Classification of human activities in uncontrolled environments

Ngoc Nguyen, Atsuo Yoshitaka
Japan Advanced Institute of Science and Technology

Abstract : Human activity recognition is an important field of computer vision research today. It has grown dramatically in the past 5 years. Recently, most work in activity recognition has concentrated on classifying simple activities, e.g. walking, jogging in controlled environments in which only a single person appears in videos taken with a simple and static background. In this paper, we focus on complex activities such as human-human interactions in realistic environments. Motivated from the well-known Independent Subspace Analysis (ISA) and deep learning networks, we introduce a three-layer ISA convolutional network to learn hierarchical invariant features. The obtained invariant features are fed into a standard bag-of-features model to recognize human interactions. Experimental results on the Hollywood2 dataset show that our approach is able to learn features which are effective to represent complex activities.

Keywords: human-human interactions, classification, convolutional network

1. Learning hierarchical invariant features

Independent subspace analysis algorithm is very well-known in natural image statistics, and it can learn features which are robust to local translation while being selective to frequency, rotation and velocity. However, the ISA algorithm becomes slow when the dimension of input data is large. In order to learn high-level concepts and solve the computational problem of the ISA when trained on video data, we combine the convolutional technique with the standard ISA algorithm to design a three-layer convolutional ISA network. This convolutional network uses PCA and ISA as sub-units.

In particular, we extract video blocks of size $w_1 \times h_1$ (spatial dimensions) and t_1 (temporal dimension) in the first layer. We preprocess these blocks by removing DC component and applying PCA to whiten. These blocks are fed into the first layer, which we call ISA1. The output of this layer is the weights W_1 and the subspace structureV_1.

Similarly, in the second layer, we extract video blocks of size $w_2 \times h_2 \times t_2$. In order to find hierarchical features, we set the dimensions of the blocks in the second layer larger than the ones in the first layer. Each block in the second layer can be seen as a collection of m overlapping blocks of size $w_1 \times h_1 \times t_1$. These small video blocks are convolved with the weights W_1, V_1 learned from the ISA1. The responses f_1, f_2 ,..., f_m are concatenated to form the inputs to train ISA weights in the second layer. The same procedure is repeated in the third layer.

3. Classification model

We apply χ^2 support vector machine (SVM) to classify human interactions. Because it is multi-class classification, we apply the one-against-rest approach and select the class with the highest score.

4. Experimental results

We evaluate our approach on the Hollywood2 dataset, which is a comprehensive benchmark for human activity recognition in realistic and challenging settings. It has 12 classes of human activities including answering, phone, driving car, eating, fighting, getting out of a car, hand shaking, hugging, and so on, which are collected from 69 Hollywood movies. In our experiments, we used the clean training dataset, which has 823 training examples while the test set has 884 samples. Our method achieves mAP of 53.7% on this dataset as shown in Table 1.

Table 1. Performance comparison on the Hollywood2 dataset

Method	mAP
Our method	53.7%
Le et al. [1]	53.3%
Sun et al. [2]	48.1%

4. Conclusion

This paper has presented a three-layer convolutional ISA network which is able to learn hierarchical invariant features. The encouraging results on the Hollywood2 dataset show that our method these features are effective to represent complex activities in realistic environments.

References

[1] Q.V. Le, W.Y. Zou, S.Y. Yeung, and A.Y. Schmid, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 3361-3368, 2011.

[2] L. Sun, K. Jia, T.H. Chan, Y. Fang, G. Wang, and S. Yan "DF-SFA: Deeply-Learned Slow Feature Analysis for Action Recognition," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 2625-2632, 2014.