterior average voting and RF histogram concatenation (Table 1). These methods perform significantly better than the combination of all STIP features. We hypothesize that each view has a distinct influence on the way that the human and action are perceived and that it is easier to generalize views at a higher level. Inspection of the confusion matrix [2] showed that the actions that are confused are very similar, e.g. cross arms and check watch.

Fusion	Combiner	Acc.
None	N/A	87.9
Early	STIPs	88.6
Interm.	set of RF hist	81.4
Interm.	concat. of RF hist	95.3
Late	av. vote	96.4
Late	max. vote	92.2
Late	best view	90.0
Late	2nd stage SVM	94.4

Table 1. Performance of the 2D view combiners compared to average result over the single-view setup per camera.

3.2. Comparison to state-of-the-art

Our best fusion methods are the RF histogram concatenation and posterior average voting. They outperform significantly the state-of-the-art method AFMKL [6] and latent kernelized SVM [5] on the four 2D side views of IXMAS, by a relative improvement of 8.2% and 4.7% resp. (Table 2).

Method	Acc.
AFMKL [6]	88.2
Latent kernel SVM [5]	91.7
Our intermediate RF concat.	95.3
Our late posterior av. voting	96.4

Table 2. Comparison of our best 2D view combiner to the best methods on IXMAS that use multiple 2D views.

3.3. Varying number of views

We are interested in how the performance of our best fusion method improves when cameras are added (Table 3). Note that the average accuracies were reported over all combinations of 4 separate cameras, 6 camera pairs and 4 camera triplets. Adding the second camera improves the action recognition for RF histogram concatenation accuracy most, a relative improvement of 6.3%, where the third and fourth camera add on average only 1.7%. Note that the standard deviation over the different pairs and triples is much smaller for histogram concatenation than for average voting.

#Cams	Cameras	Acc: interm. RF concat.	Acc: late av. vote
1	4 x separate	87.9	87.9
2	6 x pairs	92.5 ± 0.5	91.5 ± 2.4
3	4 x triplets	94.4 ± 0.3	91.3 ± 3.0
4	4 x all	95.3	96.4

Table 3. Average accuracies of action recognition. The results show an improvement when multiple 2D views are used. RF histogram concatenation and late average voting were used for fusion.

More experiments and details are provided in the AVSS paper [2].

4. Conclusions

We have investigated how much action recognition can be improved when multiple views are available. Within the bag-of-words (BoW) framework, we considered various schemes to combine camera viewpoints. Early fusion combines the STIP features, intermediate-level fusion combines BoW histograms, and late fusion combines the output of classifiers. Our experiments on the IXMAS dataset showed that action recognition can be improved significantly already by only adding one viewpoint. Furthermore, we demonstrated that the state-of-the-art on the four side-views of this dataset can be improved by 5% – achieving 96.4% accuracy – when multiple views are combined by intermediate-level fusion of the BoW representations.

Acknowledgement

This work was supported by the project Passive Sensors.

References

[1] H. Bouma, G. J. Burghouts, L. de Penning, et al. Recognition and localization of relevant human behavior in video. In *SPIE*, volume 8711, 2013.

[2] G. J. Burghouts, P. T. Eendebak, H. Bouma, and J. M. ten Hove. Improving action recognition by combining multiple 2D views in the bag-of-words model. In *IEEE Advanced Video and Signal-Based Surveillance AVSS*, pages 250–255, 2013.

[3] G. J. Burghouts, K. Schutte, R. den Hollander, and H. Bouma. Selection of negative samples and two-stage combination of multiple features for action detection in thousands of videos. *Machine Vision and Applications MVA*, 25:85–98, 2014.

[4] D. Weinland, E. Boyer, and R. Ronfard. Action recognition from arbitrary views using 3D exemplars. In *ICCV*, 2007.

[5] X. Wu and Y. Jia. View-invariant action recognition using latent kernelized structural SVM. In *ECCV*, 2012.

[6] X. Wu, D. Xu, L. Duan, and J. Luo. Action recognition using context and appearance distribution features. In *CVPR*, 2011.

[7] F. Zhu, L. Shao, and M. Lin. Multi-view action recognition using local similarity random forests and sensor fusion. *Pattern recognition letters*, 24(1):20–24, 2013.