Joint Segmentation and Pose Tracking of Human in Natural Videos*

Taegyu Lim[1,2] Seunghoon Hong[2] Bohyung Han[2] Joon Hee Han[2]
[1]DMC R&D Center, Samsung Electronics, Korea
[2]Department of Computer Science and Engineering, POSTECH, Korea
tgx.lim@samsung.com {maga33,bhhan,joonhan}@postech.ac.kr

Abstract

We propose an on-line algorithm to extract a human by foreground/background segmentation and estimate pose of the human from the videos captured by moving cameras. We claim that a virtuous cycle can be created by appropriate interactions between the two modules to solve individual problems. This joint estimation problem is divided into two subproblems, foreground/background segmentation and pose tracking, which alternate iteratively for optimization; segmentation step generates foreground mask for human pose tracking, and human pose tracking step provides foreground response map for segmentation. The final solution is obtained when the iterative procedure converges. We evaluate our algorithm quantitatively and qualitatively in real videos involving various challenges, and present its outstanding performance compared to the state-of-the-art techniques for segmentation and pose estimation.

1. Introduction

Foreground/background segmentation and human pose estimation have been studied intensively in recent years and significant performance improvement has been achieved so far. However, these problems are still known to be very challenging, especially in unconstrained videos, due to various issues in observation and inference. Existing algorithms typically assume stationary camera environment, or suffer from low accuracy with high computational cost.

Although foreground/background segmentation and human pose estimation are potentially related and complementary, the majority of algorithms attempt to solve the two problems separately and the investigation of a joint estimation technique is not active yet. We introduce an algorithm to address foreground/background segmentation and human pose tracking simultaneously in a video captured by a moving camera. The former determines shape and position of human body area in each frame, and the latter estimates temporally coherent human body configurations. Sample results of our algorithm are illustrated in Figure 1.

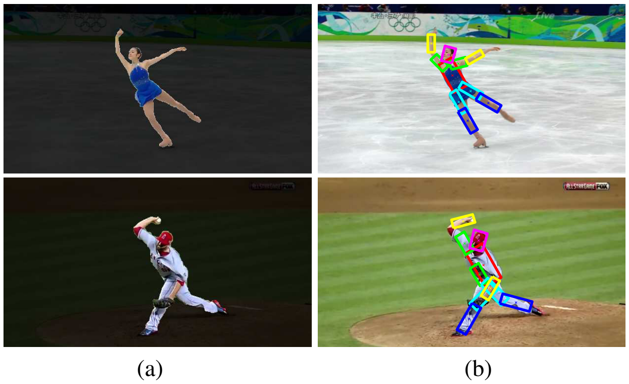

(a) (b)

Figure 1: Sample results of the proposed algorithm. Our joint estimation algorithm is accurate and robust to background noise and significant camera motion. (a) Foreground/background segmentation. (b) Pose estimation.

Foreground/background segmentation problem in moving camera environment has been studied actively these days. Sheikh *et al.* [21] proposes an algorithm to construct foreground and background appearance models for pixel-wise labeling based on a sparse set of motion trajectories. Similarly, a matrix factorization is employed in [6] to decompose a dense set of trajectories into foreground or background via low rank and group sparsity constraints. In this work, motion segmentation is also integrated to provide pixels with the binary labels. In [9], a sparse set of trajectories are clustered in a low dimensional space and labeled as foreground or background. These labeled trajectories are used to compute pixel-wise motion and appearance models for both foreground and background. In [14, 15], each frame is divided into regular grid blocks, and each block motion is estimated to propagate foreground and background models in a recursive manner. On the other hand, [17, 16] study techniques to extract human body areas from videos through the combination of various algorithms.

*This work was supported by MEST Basic Science Research Program through the NRF of Korea (NRF-2012R1A1A1043658), and Samsung Electronics Co., Ltd.

Several interesting approaches for pose estimation and tracking have been proposed. Pictorial structure [10] is a widely adopted model for human pose estimation, which computes the Maximum A Posteriori (MAP) estimate of body configuration efficiently by dynamic programming. A variation of the pictorial structure model is introduced in [19], where part-specific appearance models are learned and the pose of an articulated body is estimated by message passing in a Conditional Random Field (CRF). This approach is extended to video in [11], which improves performance by reducing search space progressively and by integrating spatio-temporal parsing. Andriluka *et al.* [1] present a discriminatively trained appearance model and a flexible kinematic tree prior on the configurations of body parts. Yang and Ramanan [23] propose a mixture-of-parts model, where a mixture of non-articulated small patches approximates various transformations of each body part effectively. Most of these approaches assume that foreground/background segmentation is given, or do not utilize the segmentation labels at all.

There are several prior studies related to the joint formulation of foreground/background segmentation and pose estimation. *ObjCut* [13] tackles a joint problem of segmentation and pose estimation in an image, where shape model is obtained from pose and segmentation is estimated given the shape model, and a similar approach is proposed by [5]. Kohli *et al.* [12] introduce *PoseCut* for simultaneous 3D pose tracking and segmentation. However, its results may be sensitive to initial pose estimation and background clutter due to weak foreground/background appearance models. Brox *et al.* [4] couple pose estimation and contour extraction problems in a multi-camera environment. In [22], the multi-level inference framework for pose estimation and segmentation from a single image is proposed. Recently, a technique for segmentation and pose estimation of human is presented in [8], where foreground area is first separated by grab-cut [20] given a bounding box, and human pose is estimated based on the foreground region. Note that most of approaches are developed for a single image or limited to naïve extension to video data.

We propose a unified probabilistic framework for foreground/background segmentation and pose tracking in videos captured by moving cameras. Our algorithm is an iterative approach that combines foreground/background segmentation and pose tracking. In each iteration of our algorithm, segmentation module propagates foreground and background models, and provides pose tracking module with foreground mask using the estimated labels. In pose tracking module, the configuration of each body part is estimated by multiple part detectors with label constraint, and gives shape prior represented by probabilistic foreground response map back to segmentation module. The refined segmentation result and the estimated pose configuration are utilized to update foreground/background motion and shape prior in the next iteration, respectively. Such iterative procedure is repeated until convergence in each frame.

