Latent Boosting for Action Recognition

Zhi Feng Huang[1]
zfh@sfu.ca
Weilong Yang[1]
wya16@sfu.ca
Yang Wang[2]
yangwang@uiuc.edu
Greg Mori[1]
mori@cs.sfu.ca

[1] School of Computing Science
Simon Fraser University,
British Columbia, Canada

[2] Department of Computer Science
University of Illinois at Urbana-Champaign,
Illinois, USA

Figure 1: Typical tracklets from the TRECVID dataset. The first row consists of two 5-frame tracklets of running people. The second row consists of two 5-frame tracklets of not-running people.

Consider the problem of recognizing an action such as *running* in a surveillance video. A typical approach involves first detecting and tracking people, followed by classification. However, accurate tracking is challenging. As illustrated in Fig. 1, trackers will suffer from jitter, especially when people are performing varied actions. Since accurate tracking is not a direct end-goal of action recognition, it is natural to consider tracking as a latent variable and train a model focused on action recognition.

In this paper we present LatentBoost, a novel learning algorithm for training models with latent variables in a boosting framework. This algorithm allows for training of structured latent variable models with boosting. The popular latent SVM [3] framework allows for training of models with structured latent variables in a max-margin framework. LatentBoost provides an analogous capability for boosting algorithms. The effectiveness of this framework is highlighted by an application to human action recognition. We show that LatentBoost can be used to train an action recognition model in which the trajectory of a person is a latent variable. This model outperforms GradientBoost [4] on a variety of datasets.

In LatentBoost, we assume that a training example $(\mathbf{x}, y)$ is associated with a set of latent variables $\mathbf{L} = \{l_1, l_2, \ldots, l_T\}$, where each latent variable takes its value from a discrete set, i.e. $l_t \in \mathscr{L}$[1]. We assume these latent variables are constrained by an undirected graph structure $\mathscr{G} = (\mathscr{V}, \mathscr{E})$, where $\mathscr{V}$ and $\mathscr{E}$ denote vertices and edges in the graph $\mathscr{G}$, respectively. For a fixed $\mathbf{L}$, the scoring function of the $(\mathbf{x}, \mathbf{L})$ pair for the k-th class can be written as the sum of a set of unary and pairwise potential functions:

$$F_k(\mathbf{x}, \mathbf{L}) = \sum_{t \in \mathscr{V}} H_k^t(\mathbf{x}, l_t) + \sum_{(t,s) \in \mathscr{E}} H_k^{t,s}(\mathbf{x}, l_t, l_s) \quad (1)$$

Under the boosting framework, we define $H_k^t(\mathbf{x}, l_t)$ and $H_k^{t,s}(\mathbf{x}, l_t, l_s)$ as linear combinations of weak learners:

$$H_k^t(\mathbf{x}, l_t) = \sum_{m=1}^{M} \rho_{k,m}^t h_{k,m}^t(\mathbf{x}, l_t) \quad (2)$$

$$H_k^{t,s}(\mathbf{x}, l_t, l_s) = \sum_{m=1}^{M} \rho_{k,m}^{t,s} h_{k,m}^{t,s}(\mathbf{x}, l_t, l_s) \quad (3)$$

Furthermore, we define the probability of an example $\mathbf{x}$ being class k as:

$$\hat{p}_k(\mathbf{x}) = \frac{\sum_L \exp(F_k(\mathbf{x}, \mathbf{L}))}{\sum_L \sum_{l=1}^{K} \exp(F_l(\mathbf{x}, \mathbf{L}))} \quad (4)$$

and we can define the loss function for LatentBoost as the negative log-likelihood of the training data:

$$loss_{LB} = \sum_{n=1}^{N} \Psi(\{y_k^{(n)}, F_k(\mathbf{x}^{(n)}, \mathbf{L})\}_{k=1}^{K}) = -\sum_{n=1}^{N} \sum_{k=1}^{K} y_k \log \hat{p}_k^{(n)}(\mathbf{x}^{(n)}) \quad (5)$$

[1] To simplify notation, we assume the same set $\mathscr{L}$. But our formulation can be generalized so that each latent variable is associated with a different label set.

