Predicting the *Where* and *What* of actors and actions through
Online Action Localization

Khurram Soomro, Haroon Idrees, Mubarak Shah  
{ksoomro, haroon, shah}@cs.ucf.edu

**Problem:**
- Prediction and Detection of Actions
- Typically action localization and recognition is performed in an offline manner
- This disallows timely localization of actions for:
  - Surveillance tasks
  - Human - Computer Interaction (HCI)

**Key Contributions:**
- Introduce a new problem of Online Action Localization for streaming videos
- Use high-level pose to learn a mid-level superpixel-based foreground model at each time instant
- Predict action label using Dynamic Programming
- Detect action location using foreground probability through Conditional Random Field

**Online Action Localization:**

Online Action Localization = Action Detection + Prediction

**Framework:**

(a) Input Stream of Video Frames  
(b) Superpixel Extraction and Pose Estimation  
(c) Learn Superpixel-based Appearance Model  
(d) Superpixel-based Foreground Likelihood  
(e) Pose Refinement  
(f) Segment Action CRF + Action Prediction using SVM

**Formulation:**

$p(X_t|S_{t-1}, P_{t-1}) = Z^{-1}(S_{t-1}) p(P_t|X_t) f(X_t|X_{t-1}) p(X_{t-1}|S_{t-1}, P_{t-1}) dX_{t-1}$

$X_t$ = Location representing bounding box at time $t$

$S_t$ represents all superpixels and $P_t$ represents all poses within the time window of $\delta$ frames \([t - \delta, t]\)

**Pose-based Foreground Likelihood:**

$p(P_t|X_t) = \exp(\alpha_{\text{pose}} \cdot H_{\text{pose}}(p_t)) = H_{\text{raw}}(p_t) + \lambda_{\text{app}}(p_t) + \lambda_{\text{loc}}(p_t) + \lambda_{\text{sc}}(p_t)$

- Appearance smoothness of joints:
  $J_{\text{app}}(p_t) = \sum_{j=1}^{\text{num_joints}} |H_{\text{app}}(\hat{q}_j) - H_{\text{app}}(\hat{q}'_j)|$

- Location smoothness of joints:
  $J_{\text{loc}}(p_t) = \sum_{j=1}^{\text{num_joints}} |\hat{q}^j - \hat{q}'^j|$  

- Scale smoothness of joints:
  $J_{\text{sc}}(p_t) = \exp\left(\left(J_{\text{max}} - J_{\text{max}}\right) - \left(J_{\text{min}} - J_{\text{min}}\right)\right)$

**Experimental Results (Qualitative):**

| Action Prediction (a) and Localization (b-c) with Time (HMDB and UCF Sports): |

| Action Prediction Analysis with Time (HMDB (a) and UCF Sports (b)): |

| Action Localization with Offline methods (HMDB (a) and UCF Sports (b-c)): |