Online Action Recognition via Nonparametric Incremental Learning

Rocco De Rosa
rocco.derosa@unimi.it
Department of Mathematics “Federigo Enriques”
Università degli Studi di Milano, Milano, Italy

Nicolò Cesa-Bianchi
nicolo.cesa-bianchi@unimi.it
Dipartimento di Informatica
Università degli Studi di Milano, Milano, Italy

Ilaria Gori
ilaria.gori@iit.it
iCub Facility
Istituto Italiano di Tecnologia, Genova, Italy

Fabio Cuzzolin
fabio.cuzzolin@brookes.ac.uk
Department of Computing and Communication Technologies
Oxford Brookes University, Oxford, UK

We introduce an *online action recognition system* that can be combined with any set of frame-by-frame feature descriptors. Our system covers the frame feature space with classifiers whose distribution adapts to the hardness of locally approximating the Bayes optimal classifier. An efficient nearest neighbour search is used to find and combine the local classifiers that are closest to the frames of a new video to be classified. The advantages of our approach are: *incremental training*, *frame by frame real-time prediction*, *nonparametric predictive modelling*, video segmentation for *continuous action recognition*, *no need to trim videos* to equal lengths and *only one tuning parameter* (which, for large datasets, can be safely set to the diameter of the feature space). Experiments on standard benchmarks (see Fig. 2 and Tab. 1) show that our system is competitive with state-of-the-art non-incremental and incremental baselines.

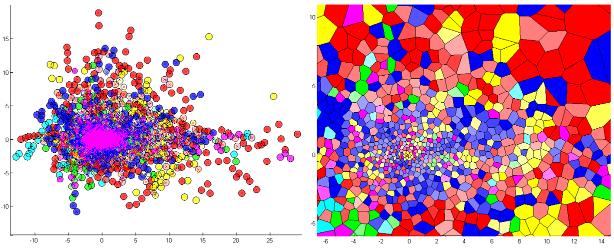
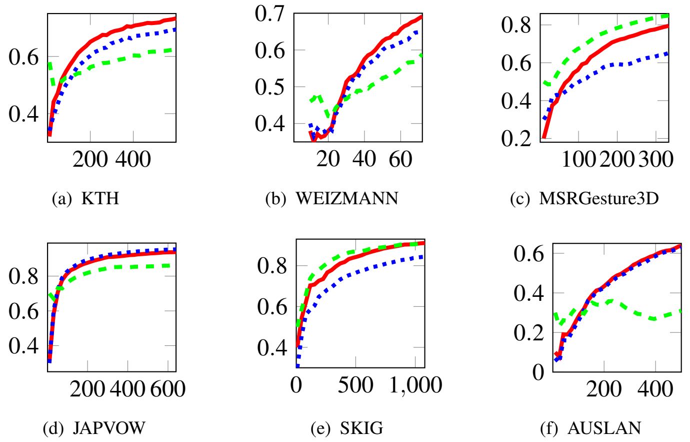


Figure 1: Left: the set of balls resulting from training on the first two principal components of local features extracted from the KTH dataset (colours denote labels, and color intensity expresses the ’purity’ of the conditional class distribution within each ball). Right: a close-up of the central area represented as the Voronoi tessellation associated with the balls shows how the regions whose class statistics are more complex are covered by a finer set of balls.

Algorithm 1 ABACOC (Adaptive Ball Cover for Classification)

```
Input: Initial radius $R > 0$, metric $\rho$
Initialize set of ball centers $\mathcal{S} = \emptyset$ and set of labels $\mathcal{Y} = \emptyset$
for $i = 1, 2, \dots$ do
    Receive labeled video $(V_i, y_i)$
    Create sequence of labeled frames $(x_1, y_i), \dots, (x_{T_i - 1}, y_i)$
    for $t = 1, \dots, T_i - 1$ do
        if $\mathcal{S} \equiv \emptyset$ then
            $\mathcal{S} = \{x_t\}$, set $\varepsilon_t = R$, and use $y_i$ to init. estimates $p_t$
        else
            Let $x_s \in \mathcal{S}$ be the nearest neighbour of $x_t$ in $\mathcal{S}$
            if $\rho(x_s, x_t) \le \varepsilon_s$ ($x_t$ belongs to current ball centered on $x_s$) then
                if $y_i \neq \operatorname{argmax}_{c \in \mathcal{Y}} p_s(c)$ then
                    Set $m_s = m_s + 1$ and update radius via $\varepsilon_s = R m_s^{-1/(2+d)}$
                end if
                Use $y_i$ to update estimates $p_s$
            else
                $\mathcal{S} = \mathcal{S} \cup \{x_t\}$, set $\varepsilon_t = R$, and use $y_i$ to init. estimates $p_t$
            end if
        end if
    end for
end for
```

The proposed method is a general framework for incremental *multivariate time series classification* (e.g. video frames) based on the following principles: (i) each video frame is a training example in a local feature space; (ii) incoming training examples are selected to cover the frame feature space with balls whose radius is adjusted according to the distribution of action classes within each ball; (iii) each ball is associated with an estimate of the conditional class probabilities, obtained by collecting statistics around its centre, which is used to make predictions on new unlabeled samples; (iv) the set of balls can be organized in a tree structure, allowing logarithmic queries in the number of balls. During training (see Alg. 1), a new ball is added whenever the input frame example does not belong to the ball whose center is the closest to the frame among the centers in the current set (Fig. 1, left). Otherwise, the ball statistics and its radius are updated. In the prediction phase, the conditional class probability estimates associated with the ball centre nearest to the input frames are used to select the action that maximises the sum of those scores (Fig. 1, right). The method allows us to work *incrementally* at frame level and *in real time*. Our learning method is also nonparametric. That is, the classifier structure is not pre-determined (as for linear classifiers), but it is inferred from the data (as for k-NN). As it handles videos on a frame-by-frame basis, the method is suitable to tackle the so-called “continuous action recognition” problem. To the best of our knowledge no other approach enjoys all these attractive features.



(a) KTH (b) WEIZMANN (c) MSRGesture3D

(d) JAPVOW (e) SKIG (f) AUSLAN

Figure 2: The plots show the online performance of ABACOC (red solid line) against SVM-b (green dashed line) and ALMA (blue dotted line). The x-axis is the number of videos fed to the algorithms and the y-axis is the average accuracy over the ten random permutations.

DATASET	**HMM-1NN**	**DTW-d**	**SVM-b**	**ABACOC**
KTH	68.28%	52.50%	69.83%	**83.20%**
Weizmann	87.50%	53.76%	97.22%	**98.61%**
SKIG	90.30%	95.74%	94.50%	**97.50%**
MSRGesture3D	78.20%	50.65%	**95.55%**	90.33%
JAPVOW	95.67%	69.72%	84.59%	**98.01%**
AUSLAN	67.07%	**83.81%**	44.78%	72.32%

Table 1: Multiclass accuracies of ABACOC compared against four baseline algorithms on the six benchmark datasets. All the methods share the same extracted features.