Action Recognition by Layout, Selective Sampling and Soft-Assignment

G.J. Burghouts, P. Eendebak, H. Bouma, R.J-M. ten Hove
TNO Intelligent Imaging
The Hague, The Netherlands
gertjan.burghouts@tno.nl

Abstract

This paper describes the TNO action recognition system that was used to generate the results for our submission to the competition track of the THUMOS '13 challenge at ICCV '13. This system deploys only the STIP features that were provided on the website of the challenge. A bag-of-features model is extended with three novelties.

1. Feature: STIP-only

As low-level features, we only use the STIP features that were provided on the website of the UCF-101 action dataset.

2. Three Novelties for the Bag-of-Features Model

We deploy a bag-of-features model to represent each clip, where the quantization is performed by a random forest, see implementation details in [1]. We include three novelties in the representation: (a) we encode the spatio-temporal layout of the features [2], (b) the random forest is trained by selectively sampling the negatives that are most similar [3], and (c) the features are assigned to the leafs of the random forest by a soft-assignment procedure to deal with their uncertainty [4]. Parameters are optimized using the leave-one-group-out cross-validation setup as provided with the UCF-101 dataset.

3. Action Prediction

For each test clip, for each of the 101 actions, a posterior is produced by an SVM with an exponential Chi-2 kernel. The feature quantization, representation and SVM classifier model are packed as a pipeline. We create many of such pipelines, by selecting random subsets of the training set to learn the parameters, obtaining multiple posteriors per action. Finally, the reported action label and posteriors for each clip are obtained from averaging the multiple posteriors for each action, which makes the prediction more robust [5].

Acknowledgements

This work is supported by DARPA (Mind's Eye program). The content of the information does not necessarily reflect the position or the policy of the US Government, and no official endorsement should be inferred.

References

[1] H. Bouma, G.J. Burghouts, L. de Penning, P. Hanckmann, J.-M. ten Hove, S. Korzec, M. Kruithof, S. Landsmeer, C. van Leeuwen, S.P. van den Broek, A. Halma, R.J. den Hollander, K. Schutte, Recognition and localization of relevant human behavior in videos, SPIE, 2013.

[2] G.J. Burghouts, K. Schutte, Spatio-Temporal Layout of Human Actions for Improved Bag-of-Words Action Detection, Pattern Recognition Letters, 2013.

[3] G.J. Burghouts, K. Schutte, Correlations Between 48 Human Actions Improve Their Detection, International Conference on Pattern Recognition, 2012.

[4] G.J. Burghouts, Soft-Assignment Random-Forest with an Application to Discriminative Representation of Human Actions in Videos, International Journal of Pattern Recognition and Artificial Intelligence, 2013.

[5] G.J. Burghouts, K. Schutte, H. Bouma, R.J.M. den Hollander, Selection of Negative Samples and Two-Stage Combination of Multiple Features for Action Detection in Thousands of Videos, Machine Vision and Applications, 2013.