Pose Estimation and Segmentation of People in 3D Movies

Karteek Alahari, Guillaume Seguin, Josef Sivic, Ivan Laptev

Pose Estimation and Segmentation of People in 3D Movies

Karteek Alahari[1,*] Guillaume Seguin[2,*] Josef Sivic[1,*] Ivan Laptev[1,*]

[1]Inria [2]École Normale Supérieure

Abstract

We seek to obtain a pixel-wise segmentation and pose estimation of multiple people in a stereoscopic video. This involves challenges such as dealing with unconstrained stereoscopic video, non-stationary cameras, and complex indoor and outdoor dynamic scenes. The contributions of our work are two-fold: First, we develop a segmentation model incorporating person detection, pose estimation, as well as colour, motion, and disparity cues. Our new model explicitly represents depth ordering and occlusion. Second, we introduce a stereoscopic dataset with frames extracted from feature-length movies "StreetDance 3D" and "Pina". The dataset contains 2727 realistic stereo pairs and includes annotation of human poses, person bounding boxes, and pixel-wise segmentations for hundreds of people. The dataset is composed of indoor and outdoor scenes depicting multiple people with frequent occlusions. We demonstrate results on our new challenging dataset, as well as on the H2view dataset from (Sheasby et al. ACCV 2012).

1. Introduction

Stereoscopic feature-length movies provide a large amount of readily available video footage of realistic indoor and outdoor dynamic scenes. Our goal is to automatically analyze people in such challenging videos. In particular, we aim to produce a pixel-wise segmentation, estimate the pose, and recover the partial occlusions and relative depth ordering of people in each frame, as illustrated in Figure 1. Our motivation is three-fold. First and foremost, we wish to develop a mid-level representation of stereoscopic videos suitable for subsequent video understanding tasks such as recognition of actions and interactions of people [42]. Human behaviours are often distinguished only by subtle cues (e.g., a hand contact) and having a detailed and informative representation of the video signal is an initial step towards their interpretation. Second, disparity cues available from stereoscopic movies are expected to improve results of person segmentation and pose estimation. Such results, in turn, can be used as a (noisy) supervisory signal for learning person segmentation and pose estimation in monocular videos or still images [1, 17, 26, 40]. Finally, pose estimation and segmentation of people will also support interactive annotation, editing, and navigation in stereo videos [16, 22], which are important tasks in post-production and home video applications.

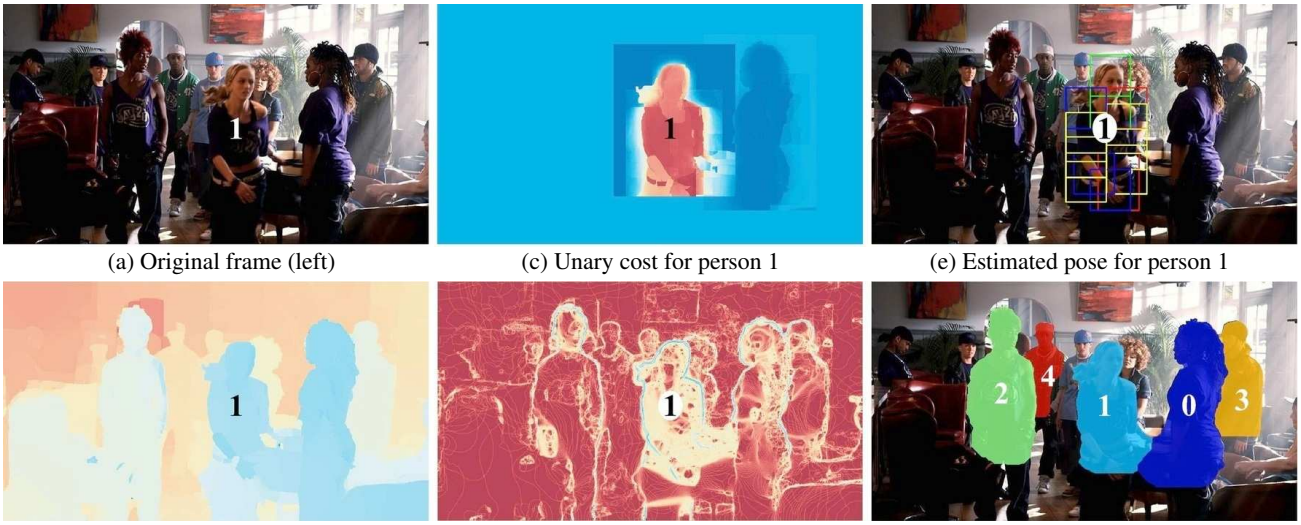
Given the recent success of analyzing people in range data from active sensors, such as Microsoft Kinect [27, 31], and a plethora of methods to estimate pixel-wise depth from stereo pairs [2], the task at hand may appear solved. However, depth estimates from stereo videos are much noisier than range data from active sensors, see Figure 1 for an example. Furthermore, we aim to solve sequences outside of the restricted "living-room" setup addressed by Kinect. In particular, our videos contain complex indoor and outdoor scenes with multiple people occluding each other, and are captured by a non-stationary camera.

