Extracting Moving People From Internet Videos

Juan Carlos Niebles[1,2] jniebles@princeton.edu
Bohyung Han[3] bohyung.han@mobileye.com
Andras Ferencz[3] andras.ferencz@mobileye.com
Li Fei-Fei[1] feifeili@cs.princeton.edu

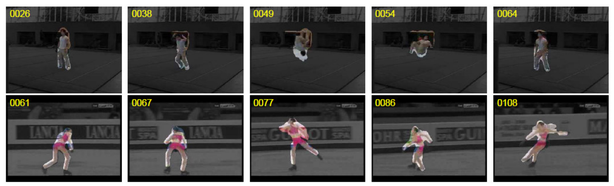
[1]Princeton University [2]Universidad del Norte, Colombia

http://www.mobileye.com http://vision.cs.princeton.edu

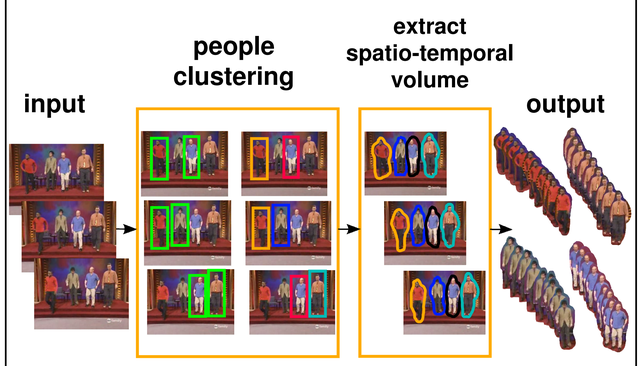
Summary

Problem: Extract the spatio-temporal volume that encloses each person on a video.

Contribution:
- An automatic system that can extract arbitrary human motion volumes from challenging YouTube videos. Some of the challenges are:
  - Highly compressed & low quality video
  - Unknown human motion and poses
  - Unknown camera parameters & motion
  - Background motion and clutter

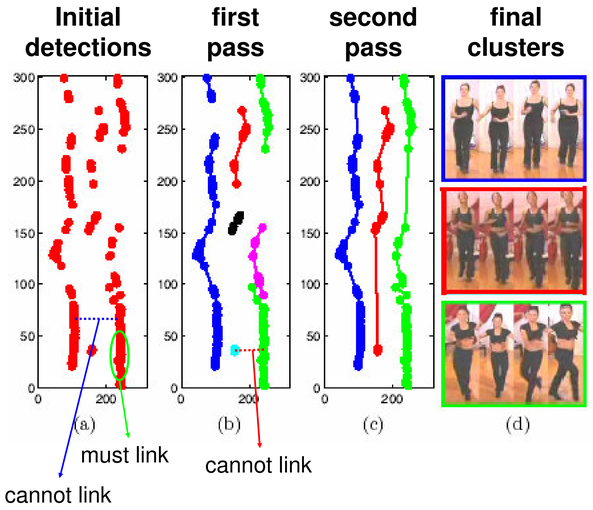


System Overview



People Clustering

- Generate hypothesis of people location in the video using a pedestrian detector.
- Impose must-link and cannot-link constrains
  - must-link: detections with high similarity and spatio-temporal coherence.
  - cannot-link: detections on the same frame.
- First pass: describe each detection window with a global histogram. Use low-level tracking to generate must-link constraints. Apply constrained clustering.
- Second pass: estimate head-torso locations and obtain more accurate appearance descriptors. Apply constrained clustering.



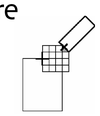
Extracting the Motion Volume

Human body model [Ramanan NIPS '07]
- Human body represented as a 10-part-tree pictorial structure model.

$$L(B \mid I) \propto \sum_{(i,j)\in E} \Psi(p_i - p_j) + \sum_i \Phi(p_i)$$

where the spatial relations among parts are

$$\Psi(p_i - p_j) = \alpha_i^T \cdot \mathrm{bin}(p_i - p_j)$$



and the measurement for each part is

$$\Phi(p_i) = \beta_i^T f_i(I(p_i))$$

- Inference can be done efficiently by message passing. The message from part i to its parent j is:

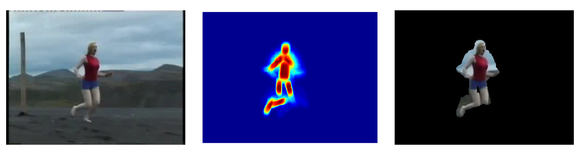
$$M_i(p_j) \propto \sum_{p_j} \Psi(p_i - p_j) O_i(p_i)$$

$$O_i(p_i) \propto \Phi(p_i) \cdot \prod_{k \in C_i} M_k(p_i)$$

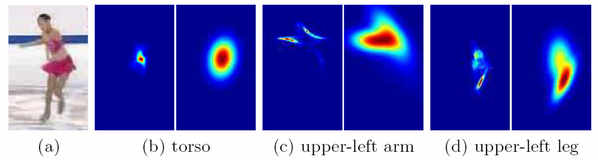
The marginal for part i is

$$P(p_i) \propto \Phi(p_i) \sum_{p_j} \Psi(p_i - p_j) P(p_j \mid I)$$

- Projecting the part posteriors into the image gives a rough segmentation of the body region



Reducing computation of measurements
- Reducing search space has several benefits
  - Measurements can be limited to a smaller space, reducing computation time
  - Distracting background observations can be avoided
  - Final estimation accuracy can be improved
- The marginal of each part is approximated with a Gaussian Mixture Model.
- The number of mixture components, means and variances automatically is found via Kernel Density Approximation (KDA) [Han et al. 'PAMI08]



- Body part densities are diffused to the next frame, which accounts for the motion of body parts.
- Densities are propagated in a Bayesian filtering framework

$$p(\mathbf{X}_t \mid \mathbf{Z}_{1:t}) \propto p(\mathbf{Z}_t \mid \mathbf{X}_t) p(\mathbf{X}_t \mid \mathbf{Z}_{1:t-1})$$

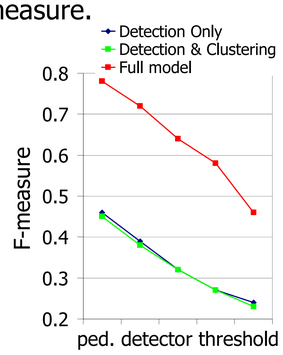
$$= \left(\sum_{i=1}^{N_1} \mathcal{N}(\kappa_i, \mathbf{x}_i, \mathbf{P}_i) \right) \left(\sum_{j=1}^{N_2} \mathcal{N}(\tau_j, \mathbf{y}_j, \mathbf{Q}_j) \right)$$

- The number of mixture components is limited by applying KDA once again.

Experimental Results

- Dataset: 50 *YouTube* videos with varying motions, backgrounds and camera settings.
- We measure the retrieval performance of our system in terms of the F-measure.

	Precision	Recall	F
Detection	0.89	0.31	0.46
Detection & Clustering	0.89	0.30	0.45
Full Model	0.83	0.73	0.78



Discussion

- Limitations
  - No exact segmentation
  - At least one frame with upright position
  - No occlusion reasoning between body parts
- Future work includes
  - Integration of the clustering and tracking steps to allow for positive feedback and improved estimation
  - Incorporation of other bottom-up cues.

Ref: J.C. Niebles, B. Han, A. Ferencz & L. Fei-Fei. Extracting Moving People from Internet Videos. *ECCV* 2008.
http://vision.cs.princeton.edu/projects/extractingPeople.html