Comparing Evaluation Protocols on the KTH Dataset

Zan Gao¹, Ming-yu Chen², Alexander G. Hauptmann², and Anni Cai¹

¹ School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, P.R. China
² School of Computer Science, Carnegie Mellon University, 15213, PA, USA
zangaonsh4522@gmail.com, {mychen,alex}@cs.cmu.edu, annicai@bupt.edu.cn

Abstract. Human action recognition has become a hot research topic, and a lot of algorithms have been proposed. Most of researchers evaluated their performances on the KTH dataset, but there is no unified standard how to evaluate algorithms on this dataset. Different researchers have employed different test setups, so the comparison is not accurate, fair or complete. In order to know how much difference there is when different experimental setups are used, we take our own spatio-temporal MoSIFT feature as an example to assess its performance on the KTH dataset using different test scenarios and different partitioning of the data. In all experiments, support vector machine (SVM) with a chi-square kernel is adopted. First, we evaluate performance changes resulting from differing vocabulary sizes of the codebook, and then decide on a suitable vocabulary size of codebook. Then, we train the models using different training dataset partitions, and test the performances one the corresponding held-out test sets. Experiments show that the best performance of MoSIFT can reach 96.33% on the KTH dataset. When different n-fold cross-validation methods are used, there can be up to 10.67% difference in the result. And when different dataset segmentations are used (such as KTH1 and KTH2), the difference in results can be up to 5.8% absolute. In addition, the performance changes dramatically when different scenarios are used in the training and test dataset. When training on KTH1 S1+S2+S3+S4 and testing on KTH1 S1 and S3 scenarios, the performance can reach 97.33% and 89.33% respectively. This paper shows how different test configurations can skew results, even on standard data set. The recommendation is to use a simple leave-one-out as the most easily replicable clear-cut partitioning.

Keywords: Action Recognition, training/test data sets, partitioning, experimental methods.

1 Introduction

The problem of data sets and partitioning the data sets has confronted every researcher. Some standard evaluation efforts, such as TRECVID, organized by the U.S. National Institute of Standards (NIST), provide enough data and specify a partitioning of the data into training and testing for all published experiments. While researchers still at times evaluation only subsets of the data that perhaps are most suitable to their approaches, the basic partitioning ensures at least a clear method to duplicate and
validate any experiments, as well as developing improved methods. The problem is more acute, when there is not enough data for reliable trainings. This is the case with the well-known Weizmann and KTH action data sets, where the data sets have been made public, but no standard has been agreed for evaluation.

The typical research approaches, discussed in more detail in section 2, try multiple folder cross-validation, where some data is used for training, and some held out for testing. In addition, each KTH video includes multiple repetitions of an action, and some papers treat adjacent scenes as separate instances in the data, using one for training and the other for testing, while others partition the data at the complete video file level. Similar choices exist with the same person performing an action in different camera settings, etc. While one might argue that this is unlikely to result in major different, this paper intends to show that the differences are quite significant and we argue for a uniform leave-one-file-out approach in these situations, which is easy to understand and replicate, but also provides a relatively large amount of training data in each iteration.

In general, action recognition has been widely researched and applied in many domains, such as visual surveillance, human computer interaction and video retrieval etc. Many schemes have been proposed for the human action recognition, and we give a brief overview over some of the more frequently cited approaches in the literature.


Jia and Yeung [29] use a dimensionality reduction approach called LSTDE to recognize silhouette-based human action. Gorelick et al. [28] regard human actions as