Draft - Hierarchical Discriminative learning of Activities and Actions
Adding latent actions and multiple classifiers per action

November 1, 2015

1 Background

As before, the idea is to build an hierarchical model to be applied in video sequences of length T (number of frames) for classification of human activities, each video spatially and temporally composed by a variable number of atomic actions. In each frame, we define R regions, each characterized by a vector x_r, with $r = 1, \ldots, R$. A different action could be assigned to each region in each frame, and the complete sequence must be associated to a single activity. **We modify the original hierarchical model to include multiple linear classifiers per action**. We create two new concepts: **semantic actions**, that refer to actions *names* that compose an activity; and **atomic sequence**, that refers to the sequence of poses that conform an action. Several atomic sequences can be associated to a single semantic actions, so we can create disjoint sets of atomic sequences, each set associated to a single semantic action. The main idea is that the action annotations in the datasets are associated to semantic actions, whereas for each semantic action we learn several atomic sequence classifiers. With this formulation, we can handle the multimodal nature of semantic actions, which with a single linear classifier is in general not enough to cover the changes in motion, poses , or even when some actions labels have different meaning depending on the context (e.g. the action "open" can be associated to open a can, open a door, etc.).

We also include the formulation using three levels of knowledge of semantic actions:

i. Knowing the temporal spanning and spatial region for the actions in each video (the original model used along with Composable Activities dataset)

ii. Only using the temporal annotations od semantic actions

iii. Only knowing that the action exists in the video, like an attribute in the video but with no temporal or spatial annotations.

1.1 Definitions

- T: number of frames of the sequence.
- R: number of spatial regions of each frame.
- C: number of activities
- G: number of semantic actions
- A: total number of atomic action classifiers
- K: size of pose dictionary

- t: index of frame, $t \in \{1, \ldots, T\}$.
- r: index of region, $r \in \{1, \ldots, R\}$.
- c: index of activity, $c \in \{1, \ldots, C\}$.
- g, g': index of semantic action, $g, g' \in \{1, \ldots, G\}$.
- a: index of atomic sequence, $a \in \{1, \ldots, A\}$
- k: index of pose, $k \in \{1, \ldots, K\}$.

1.2 Notation used in this document

Three shortened notations are used across the document:

- $\sum_x$ refers to $\sum_{x=1}^{X}$, with x some index defined in the last section.
- $\sum_{x,y,\ldots}$ refers to $\sum_x \sum_y \cdots$.
- δ_a^b refers to Kronecker delta function $\delta(a = b)$.

2 Model for multiple atomic sequence classifiers per semantic action

Energy function:

$$E = E_{\text{activity}} + E_{\text{action}} + E_{\text{pose}} + E_{\text{action transition}} + E_{\text{pose transition}}. \tag{1}$$

At the lowest level of the hierarchy, $Z^\top = (z_{1,1} \ldots z_{T,R})$ assigns poses to entries in the dictionary. In particular $z_{t,r} = k$ assigns pose (dictionary word) k to region r in frame t.

$$E_{\text{pose}} = \sum_{r,t} {w^r_{z_{t,r}}}^\top x_{t,r} = \sum_{r,t,k} {w^r_{k,r}}^\top x_{t,r} \delta^k_{z_{t,r}} \tag{2}$$

At second level, we assume that we have atomic sequence labels for each frame and region, grouped in a vector $V^\top = [v_{1,1}, \ldots, v_{T,R}]$ (if we do not have that information, we must estimate the atomic sequence labels). $h^{a,r}(Z, V)$ is the histogram over the pose dictionary, at those frames assigned to atomic sequence a in region r. Each entry k in $h^{a,r}(Z, V)$ is given by:

$$h^{a,r}_k(Z, V) = \sum_t \delta^k_{z_{t,r}} \delta^a_{v_{t,r}} \tag{3}$$

$$E_{\text{action}} = \sum_{r,a} {\beta^r_a}^\top h^{a,r}(Z, V) = \sum_{r,a,t,k} \beta^r_{a,k} \delta^k_{z_{t,r}} \delta^a_{v_{t,r}} \tag{4}$$

To associate the atomic sequences to semantic actions, each frame is annotated with a new label vector $U^\top = [u_{1,1}, \ldots u_{T,R}]$, which indicates the semantic action that the action sequence label $v_{t,r}$ belongs in each frame and region. Recall that the action sequence form disjoint groups, each group associated to a semantic action, so many atomic sequence labels values $a \in \{1, \ldots, A\}$ are associated to a single semantic label value $g \in \{1, \ldots, G\}$.

