Int J Comput Vis (2012) 100:1–15
DOI 10.1007/s11263-012-0534-7

Sparse Modeling of Human Actions from Motion Imagery

Alexey Castrodad · Guillermo Sapiro

Received: 31 July 2011 / Accepted: 27 April 2012 / Published online: 6 June 2012

Abstract An efficient sparse modeling pipeline for the classification of human actions from video is here developed. Spatio-temporal features that characterize local changes in the image are first extracted. This is followed by the learning of a class-structured dictionary encoding the individual actions of interest. Classification is then based on reconstruction, where the label assigned to each video comes from the optimal sparse linear combination of the learned basis vectors (action primitives) representing the actions. A low computational cost deep-layer model learning the inter-class correlations of the data is added for increasing discriminative power. In spite of its simplicity and low computational cost, the method outperforms previously reported results for virtually all standard datasets.

Keywords Action classification · Sparse modeling · Dictionary learning · Supervised learning

1 Introduction

We are living in an era where the ratio of data acquisition over exploitation capabilities has dramatically exploded. With this comes an essential need for automatic and semi-automatic tools that could aid with the processing requirements in most technology-oriented fields. A clear example pertains to the surveillance field, where video feeds from possibly thousands of cameras need to be analyzed by a limited amount of operators on a given time lapse. As simple as it seems for us to recognize human actions, it is still not well understood how the processes in our visual system give our ability to interpret these actions, and consequently is difficult to effectively emulate these through computational approaches. In addition to the intrinsic large variability for the same type of actions, factors like noise, camera motion and jitter, highly dynamic backgrounds, and scale variations, increase the complexity of the scene, therefore having a negative impact in the performance of the classification system. In this paper, we focus in a practical design of such a system, that is, an algorithm for supervised classification of human actions in motion imagery.

There are a number of important aspects of human actions and motion imagery in general that make the particular task of action classification very challenging:

1. Data is very high dimensional and redundant: Each video will be subdivided into spatio-temporal patches which are then vectorized, yielding high-dimensional data samples. Redundancy occurs from the high temporal sampling rate, allowing relatively smooth frame-to-frame transitions, hence the ability to observe the same object many times (not considering shot boundaries). In addition, many (but not all) of the actions have an associated periodicity of movements. Even if there is no periodicity associated with the movements, the availability of training data implies that the action of interest will be observed redundantly, since overlapping patches characterizing a specific spatio-temporal behavior are generally very similar, and will be accounted multiple times with relatively low variation. These properties of the data allow the model to benefit from the *blessings* of high dimensionality (Donoho 2000), and will be key to over-

Alexey Castrodad is also with NGA.

A. Castrodad (✉) · G. Sapiro
Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455, USA
e-mail: castr103@umn.edu

G. Sapiro
e-mail: guille@umn.edu