Our joint human segmentation and pose tracking algorithm has the following contributions and characteristics:

- We formulate a probabilistic framework of a joint and iterative optimization procedure for foreground-background segmentation and pose tracking.
- We propose an online algorithm based on a recursive foreground/background appearance modeling and sequential Bayesian filtering for pose tracking.
- Our algorithm is applied to natural videos and improves both segmentation and pose tracking performance significantly.

The rest of paper is organized as follows. We first describe the objective and main formulation of our algorithm in Section 2. Foreground/background segmentation in a moving camera environment is discussed in Section 3, and our pose estimation technique is presented in Section 4. Section 5 describes model update strategy in each frame. Section 6 illustrates experimental results and evaluates the performance of the proposed algorithm.

2. Objective and Main Formulation

Our goal is to perform foreground/background segmentation and human pose tracking jointly and sequentially in a video captured by a moving camera. For the purpose, we estimate the MAP solution over pose parameters and segmentation labels at each frame given observation history, which is formally given by

$$(\mathbf{X}_t^*, \mathcal{L}_t^*) = \arg\max_{\mathbf{X}_t, \mathcal{L}_t} p(\mathbf{X}_t, \mathcal{L}_t | \mathcal{I}_{1:t}), \tag{1}$$

where $\mathbf{X}_t$ and $\mathcal{L}_t$ denote human body configuration and pixel-wise segmentation labels at time t, respectively, and $\mathcal{I}_{1:t}$ represents all image evidence. The human body configuration denoted by $\mathbf{X}_t \triangleq \{\mathbf{x}_{1,t}, \ldots, \mathbf{x}_{m,t}\}$ is composed of a set of pose parameters for individual body parts, where m is the number of parts[1]. The pose of each body part, $\mathbf{x}_{i,t}$, is represented by location, orientation and scale information. The segmentation of an image with n pixels is given by $\mathcal{L}_t \triangleq \{\ell_{1,t}, \ldots, \ell_{n,t}\}$, where the label in the k-th pixel $\ell_{k,t}$ is either 0 (background) or 1 (foreground).

The optimization problem in Eq. (1) involves a very high dimensional search space, and $\mathbf{X}_t$ and $\mathcal{L}_t$ have significant mutual dependency. Therefore, we divide the original problem into two subproblems—foreground/background seg-

[1] The number of parts is 10 in this work: head, torso, left/right upper arms, left/right lower arms, left/right upper legs, and left/right lower legs.

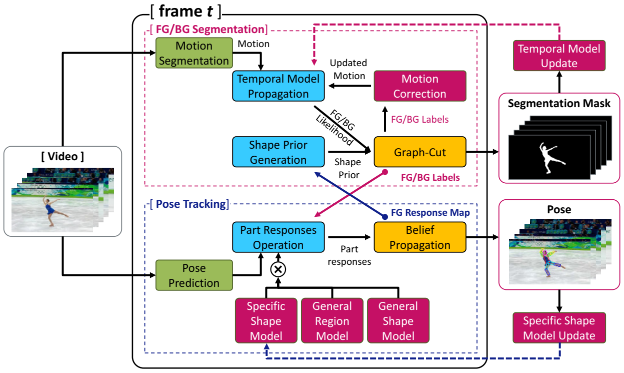


Figure 2: Overview of our algorithm, which is composed of two modules: foreground/background segmentation and pose tracking. Segmentation is refined with shape information given by pose tracking while pose tracking utilizes segmentation mask. These two tasks are optimized jointly in a single framework.

mentation and pose tracking—and solve the following energy minimization problem:

$$\min_{\mathcal{L}_t,\mathbf{X}_t,\mathcal{Y}_t} \; E_{\text{seg}}(\mathcal{L}_t,\mathcal{Y}_t,\mathcal{I}_t)+E_{\text{pose}}(\mathbf{X}_t,\mathcal{L}_t,\mathcal{I}_t), \tag{2}$$

where $E_{\text{seg}}(\mathcal{L}_t,\mathcal{Y}_t,\mathcal{I}_t)$ and $E_{\text{pose}}(\mathbf{X}_t,\mathcal{L}_t,\mathcal{I}_t)$ denote energy functions for segmentation and pose tracking, respectively. $\mathcal{Y}_t$ denotes a foreground response map for human body area generated from the pose variable $\mathbf{X}_t$. The energy terms in Eq. (2) are defined probabilistically as follows:

$$E_{\text{seg}}(\mathcal{L}_t,\mathcal{Y}_t,\mathcal{I}_t) = -\log\left(p(\mathcal{L}_t|\mathcal{Y}_t,\mathcal{I}_t)\right), \tag{3}$$

$$E_{\text{pose}}(\mathbf{X}_t,\mathcal{L}_t,\mathcal{I}_t) = -\log\left(p(\mathbf{X}_t|\mathcal{L}_t,\mathcal{I}_{1:t})\right). \tag{4}$$

For the optimization of the problem in Eq. (2), we employ an EM-style iterative method; we first fix E_{pose} and minimize E_{seg}, and then minimize E_{pose} with E_{seg} fixed. This iterative procedure is repeated until convergence. The outline of overall system is illustrated in Figure 2.