To optimize the loss function, we learn the weak learners (both unary and pairwise) and their associated weights in an iterative fashion. At the m-th iteration, we compute the gradient of the loss function $loss_{LB}$ with respect to the current unary potential functions, $g_{k,m}^t(\mathbf{x}^{(n)}, l_t)$. We then pick the weak learner $h_{k,m}^t$ that is the most parallel in the N-dimensional data space with the negative gradient $\{-g_{k,m}^t(\mathbf{x}^{(n)}, l_t)\}_1^N$ by a least-squares minimization problem. Updating the pairwise functions is done in a similar fashion. The weights $\rho_{k,m}^t$ and $\rho_{k,m}^{t,s}$ can be simply computed by a line search algorithm. Putting everything together, we have the LatentBoost algorithm illustrated in Algorithm 1.

Algorithm 1 LatentBoost

$F_{k,0} = 0, k = 1, \ldots, K$
for $m = 1$ to M **do**
  Compute $\Pr(l_t|\mathbf{x}^{(n)})$, $\Pr(y_k^{(n)} = 1, l_t|\mathbf{x}^{(n)})$, $\Pr(l_t, l_s|\mathbf{x}^{(n)})$ and $\Pr(y_k^{(n)} = 1, l_t, l_s|\mathbf{x}^{(n)})$ $\forall n, \forall t \in \mathscr{V}, \forall (t,s) \in \mathscr{E}$
  for $k = 1$ to K **do**
    //update unary potentials
    for $t \in \mathscr{V}$ **do**
      $h_{k,m}^t = \arg\min_{h^t} \sum_{n=1}^{N} \sum_{l_t} [-g_{k,m}^t(\mathbf{x}_n, l_t) - h^t(\mathbf{x}^{(n)}, l_t)]^2$
      $\rho_{k,m}^t = \arg\min_{\rho^t} \sum_{n=1}^{N} \Psi(y_k^{(n)}, h_{k,m-1}^t(\mathbf{x}^{(n)}, l_t) + \rho^t h_{k,m}^t(\mathbf{x}^{(n)}, l_t))$
      $F_{k,m}^t(\mathbf{x}, l_t) = F_{k,m-1}^t(\mathbf{x}, l_t) + \rho_{k,m}^t h_{k,m}^t(\mathbf{x}, l_t)$
    end for
    //update pairwise potentials
    for $(t,s) \in \mathscr{E}$ **do**
      $h_{k,m}^{t,s} = \arg\min_{h^{t,s}} \sum_{n=1}^{N} \sum_{l_t, l_s} [-g_{k,m}^{t,s}(\mathbf{x}^{(n)}, l_t, l_s) - h^{t,s}(\mathbf{x}^{(n)}, l_t, l_s)]^2$
      $\rho_{k,m}^{t,s} = \arg\min_{\rho^{t,s}} \sum_{n=1}^{N} \Psi(y_k^{(n)}, F_{k,m-1}^{t,s}(\mathbf{x}^{(n)}, l_t, l_s) + \rho^{t,s} h_{k,m}^{t,s}(\mathbf{x}^{(n)}, l_t, l_s))$
      $F_{k,m}^{t,s}(\mathbf{x}, l_t, l_s) = F_{k,m-1}^{t,s}(\mathbf{x}, l_t, l_s) + \rho_{k,m}^{t,s} h_{k,m}^{s,t}(\mathbf{x}, l_t, l_s)$
    end for
  end for
end for

We test LatentBoost on the task of human action recognition. Our method for human action recognition operates on a figure-centric representation of the human figure extracted from an input video. We use the optical flow features in [2] on unary potentials and colour histogram features on pairwise potentials. We show that LatentBoost outperforms GradientBoost on two publicly available datasets: Weizmann human action dataset [1] and TRECVID surveillance event detection [5].

[1] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. *ICCV*, 2005.

[2] A.A. Efros, A.C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. *ICCV*, 2003.

[3] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 2010.

[4] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Technical report, Stanford University*, 1999.

[5] Alan F. Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and trecvid. In *the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, 2006.