At third level, $h^r(U)$ is the histogram corresponding to region r over the semantic action labels accumulated over all frames. Each entry g in $h^r(U)$ is given by:

$$h^r_g(U) = \sum_t \delta^g_{u_{t,r}} \tag{5}$$

So the energy in the activity level is

$$E_{\text{activity}} = \sum_r {\alpha^r_y}^\top h^r(U) = \sum_{r,g,t} \alpha^r_{y,g} \delta^g_{u_{t,r}} \tag{6}$$

One alternative formulation for the energy in this level is to use the aggregated histograms of actions and use a single classifier per activity using all regions, in contrast to a different activity classifier for every region.

$$E_{\text{activity\_alternative}} = \sum_r \alpha_y{}^\top h^r(U) = \sum_{t,g} \alpha_{y,g} \sum_r \delta^g_{u_{t,r}} \tag{7}$$

In terms of atomic sequences and pose transition, energies are defined as histograms over two consecutive frames, involving atomic sequences (V) and poses (Z):

$$E_{\text{action transition}} = \sum_{r,a,a'} \gamma^r_{a',a} \sum_t \delta^{a'}_{v_{t-1,r}} \delta^a_{v_{t,r}} \tag{8}$$

$$E_{\text{pose transition}} = \sum_{r,k,k'} \eta^r_{k',k} \sum_t \delta^{k'}_{z_{t-1,r}} \delta^k_{z_{t,r}} \tag{9}$$

Summarizing, we have three levels in the hierarchy. At the lowest level, we use the region descriptors to associate the region to a pose. In the middle level, we use pose histograms to associate poses to atomic sequences. Then, the atomic sequence labels are summarized into semantic actions, and at the highest level the histogram of semantic actions are used. The transition terms are computed over transitions of atomic sequences and poses.

We can use the same base math of the previous model, since we are dealing with the same labels (poses and actions), and in this new model we just combine action sequences to conform semantic actions, using the mid-level annotations as semantic actions.

One problem we have to deal is how to assign the labels V in each video. Assume for the moment we have the same annotations as Composable Activity, which have a single label for activity, and multiple spatial and temporal annotations for actions for each video. In terms of labels, this means we have the ground truth for the labels U (semantic actions) for each video, and we need to estimate the labels V of atomic sequences. In general, we need to:

1. Find a suitable number of atomic sequences for each semantic action
2. Assign the atomic sequences to each action annotation for all videos. Each atomic sequence must lie in the group conforming the same semantic action.

A simple method to find the number of atomic sequences for each semantic actions could be first finding an appropriate number of clusters using Cattell's scree test as used by [Raptis et. al. 2012] to find the number of clusters for each video, which is not perfect but will give a high number of clusters to highly variant semantic actions, and a low number of clusters if the semantic actions are similar in all videos. Then, we can find V using the cluster assignments.

To learn the model parameters, we can, as usual, formulate the model in an optimization framework, where the goal is to find the best parameters α, β, w, γ and η that yield to the best classification of the training set of M videos. We can formulate our model using the following objective function:

$$\min_{\alpha,\beta,w,\gamma,\eta} \Omega(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{w}, \boldsymbol{\gamma}, \boldsymbol{\eta}) + \frac{C}{M} \sum_{i=1}^{M} \max_{Z,V,y} \left(E(X_i, Z, V, U, y) + \Delta((y_i, U_i), (y, U)) - \max_{Z_i} E(X_i, Z_i, U_i, V_i, y_i) \right), \tag{10}$$

where $\Omega(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{w}, \boldsymbol{\gamma}, \boldsymbol{\eta})$ is a regularization term over the coefficients of the classifiers, visual dictionary and action/pose transition terms. $\Delta((y_i, U_i), (y, U))$ is the loss function that measures the labeling performance of activities and actions. As we plan to use three levels of knowledge of semantic actions, we must use a different loss function according to what we know in advance:

i. If we know the temporal spanning of semantic actions along with the spatial region each action is performed, then a loss function that favors predicting the correct labels at activity and semantic action levels is given by:

$$\Delta((y_i, U_i), (y, U)) = \lambda_1 \delta(y_i \neq y) + \frac{\lambda_2}{T} \sum_{r,t} \delta(u_{(t,r)_i} \neq u_{t,r}). \tag{11}$$

ii. If we only know the temporal spanning of the semantic actions, then we can no longer use the labels $u_{(t,r)_i}$ as known. Then, we only know which actions are performed in each frame. If we use the temporal information, we can form groups S_t of semantic action labels for each frame, and use a loss function that favors predicting semantic actions to each frame belonging to the group S_t:

$$\Delta((y_i, U_i), (y, U)) = \lambda_1 \delta(y_i \neq y) + \frac{\lambda_2}{T} \sum_{r,t} \delta(u_{(t,r)} \notin S_t). \tag{12}$$