3. Foreground/Background Segmentation

We first need to identify foreground/background label for each pixel by minimizing $E_{\text{seg}}(\mathcal{L}_t,\mathcal{Y}_t,\mathcal{I}_t)$, equivalently maximizing the segmentation posterior $p(\mathcal{L}_t|\mathcal{Y}_t,\mathcal{I}_t)$, which is given by

$$p(\mathcal{L}_t|\mathcal{Y}_t,\mathcal{I}_t) \propto p(\mathcal{I}_t|\mathcal{L}_t,\mathcal{Y}_t)p(\mathcal{L}_t|\mathcal{Y}_t), \tag{5}$$

where $p(\mathcal{I}_t|\mathcal{L}_t,\mathcal{Y}_t) = p(\mathcal{I}_t|\mathcal{L}_t)$ is an observation likelihood given segmentation label, and $p(\mathcal{L}_t|\mathcal{Y}_t)$ is a prior of segmentation given pose. To the end, we employ a slightly modified version of [15]; spatial model composition step is removed but pose tracking feedback is taken into account for the joint formulation. Foreground and background models in the previous frame are propagated to the current frame using motion information through an iterative procedure. The segmentation labels, $\mathcal{L}_t^*$, are obtained efficiently by graph-cut algorithm [2, 3] based on $p(\mathcal{L}_t|\mathcal{Y}_t,\mathcal{I}_t)$, which depends on the two terms—observation likelihood and segmentation prior given pose. We mainly discuss how these two components in Eq. (5) are estimated in the rest of this section, and recommend to refer to [15] for more details about foreground/background modeling and segmentation.

3.1. Estimation of Observation Likelihood

We obtain the observation likelihood $p(\mathcal{I}_t|\mathcal{L}_t)$ from the probabilistic models of foreground and background appearances. We divide a frame into N regular grid blocks[2] and construct foreground and background models in each block by kernel density estimation. Suppose that foreground model $\boldsymbol{\varphi}_{f,t-1}^k$ and background model $\boldsymbol{\varphi}_{b,t-1}^k$ for the k-th block B_{t-1}^k at time $t-1$ are already given, where $\{\mathbf{y}_{\xi,t-1}^1,\ldots,\mathbf{y}_{\xi,t-1}^{n_\xi}\}$, $\xi\in\{b,f\}$, are sample data comprising the models. The foreground and background likelihoods of an observed pixel $\mathbf{z}_{t-1}$ are respectively given by

$$p(\mathbf{z}_{t-1}|\boldsymbol{\varphi}_{f,t-1}^k) = \alpha\mathbf{U}(\mathbf{z}_{t-1}) + \frac{1-\alpha}{n_f}\sum_{i=1}^{n_f} K_{\mathbf{H}}(\mathbf{z}_{t-1}-\mathbf{y}_{f,t-1}^i), \tag{6}$$

$$p(\mathbf{z}_{t-1}|\boldsymbol{\varphi}_{b,t-1}^k) = \frac{1}{n_b}\sum_{i=1}^{n_b} K_{\mathbf{H}}(\mathbf{z}_{t-1}-\mathbf{y}_{b,t-1}^i), \tag{7}$$

where $\mathbf{U}(\cdot)$ is a uniform distribution, $\alpha\in[0,1]$ is a mixture weight for the uniform distribution, and $K_{\mathbf{H}}(\cdot)$ denotes a kernel function with bandwidth $\mathbf{H}$.

To construct foreground and background models at time t based on the earlier ones, we compute foreground and background motion vectors in each block, and propagate models from the previous frame using the block motions. A block motion at time t, $\mathbf{V}_{\xi,t}^k$, is given by

$$\mathbf{V}_{\xi,t}^k = \frac{1}{|\chi_{\xi,t}^k|}\sum_{\mathbf{v}_t^i\in\chi_{\xi,t}^k}\mathbf{v}_t^i, \tag{8}$$

where $\mathbf{v}_t^i$ denotes the backward motion of the i-th pixel and $\chi_{\xi,t}^k$ is a set of backward motions of the pixels labeled as ξ in B_t^k. If the motion observation in a block is insufficient due to occlusion, background block motion is estimated by the average motion of adjacent blocks and foreground block motion is set to zero. Note that $\chi_{\xi,t}^k$ depends on the segmentation labels and is updated in each iteration due to potential label changes of the pixels within the block.

Through the iterative model propagation with respect to backward block motion $\mathbf{V}_{\xi,t}^k$, the likelihood of an observed

[2] The size of each block is 24 × 24 in our experiment.

pixel $\mathbf{z}_t$ in B^k_t is determined by the new model, which is given by a mixture of block models at time $t-1$, as

$$p(\mathbf{z}_t|\tilde{\boldsymbol{\varphi}}^k_{\xi,t}) = \sum_{B^j_{t-1}\in\mathcal{A}^k_{\xi,t-1}} \omega^j_\xi p(\mathbf{z}_t|\boldsymbol{\varphi}^j_{\xi,t-1}), \tag{9}$$

where $\mathcal{A}^k_{\xi,t-1}$ is a set of blocks at time $t-1$ overlapped with transformed block, $T(B^k_t;\mathbf{V}^k_{\xi,t})$, and ω^j_ξ is the normalized mixture weight. Note that the mixture weight ω^j_ξ is determined by the relative overlapping ratio between $T(B^k_t;\mathbf{V}^k_{\xi,t})$ and B^j_{t-1}.

Once foreground and background models are constructed at time t, we can compute $p(\mathcal{I}_t|\mathcal{L}_t)$. Assuming that the pixels $\mathbf{z}_{1,t},\ldots,\mathbf{z}_{n,t}$ in $\mathcal{I}_t$ are conditionally independent given $\mathcal{L}_t$, the observation likelihood is finally computed by

$$p(\mathcal{I}_t|\mathcal{L}_t) = \prod_{k=1}^{N}\prod_{j=1}^{n_k} p(\mathbf{z}_{j,t}|\tilde{\boldsymbol{\varphi}}^k_{b,t})^{(1-\ell_{j,t})} p(\mathbf{z}_{j,t}|\tilde{\boldsymbol{\varphi}}^k_{f,t})^{\ell_{j,t}}, \tag{10}$$

where n_k is the number of pixels in the k-th block.

