MindLAB at the THUMOS Challenge

Fabián Páez
MindLAB Research Group
Bogotá, Colombia
fmelpaezr@unal.edu.co

Jorge A. Vanegas
MindLAB Research Group
Bogotá, Colombia
javanegasr@unal.edu.co

Fabio A. González
MindLAB Research Group
Bogotá, Colombia
fagonzalezo@unal.edu.co

Abstract

In this notebook paper we describe the MindLAB research group participation at the THUMOS challenge held as part of the ICCV 2013 conference. Two runs were submitted using different methods (SVM, OMF) with the features provided by the challenge organizers (DTF). The performance obtained shows an improvement over the baseline method.

1. Introduction

Action recognition has received a lot of attention by the computer vision and machine learning community. This interest has contributed to the generation of different datasets targeted to this task [2]. Most of those datasets contain a limited number of actions and samples, and in some cases belong to unrealistic settings. The UCF101 dataset [6] aims to provide a large scale dataset for human action recognition. This poses new problems to efficiently process the amount of samples provided along with the high dimensionality of the used features.

2. Methods

Two methods were evaluated on the recognition task, Support Vector Machines (SVM) and Online Matrix Factorization (OMF).

2.1. SVM

2.1.1 Kernels

Kernels provide a mapping to a higher dimensional space in which patterns are easier to notice. The $\chi^2$ kernel [1] has been widely used on action recognition tasks [8], and the histogram intersection kernel [2] is an effective kernel used on texture classification [3] and other computer vision tasks.

\[
\phi_{\chi^2}(x, y) = \exp\left(\frac{-1}{2A} \sum_{i=1}^{D} \frac{(x_i - y_i)^2}{x_i + y_i}\right)
\]

\[
\phi_{HI}(x, y) = \sum_{i=1}^{D} \min(x_i, y_i)
\]

2.2. OMF

Online Matrix Factorization [1] aims to find a semantic space in which to map different data modalities using an online approach. Projection matrices for each modality are calculated based on training samples. Once the matrices have been found, they are used to map the visual features of the test samples to the semantic space and then, from this space to the action label space.

The input to the algorithm are two matrices made of the visual ($V$) and textual ($T$) representation of the training samples. The aim is to find the corresponding $P$ and $Q$ projection matrices, that map to the semantic space, to obtain a representation in this space $H$.

\[
V \approx PH
\]

\[
T \approx QH
\]

The objective function to simultaneously solve the factorization is

\[
\min_{P,Q,H} (1-\alpha)\|V-PH\|^2_F + \alpha\|T-QH\|^2_F + \lambda(\|P\|^2_F + \|Q\|^2_F + \|H\|^2_F)
\]

where $\lambda$ is a regularization parameter and $\alpha$ is a parameter determining the weight given to each modality. An implicit parameter is the number of latent factors, corresponding to the dimensionality of the $H$ matrix.

The online approach is developed in [1]. It applies stochastic gradient descent to solve the optimization problem. This introduces new parameters, such as the minibatch size (number of samples used on each iteration), gradient descent step size ($\gamma$), and the number of epochs.
3. Experiments

3.1. Features

The experiments were performed using the features provided by the challenge organizers [3]. Specifically, DTF \cite{7} features were used, which consist of a 4000 bin bag-of-features (BOF) histogram for each of HOG, HOF, MBH and trajectories descriptors. As the dimension of the provided features is high, we decided to use only the most descriptive feature to allow shorter runtimes. According to [7], MBH is the feature that in general gives the best results when used on its own.

We performed experiments to verify if MBH yield the best results for this dataset. These experiments are presented on section 3.2.

3.2. SVM

The aforementioned kernels were tested with the each descriptor (HOG, HOF, MBH and TR). First, 3-fold cross-validation was performed with the first training partition. Each fold was made following the recommendation to leave clips from the same group within the same partition. The folds were made each of 6 groups from the 18 groups belonging to the training partition.

Tables 1 and 2 show the results of 3-fold cross-validation on the first training partition for each feature and kernel. With these experiments, the effectiveness of the MBH descriptor was verified, and it was chosen as the feature to perform experiments on the test partitions. Also, the histogram intersection kernel showed a better performance and was therefore selected to carry the tests.

Having selected a feature for testing, 3-fold cross-validation was used again as described before, but this time on each training partition. The purpose was to select an appropriate value for the complexity of the SVM. The best performing parameter in validation was then used on the corresponding test partition.

For these experiments, \cite{4} was used, as it provides an SVM implementation, among other tools like cross-validation.

3.3. OMF

There are 6 parameters to tune on OMF, making an exhaustive parameter search unfeasible. Therefore, only a small subset of the parameters was explored and only for one of the training partitions, the first one. The best performing parameters on this partition were used for all the test partitions. For the parameter search within the first training partition, the same 3-fold cross-validation setup as described in section 3.2 was followed.

The data modalities for this dataset are the visual features and the corresponding action label. For training, both modalities are used to obtain the projection matrices. Testing only uses the visual modality, no label is used on this phase. Using only the visual modality, and the projection matrices, the corresponding test label can be predicted.

OMF allows to perform multilabeling, but as the aim here is to obtain only one label, the label with the greatest value for a given test sample is used as the action present in the video.

Table 3 shows the best performing parameters for OMF using 3-fold cross-validation. The corresponding accuracy for each fold is presented on Table 4.

4. Results

SVMs achieved a better performance than OMF\textsuperscript{1}. This can be attributed to the difference among the methods, as SVMs are discriminative while OMF is a generative method. As the task addressed here is classification, a discriminative method is expected to obtain better results.

\textsuperscript{1}Due to a bug in the OMF testing code, the official challenge results differ from the actual ones. After fixing the bug, 54.78% accuracy was obtained.
References


