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.

Algorithm 1 ABACOC (Adaptive Ball Cover for Classification)

**Input:** Initial radius $R > 0$, metric $\rho$

1. Initialize set of ball centers $S = \emptyset$ and set of labels $\mathcal{Y} = \emptyset$
2. for $i = 1, 2, \ldots$ do
3.   Receive labeled video $(V_i, y_i)$
4.   Create sequence of labeled frames $(x_1, y_1), \ldots, (x_T, y_T)$
5.   for $t = 1, \ldots, T - 1$ do
6.     if $S = \emptyset$ then
7.       $S = \{x_t\}$, set $\epsilon_t = R$, and use $y_t$ to init. estimates $p_t$
8.     else
9.       Let $x_t \in S$ be the nearest neighbour of $x_t$ in $S$
10.      if $\rho(x_t, x_t) \leq \epsilon_t$ ($x_t$ belongs to current ball centered on $x_t$) then
11.         if $y_t \neq \text{argmax}_{c \in \mathcal{Y}} p_t(c)$ then
12.             Set $m_0 = m_0 + 1$ and update radius via $\epsilon_t = R m_0^{-1/(2+d)}$
13.         end if
14.      else
15.         Use $y_t$ to update estimates $p_t$
16.      end if
17.     end if
18.     $S = S \cup \{x_t\}$, set $\epsilon_t = R$, and use $y_t$ to init. estimates $p_t$
19.   end if
20. 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.