3.2. Prior of Segmentation Given Pose

For the alternating procedure between segmentation and pose tracking, feedback from pose tracking needs to be incorporated for label estimation. In this work, pixel-wise foreground response map $\mathcal{Y}_t$ plays this role, and the prior of segmentation given human pose introduced in Eq. (5) is given by

$$p(\mathcal{L}_t|\mathcal{Y}_t) \propto \underbrace{\prod_{i=1}^{n}\prod_{j=1}^{n} p(\ell_{i,t}|\ell_{j,t})}_{\text{spatial smoothness}} \; \underbrace{\prod_{i=1}^{n} p(\ell_{i,t}|\mathcal{Y}_t)}_{\text{pose consistency}}, \tag{11}$$

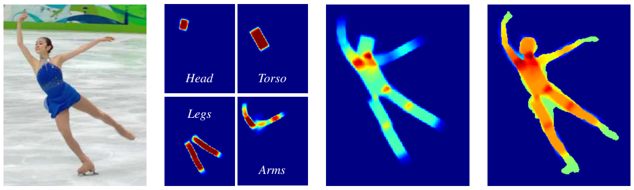
where spatial smoothness term defines the relationship between adjacent pixels and pose consistency term corresponds to the coherency between the labels from segmentation and pose tracking.

The spatial smoothness term penalizes inconsistent labels of neighboring pixels[3] in the MRF framework as

$$\prod_{i=1}^{n}\prod_{j=1}^{n} p(\ell_{i,t}|\ell_{j,t}) \propto \exp\left(\sum_{i=1}^{n}\sum_{j=1}^{n} (\ell_{i,t}\ell_{j,t} + (1-\ell_{i,t})(1-\ell_{j,t}))\right), \tag{12}$$

where $\ell_{i,t}$ denotes the label of the i-th pixel adjacent to the j-th pixel.

On the other hand, pixel-wise foreground response map, $\mathcal{Y}_t$, is estimated based on the response maps of individual body parts obtained from pose tracking as well as the segmentation mask inferred in the previous iteration. To construct the foreground response map, we marginalize the responses of all body parts in a 2D image space, and then normalize the marginalized responses with the maximum value. Also, the label likelihood given the foreground response, $p(\ell_{i,t}|\mathcal{Y}_t)$ in Eq. (11), is computed by distance transform [12] as



(a) (b) (c) (d)

Figure 3: Foreground response map. (a) Input image. (b) Detector response of each body part (head, torso, upper and lower legs, upper and lower arms). (c) Marginalized part response in 2D image space. (d) Foreground response map is generated by the sum of the marginalized response and the segmentation mask in the previous iteration.

$$p(\ell_{i,t}=1|\mathcal{Y}_t) = \frac{1}{1+\exp(\nu\cdot d(i,\mathcal{Y}_t))}, \tag{13}$$

$$p(\ell_{i,t}=0|\mathcal{Y}_t) = 1 - p(\ell_{i,t}=1|\mathcal{Y}_t), \tag{14}$$

where $d(i,\mathcal{Y}_t)$ is the distance from the i-th pixel to the closest pixel that have non-zero value in $\mathcal{Y}_t$, and ν controls the penalty of foreground misdetection. Figure 3 visualizes the construction of the foreground response map.

4. Pose Tracking

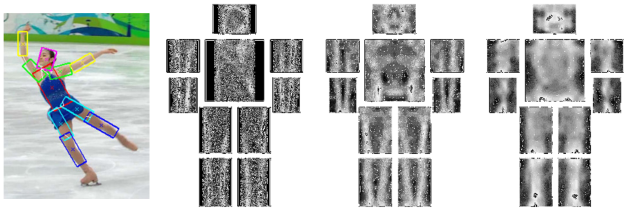
Pose tracking sequentially estimates $p(\mathbf{X}_t|\mathcal{L}_t,\mathcal{I}_{1:t})$, posterior distribution over the current human body configuration $\mathbf{X}_t$ at time t given segmentation result $\mathcal{L}_t$ and all image evidences $\mathcal{I}_{1:t}$. The posterior probability is factorized by Bayesian filtering as

$$p(\mathbf{X}_t|\mathcal{L}_t,\mathcal{I}_{1:t}) \propto p(\mathcal{L}_t,\mathcal{I}_t|\mathbf{X}_t)p(\mathbf{X}_t|\mathcal{I}_{1:t-1}), \tag{15}$$

where $p(\mathcal{L}_t,\mathcal{I}_t|\mathbf{X}_t)$ is the likelihood of image evidence and segmentation given a particular body part configurations, and $p(\mathbf{X}_t|\mathcal{I}_{1:t-1})$ is the prior of $\mathbf{X}_t$. The configurations of body parts are estimated by tracking-by-detection paradigm, in which the responses of part detectors serve as observation for tracking.

We restrict search space for each body part to the intersection of foreground area and predicted region corresponding to each part area. The restricted search region significantly reduces ambiguities and false positive detections caused by the features similar to human body part in background clutter. Also, it obviously alleviates computational burden to obtain body part responses. The inference

[3] We used the standard four neighborhood system.



(a) (b) (c) (d)

Figure 4: The specific and general part models. (a) Result of pose estimation at frame 125 in *Skating* sequence. (b) Specific shape model for *Skating* sequence at $t = 125$. (c) General shape model (d) General region model.

for finding the optimal body part configuration at time t is performed by sum-product belief propagation [18].

4.1. Likelihood by Multiple Part Detectors

The likelihood term, $p(\mathcal{L}_t, \mathcal{I}_t|\mathbf{X}_t)$, is a crucial component in our formulation. It is approximated with a product of likelihoods of individual body parts, which is given by

$$p(\mathcal{L}_t, \mathcal{I}_t|\mathbf{X}_t) \propto \prod_{i=1}^{m} p(\mathcal{L}_t|\mathbf{x}_{i,t}) p(\mathcal{I}_t|\mathbf{x}_{i,t}, \mathcal{L}_t), \tag{16}$$

where $p(\mathcal{L}_t|\mathbf{x}_{i,t})$ and $p(\mathcal{I}_t|\mathbf{x}_{i,t}, \mathcal{L}_t)$ denote region and shape likelihood of each part, respectively. These likelihoods are computed by the responses of three different kinds of filters, which model general shape and region, and specific shape.

