Joint Action Recognition and Pose Estimation From Video

Bruce Xiaohan Nie[1], Caiming Xiong[1], Song-Chun Zhu[1]
[1]Department of Statistics, University of California, Los Angeles.

Action recognition and pose estimation from video are closely related tasks for understanding human motion, however most methods learn two separate models and then combine them in a sequential order. In this paper, we propose a framework to integrate training and testing of the two tasks. A spatial-temporal And-Or graph model is introduced to represent action at three scales. Specifically the action is decomposed into poses which are further divided to mid-level ST-parts and then parts. The hierarchical structure of our model captures the geometrical and appearance variation of pose at each frame and lateral connections between ST-parts at adjacent frames capture the action-specific motion information. The model parameters for three scales are learned discriminatively and the inference of action label and poses are done efficiently by Dynamic Programming. The experiments demonstrate that our approach can not only achieve state-of-art accuracy in action recognition but also improve pose estimation.

We start by building a spatial-temporal And-Or graph model[4] to represent action and poses together. Hierarchical structure of our model can represent top-down part geometric configuration in a single frame and lateral temporal pose relation in subsequent frames. On the top layer, the low-scale action information is captured by coarse-level features and the action is constitute of poses at each frame. Each pose is decomposed into five independent mid-level parts named 'ST-parts' which cover large portion of human bodies and are more robust to image variations. All fine parts are conditioned on their ST-part parents. Each ST-part is discretized into several components by clustering. The ST-parts with the same component can be seen as a poselet[1] that has small variation of appearance and deformation and each component is represented by mid-level features and fine-level part features from single image pose estimation.

In order to capture the specific motion information of each action, the same ST-parts at adjacent frames are connected to represent temporal co-occurrence and deformation. The model parameters at three levels are first trained separately by S-SVM and then combined by mixture of experts method. Due to the independence between ST-parts of each pose, we can infer both action label and poses efficiently by DP.

The fine-level score of one image sequence is written as follows,

$$S_H(A) = \sum_{i=1}^{M}(\sum_{t=1}^{T} S(l_i^t) + \sum_{t=1}^{T-1} S(l_i^{t+1}|l_i^t)) \quad (1)$$

Here A is an action example. $S(l_i^t)$ is the score of ST-part i at time t. $S(l_i^{t+1}|l_i^t)$ is the transition score between ST-parts l_i^t and l_i^{t+1}. T is the total length of the action example and M is the number of ST-parts.

The score of ST-part i is defined by:

$$S(l_i) = S_d(l_i) + S_h(l_i) + \lambda \sum_{j=0}^{N_i} S(o_j) + \lambda \sum_{j=1}^{N_i} S(o_j, o_0) \quad (2)$$

There are four terms contributing to ST-part score. The first two terms are classification scores and the last two terms are detection scores. $S_d(l_i) =< \omega_d^{l_i}, \psi(l_i) >$ measures the compatibility of component c_i. $S_h(l_i) =< \omega_h^{l_i}, \psi(c_i) >$ is histogram score of component c_i. $S(o_j)$ is score of part j and $S(o_j) = P(o_j)$ where $P(o_j)$ is part marginal score from pose estimation. $S(o_j, o_0) =< \omega^{ij}, \psi(E_d^{ij}) >$ is deformation score of part j related to the root part. Parameter λ is the weight for detection score. The inference algorithm will search all possible ST-parts in feature pyramid and output a top candidate list for each frame.

The transition score between two ST-parts is defined as:

$$S(l_i^{t+1}|l_i^t) = S(c_i^t, c_i^{t+1}) + \beta d(l_i^t, l_i^{t+1}) \quad (3)$$

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

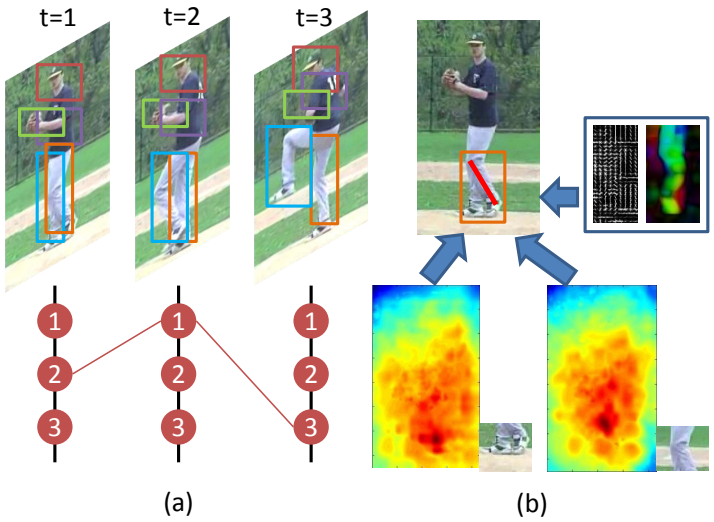

Figure 1: An example of our inference method. (a) For each frame we generate several ST-part candidates and obtain the best path for each ST-part by DP. (b) The ST-part is represented by the mid-level features (HOG and HOF template) and fine-level features (scores of knee and ankle).

It includes two components: the co-occurrence score $S(c_i^t, c_i^{t+1}) =< \omega_o^{l_i}, \psi(E_o^{c_i^t, c_i^{t+1}}) >$ and smoothness score $\beta d(l_i^t, l_i^{t+1})$, where β is the weight for the smoothness.

With coarse-level and mid-level scores, the action score can be written in the following,

$$S(A) = \pi_L(A)S_L(A) + \pi_M(A)S_M(A) + \pi_H(A)S_H(A) \quad (4)$$

$S_L(A) =< \omega_L, \psi_L(A) >$ is coarse-level score and $\psi_L(A)$ is the bag-of-words feature from [2]. $S_M(A) =< \omega_M, \psi_M(A) >$ is mid-level score and $\psi_M(A)$ is from [3]. The weights $\pi_L(A) =< \omega_L', \phi_L'(A) >$, $\pi_M(A) =< \omega_M', \phi_M'(A) >$ and $\pi_H(A) =< \omega_H', \phi_H'(A) >$ are linear functions on features of action example A.

The inference procedure is illustrated in Fig.1. The fine-level action score $S_H(A)$ is divided into M independent terms each of which corresponds to the summation of unary scores and binary transition scores for one ST-part, thus the dynamic programming can be used to find the best ST-part path. This procedure is repeated M times to find the total M best paths for each action label. Finally the action label with maximum score is obtained. With the best action label, we trace back to the best ST-part paths for this action and then obtain all joint locations. To speed up computation, we pick the ST-part candidates that have score above a threshold at each frame. We connect all candidates on consecutive frames and compute their unary scores and binary transition scores.

[1] L. Bourdev and J. Malik. Poselets: Body Part Detectors Trained Using 3D Human Pose Annotations. In *ICCV*, 2009.

[2] H. Wang, A. Klaser, C. Schmid, and C. Liu. Dense trajectories and motion boundary descriptors for action recognition. *IJCV*, 103(1):60–79, 2013.

[3] J. Wang, B. X. Nie, Y. Xia, Y. Wu, and S. C. Zhu. Cross-view Action Modeling, Learning and Recognition. In *CVPR*, 2014.

[4] S. C. Zhu and D. Mumford. A stochastic grammar of images. *Foundations on Trends in Computer Graphics and Vision*, 2(4):259–362, 2006.