The general shape and region models, which are built based on training dataset [19], represent common characteristics in shape and region of each part, respectively. These general models detect parts with unknown appearance in an image, but are too generic to handle scene-specific variations. To overcome the limitation, we incorporate an additional part model, which represents scene-specific shape of a part adaptively. The examples of the three part models are illustrated in Figure 4. Note that the specific shape model is updated online while the general models are fixed.

The shape likelihood of the part $\mathbf{x}_{i,t}$ is computed by the convolution of shape model filters and local foreground edge map with respect to $\mathbf{x}_{i,t}$, which is given by

$$\begin{aligned} p(\mathcal{I}_t|\mathbf{x}_{i,t}, \mathcal{L}_t) \propto \exp\big(-\eta_i \left(f(\mathcal{I}_t; \mathbf{x}_{i,t}, \mathcal{L}_t) * \rho_i^{\mathrm{g}}\right) \\ -(1-\eta_i)\left(f(\mathcal{I}_t; \mathbf{x}_{i,t}, \mathcal{L}_t) * \rho_{i,t}^{\mathrm{s}}\right)\big), \end{aligned} \tag{17}$$

where $\rho_{i,t}^{\mathrm{s}}$ denotes the specific shape model at time t, ρ_i^{g} denotes general shape model, and $*$ indicates convolution operator. The relative weights of the two models are determined by η_i. To minimize the effect of the features similar to human body parts within background region, a local foreground edge map $f(\cdot)$ is obtained by the intersection of segmentation mask and edge map in the local region, which is defined by

$$f(\mathcal{I}_t; \mathbf{x}_{i,t}, \mathcal{L}_t) = E(\mathcal{I}_t^i) \cap \mathcal{L}_t^i, \tag{18}$$

where $\mathcal{I}_t^i$ and $\mathcal{L}_t^i$ denote image and foreground mask for the area defined by $\mathbf{x}_{i,t}$ and $E(\mathcal{I}_t^i)$ denotes an edge map extracted within $\mathcal{I}_t^i$.

On the other hand, the region likelihood of the part $\mathbf{x}_{i,t}$ is given by

$$p(\mathcal{L}_t|\mathbf{x}_{i,t}) \propto \exp\left(-c \cdot \left(\mathcal{L}_t^i * \rho_i^{\text{region}}\right)\right), \tag{19}$$

where c is a constant and ρ_i^{region} is the general region model of the i-th part. As mentioned earlier, the detector responses for each body part, which correspond to Eq. (17) and (19), are used to construct foreground response map $\mathcal{Y}_t$.

4.2. Pose Prediction

The prediction $p(\mathbf{X}_t|\mathcal{I}_{1:t-1})$ is composed of a spatial prior on the relative position between parts and a temporal prior of an individual body part as

$$\begin{aligned} p(\mathbf{X}_t|\mathcal{I}_{1:t-1}) \propto \prod_{(i,j)\in\mathcal{E}} p(\mathbf{x}_{i,t}|\mathbf{x}_{j,t}) \times \\ \int_{\mathbf{X}_{t-1}} \prod_{i=1}^{m} p(\mathbf{x}_{i,t}|\mathbf{x}_{i,t-1}) p(\mathbf{X}_{t-1}|\mathcal{I}_{1:t-1})\, d\mathbf{X}_{t-1}, \end{aligned} \tag{20}$$

where $p(\mathbf{x}_{i,t}|\mathbf{x}_{j,t})$ corresponds to the spatial prior based on the kinematic constraints between two adjacent parts (*e.g.* upper arms must be connected to torso), $p(\mathbf{x}_{i,t}|\mathbf{x}_{i,t-1})$ denotes the temporal prior for an individual part, and $p(\mathbf{X}_{t-1}|\mathcal{I}_{1:t-1})$ is the posterior of the pose parameters at the previous time step $t-1$.

The spatial prior on body part configurations is based on a tree structure and represents the kinematic dependencies between body parts. To deal with various changes of pose and appearance in image, we adopt the spatial prior model by discrete binning [19], which is given by

$$\prod_{(i,j)\in\mathcal{E}} p(\mathbf{x}_{i,t}|\mathbf{x}_{j,t}) \propto \exp\left(\sum_{(i,j)\in\mathcal{E}} \beta_i^{\top} bin(\mathbf{x}_{i,t} - \mathbf{x}_{j,t})\right), \tag{21}$$

where $\mathcal{E}$ is a set of edges representing kinematic relationship between parts, $bin(\cdot)$ is a vectorized form of spatial and angular histogram bins, and β_i is a model parameter that favors certain spatial and angular bins for the i-th part with respect to the j-th part.

The temporal prior $p(\mathbf{x}_{i,t}|\mathbf{x}_{i,t-1})$ of each body part predicts the distribution of current configuration $\mathbf{x}_{i,t}$ given the previous configuration $\mathbf{x}_{i,t-1}$, which is modeled as

$$p(\mathbf{x}_{i,t}|\mathbf{x}_{i,t-1}) = \mathcal{N}\left(\mathbf{x}_{i,t}|\mathbf{x}_{i,t-1}, \mathbf{\Sigma}_i\right), \tag{22}$$

where $\mathcal{N}(\cdot|\boldsymbol{\mu}, \mathbf{\Sigma})$ denotes a 4D Gaussian distribution with mean $\boldsymbol{\mu}$ and diagonal covariance matrix $\mathbf{\Sigma}$.

Algorithm 1: Joint segmentation and pose tracking

```
Input: $\mathcal{I}_t$, $\boldsymbol{\varphi}^k_{\xi,t-1}$, $\mathbf{X}_{t-1}$, $\rho^{\mathrm{s}}_{i,t-1}$
Output: $\mathcal{L}_t$, $\boldsymbol{\varphi}^k_{\xi,t}$, $\mathbf{X}_t$, $\rho^{\mathrm{s}}_{i,t}$
iterate
    FG/BG Segmentation:
        FG/BG model propagation (Eq. (9))
        Labeling by graph-cut to maximize Eq. (5) based on
            – FG/BG likelihood (Eq. (10))
            – Prior of segmentation given pose (Eq. (11))
    Pose Tracking:
        Estimation of pose posterior (Eq. (15)) based on
            – Pose likelihood (Eq. (16)): combination of shape
              (Eq. (17)) and region (Eq. (19)) likelihoods
            – Pose prediction (Eq. (20))
until converge;
Model Update:
    FG/BG model update (Eq. (23))
    Specific shape model update (Eq. (24))
```

5. Model Updates

After the iterative procedure at each frame, we obtain foreground/background labels and human body configuration. To propagate the labels and pose parameters accurately, foreground/background models and specific shape model should be updated in each frame based on the converged segmentation labels. The foreground and background models are recursively updated using the propagated models from the previous time step $t-1$ and the observations in the current time step t, which are given by

$$p(\mathbf{z}_t|\hat{\boldsymbol{\varphi}}^k_{\xi,t}) = \tau_{\text{seg}} \cdot p(\mathbf{z}_t|\tilde{\boldsymbol{\varphi}}^k_{\xi,t}) + \frac{1-\tau_{\text{seg}}}{n_\xi}\sum_{i=1}^{n_\xi} K_{\mathbf{H}}(\mathbf{z}_t - \mathbf{y}^i_{\xi,t}), \quad (23)$$

where τ_{seg} is a forgetting factor.

The specific shape model of each body part is also updated incrementally based on the local foreground edge map at the current time step t, which is given by

$$\hat{\rho}^{\mathrm{s}}_{i,t} = \tau_{\text{pose}} \cdot \hat{\rho}^{\mathrm{s}}_{i,t-1} + (1-\tau_{\text{pose}}) \cdot f(\mathcal{I}_t; \mathbf{x}_{i,t}, \mathcal{L}_t), \quad (24)$$

where τ_{pose} is a forgetting factor for specific shape model.

The pseudo code of overall joint segmentation and pose tracking algorithm is described in Algorithm 1.

6. Experiments

We evaluated the performance of our algorithm qualitatively and quantitatively in real videos downloaded from public websites. Our results are compared with existing methods for foreground/background segmentation and pose estimation such as [15, 8, 19]. Since the proposed technique combines segmentation and pose estimation, the two subproblems are evaluated separately.

6.1. Datasets and Evaluation Methods

We employ five challenging videos captured by a moving camera for experiment. All the sequences involve various pose changes and substantial camera motions. Additionally, *Skating* and *Dunk* sequences contain scale and illumination changes, *Pitching* and *Javelin* sequences suffer from self-occlusions, and *Jumping* sequence involves scale changes as well as self-occlusion.

To evaluate the performance of foreground/background segmentation, we compute precision, recall, and F-measure based on ground-truth annotation. On the other hand, pose estimation is evaluated by the *Percentage of Correctly estimated body Parts* (PCP) [7]. By this measure, an estimated body part is considered correct if its segment endpoints lie within the fraction of the length of the ground-truth segment; a larger value means a more accurate result. We compute PCP values for individual body parts, and the performance of entire human body is estimated based on the average of the PCP measures of all body parts. Note that we annotated foreground/background masks and human body poses manually in every 5 frame.

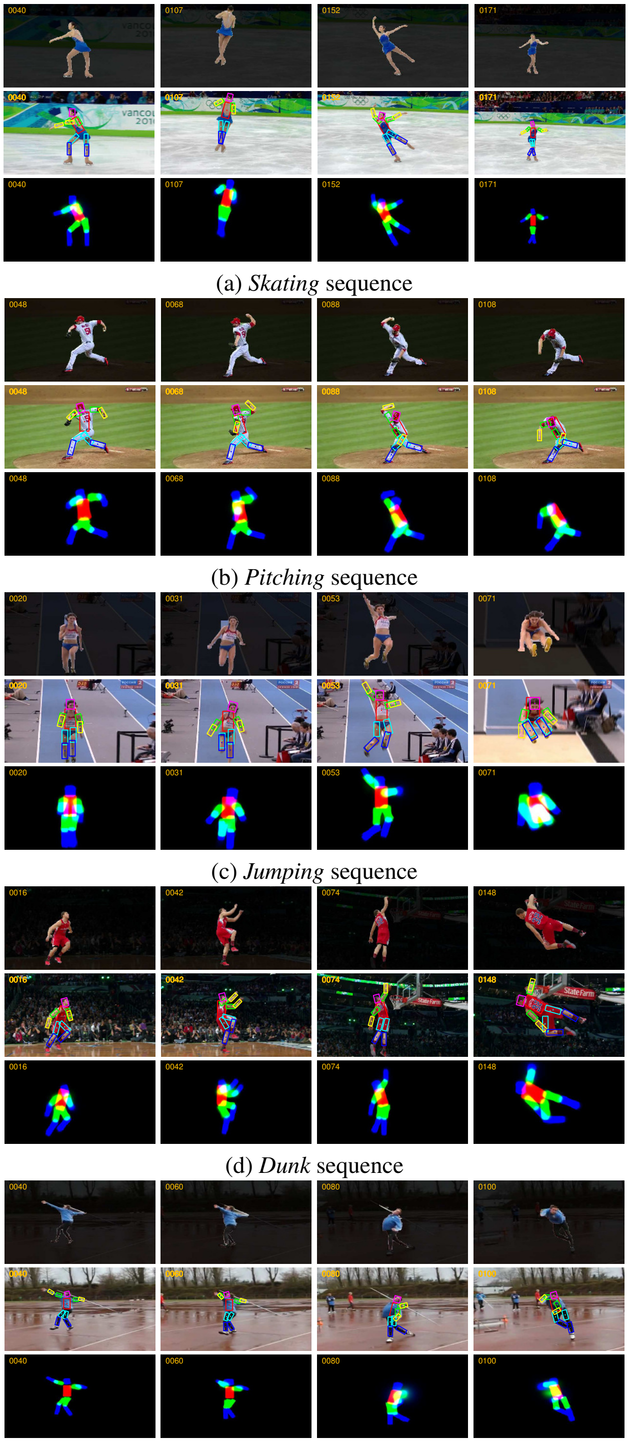
6.2. Results

We present our foreground/background segmentation and human pose tracking results in Figure 5. The experimental results illustrate that our algorithm produces accurate and robust outputs in the presence of background clutter, significant pose variations, fast camera motions, occlusions, scale changes, and so on. Note that our algorithm successfully handles dynamic human body configurations involving foreshortening, full stretching, and self occlusion.

To demonstrate the effectiveness of our joint estimation algorithm, we first compare our foreground/background segmentation algorithm with other methods such as progressive pruning [8], motion segmentation, and our algorithm with segmentation only. Then, our pose tracking algorithm is also compared with progressive pruning [8], iterative learning algorithm with our foreground/background segmentation [19], and our algorithm with pose tracking only. As illustrated in Figure 6, our joint estimation algorithm performs better than all other methods significantly and is robust to the background features similar to human body parts.

The quantitative performance of foreground/background segmentation algorithm are summarized in Figure 7, where our algorithm is compared with a simple motion segmentation, our algorithm with segmentation only and an existing techniques [8]. Our algorithm outperforms all other algorithms substantially in all three tested measures.

The quantitative performance of human pose estimation is evaluated based on the PCP-curves, which are presented in Figure 8. The PCP-curves are obtained by varying the fraction (PCP-threshold) from 0 to 1, where the threshold



(a) *Skating* sequence

(b) *Pitching* sequence

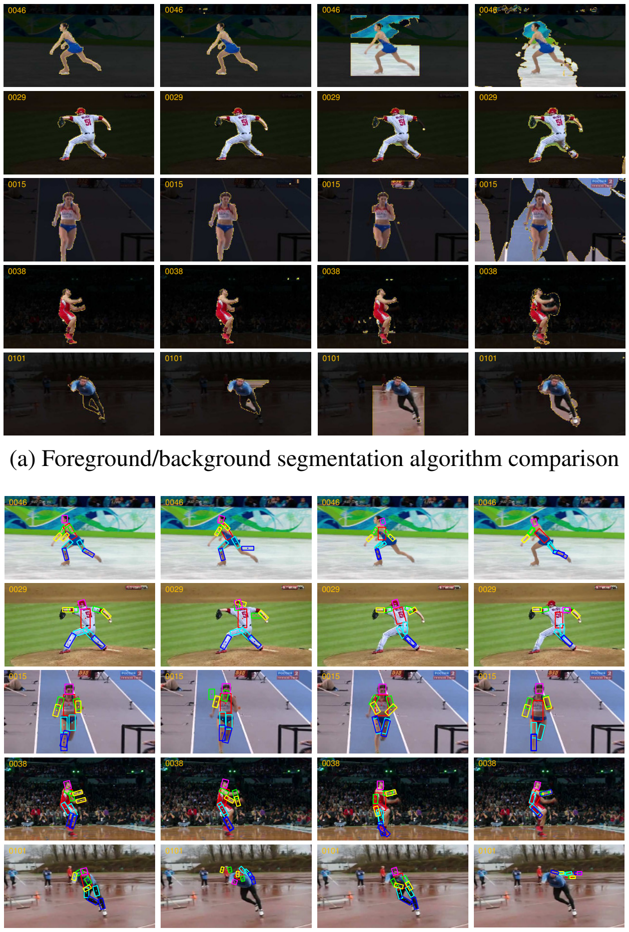
(c) *Jumping* sequence

(d) *Dunk* sequence

(e) *Javelin* sequence

Figure 5: Results of all video sequences. **(Row1)** Foreground/background segmentation. **(Row2)** Pose estimation. **(Row3)** Augmented likelihood maps for pose estimation.

denotes error allowance for correctness. Our algorithm performs better than all other algorithms throughout the range of all threshold values.



(a) Foreground/background segmentation algorithm comparison

(b) Pose estimation algorithm comparison

Figure 6: Qualitative comparisons of all five video sequences. (a) **(Col1)** Our full algorithm. **(Col2)** Our foreground/background segmentation only. **(Col3)** Eichner *et al*. [8]. **(Col4)** Motion segmentation. (b) **(Col1)** Our full algorithm. **(Col2)** Our pose tracking only. **(Col3)** Eichner *et al*. [8]. **(Col4)** Ramanan [19] with our foreground/background segmentation.

7. Conclusion

We presented a unified probabilistic framework to perform foreground/background segmentation and human pose tracking jointly in an on-line manner. The proposed algorithm presents outstanding segmentation and pose estimation performance by mutual interactions between the two complementary subsystems; they alternate each other and improve the quality of solution in each iteration. We showed the robustness of our foreground/background segmentation and pose tracking algorithms to background clutter, pose changes, object scale changes, and illumination variations through qualitative and quantitative validation in real videos.

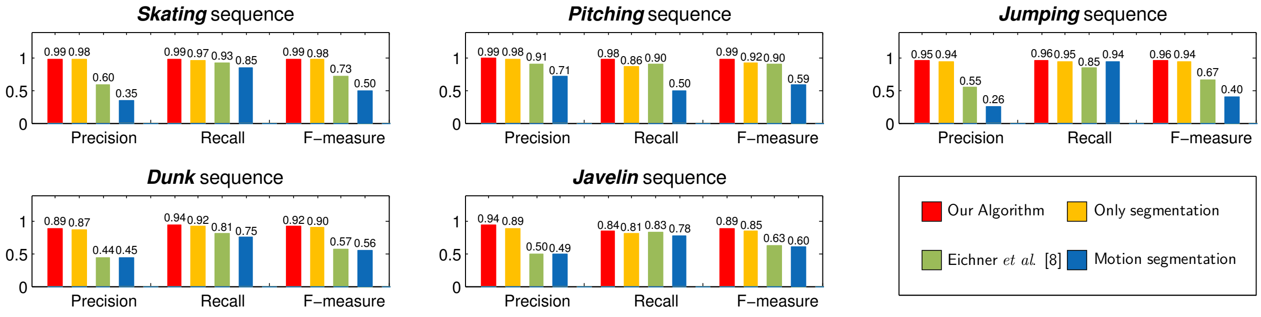
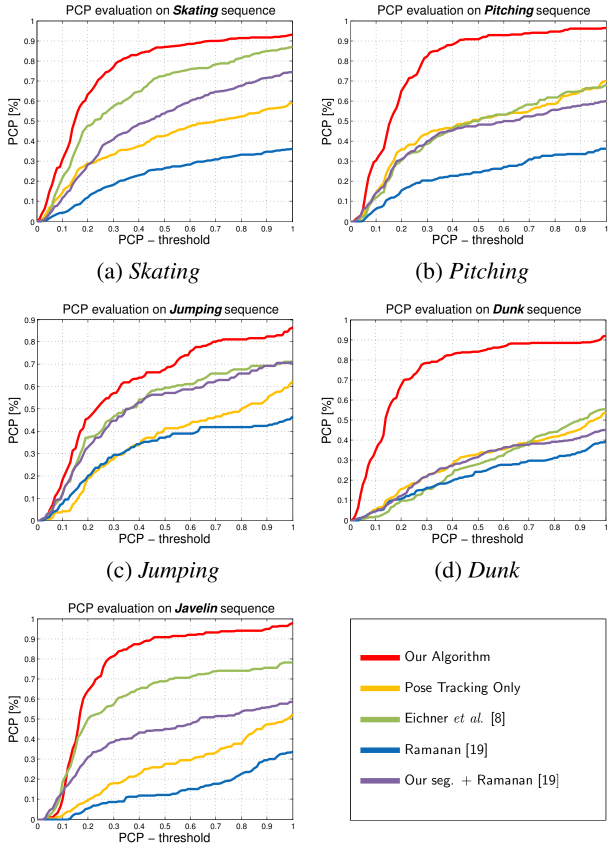


Figure 7: Quantitative performance evaluation results of foreground/background segmentation.



(a) *Skating*

(b) *Pitching*

(c) *Jumping*

(d) *Dunk*

(e) *Javelin*

Figure 8: Quantitative performance evaluation results of pose estimation evaluation by PCP curves.

References

[1] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, 2009. 2

[2] A. Blake, C. Rother, M. Brown, P. Perez, and P. Torr. Interactive image segmentation using an adaptive gmmrf model. In *ECCV*, 2004. 3

[3] Y. Boykov and M.-P. Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in n-d images. In *ICCV*, 2001. 3

[4] T. Brox, B. Rosenhahn, and D. Cremers. Contours, optic flow, and prior knowledge: cues for capturing 3D human motion in videos. In *Human Motion - Understanding, Modeling, Capture, and Animation*. Springer, 2007. 2

[5] C. Chen and G. Fan. Hybrid body representation for integrated pose recognition, localization and segmentation. In *CVPR*, 2008. 2

[6] X. Cui, J. Huang, S. Zhang, and D. N. Metaxas. Background subtraction using low rank and group sparsity constraints. In *ECCV*, 2012. 1

[7] M. Eichner and V. Ferrari. Better appearance models for pictorial structures. In *BMVC*, 2009. 6

[8] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari. 2d articulated human pose estimation and retrieval in (almost) unconstrained still images. *IJCV*, 99:190–214, 2012. 2, 6, 7

[9] A. Elqursh and A. M. Elgammal. Online moving camera background subtraction. In *ECCV*, 2012. 1

[10] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61:55–79, 2005. 2

[11] V. Ferrari, M. Marin-Jiminez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, 2008. 2

[12] P. Kohli, J. Rihan, M. Bray, and P. Torr. Simultaneous segmentation and pose estimation of humans using dynamic graph cuts. *IJCV*, 79:285–298, 2008. 2, 4

[13] M. Kumar, P. Ton, and A. Zisserman. Obj Cut. In *CVPR*, 2005. 2

[14] S. Kwak, T. Lim, W. Nam, B. Han, and J. H. Han. Generalized background subtraction based on hybrid inference by belief propagation and bayesian filtering. In *ICCV*, 2011. 1

[15] T. Lim, B. Han, and J. H. Han. Modeling and segmentation of floating foreground and background in videos. *Pattern Recognition*, 45(4):1696–1706, 2012. 1, 3, 6

[16] J. C. Niebles, B. Han, and L. Fei-Fei. Efficient extraction of human motion volumes by tracking. In *CVPR*, 2010. 1

[17] J. C. Niebles, B. Han, A. Ferencz, and L. Fei-Fei. Extracting moving people from internet videos. In *ECCV*, 2008. 1

[18] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1998. 5

[19] D. Ramanan. Learning to parse images of articulated bodies. In *NIPS*, 2006. 2, 5, 6, 7

[20] C. Rother, V. Kolmogorov, and A. Blake. ”GrabCut”: interactive foreground extraction using iterated graph cuts. In *ACM SIGGRAPH*, 2004. 2

[21] Y. Sheikh, O. Javed, and T. Kanade. Background subtraction for freely moving cameras. In *ICCV*, 2009. 1

[22] H. Wang and D. Koller. Multi-level inference by relaxed dual decomposition for human pose segmentation. In *CVPR*, 2011. 2

[23] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011. 2