In this paper, we develop a segmentation model in the context of stereoscopic videos, which addresses challenges such as: (i) handling non-stationary cameras, by incorporating explicit person detections and pose estimates; (ii) the presence of multiple people in complex indoor and outdoor scenarios, by incorporating articulated person-specific segmentation masks (Section 3) and explicitly modelling occlusion relations among people; and finally (iii) the lack of accurate stereo estimates, by using other cues, such as colour and motion features. We cast the problem as a discrete labelling task involving multiple person labels, devise a suitable cost function (Section 2), and optimize it efficiently (Section 4). We evaluate the proposed model on the new Inria 3DMovie Dataset with challenging realistic dynamic scenes from two stereoscopic feature-length movies "StreetDance" [Giwa and Pasquini, 2010] and "Pina" [Wenders, 2011]. Additionally, we compare our results on the Humans in Two Views (H2view) dataset [30] (Section 5).

1.1. Related work

The problem of segmenting a stereo video into foreground-background regions has been addressed for a teleconferencing set-up in [21]. The sequences considered in this work involved only one or two people seated in front of a webcam, i.e., a restricted set of poses and at best, simple occlusions. Also, no articulated person model was used. Many recent works have investigated the use of stereo (or depth) signal in tasks such as person detection [19, 29, 32, 35], pose estimation [31], and segmenta-

*WILLOW project-team, Département d'Informatique de l'École Normale Supérieure, ENS/Inria/CNRS UMR 8548, Paris, France.



(a) Original frame (left) (c) Unary cost for person 1 (e) Estimated pose for person 1

(b) Disparity (d) Smoothness cost (f) Segmentation result

Figure 1. *Illustration of the steps of our proposed framework on a sample frame from the movie "StreetDance". We compute the disparity map (b) from the stereo pair. Occlusion-aware unary costs based on disparity and articulated pose mask are computed for all the people detected in the scene. In (c) we show the unary cost for the person labelled* 1. *Pairwise smoothness costs computed from disparity, motion, and colour features are shown in (d). The range of values in (b,c,d) is denoted by the red (low) - blue (high) spectrum of colours. The estimated articulated pose for person 1 is shown in (e). (f) shows the final segmentation result, where each colour represents a unique person, and the numbers denote the front (0) to back (4) ordering of people.* (**Best viewed in colour.**)

tion [21]. Given the success in these individual tasks, the challenge now is to take a step further, and look at these problems jointly in scenarios involving multiple interacting people (see Figure 1).

In addition to the significant progress in human pose estimation in still images and videos [15, 18, 28, 41], there has been some work in joint pose estimation and segmentation [20, 24, 30, 36]. However, these works are limited to cases involving isolated people, and extending them to situations with multiple interacting people is not straightforward. Recently, a model for joint reasoning about poses of multiple upright people has been proposed in [12]. This framework does not output segmentations of people, but can be adapted to do so. We experimentally compare [12] to results of our method in Section 5.

The proposed method not only computes a segmentation of people and their poses, but also estimates their depth ordering and occlusion. This relates our work to layered representations [23, 33, 34, 37, 39]. For example, Kumar *et al*. [23] demonstrate detection and tracking of articulated models of walking people and animals. The method assumes consistent appearance and a locally affine parametric motion model of each object part. Layered representations can also explicitly model occlusions and depth ordering [33]. In a similar spirit, Yang *et al*. [40] apply a layered model to recover occlusions and depth ordering of multiple overlapping object detections in one image. These methods do not, however, recover any pose information, as we do.

Contributions. The main contribution of this paper is a multi-person segmentation model for stereoscopic video data. The model incorporates person detections and learnt articulated pose-specific segmentation masks, as well as colour, motion, and stereo disparity cues. The model also explicitly represents depth ordering and occlusion. As a second contribution, we introduce a new annotated dataset with more than 400 pixel-wise segmentations of people in frames extracted from a stereoscopic movie. We demonstrate the benefits of the proposed approach on this new challenging data.

2. The Segmentation Model

We segment a given stereoscopic video sequence extracted from a 3D movie into regions representing individual people. Figure 1 illustrates an overview of our method on a sample frame from a video. Here we consider a stereo pair (only the left image is shown in the figure), estimate the disparity for every pixel, and use it together with person detections, colour and motion features, and pose estimates, to segment individual people, as shown in Figure 1(f).

We initialize our model using automatically obtained person detections and assign every detection to a person, i.e., we assume a one-to-one mapping between people and detections. Each pixel i in the video takes a label from the set $\mathcal{L} = \{0, 1, \ldots, L\}$, where $\{0, 1, \ldots, L-1\}$ represents the set of person detections and the label L denotes the "background".[1] The cost of assigning a person (or background) label, from the set $\mathcal{L}$, to every pixel i, $E(\mathbf{x}; \mathbf{\Theta}, \tau)$, is given by:

[1] With a slight abuse of terminology we refer to image regions that correspond to other objects, which may lie in front of or behind people, as background.

$$E(\mathbf{x};\mathbf{\Theta},\tau) = \sum_{i\in\mathcal{V}} \phi_i(x_i;\mathbf{\Theta},\tau) + \sum_{(i,j)\in\mathcal{E}} \phi_{ij}(x_i,x_j) + \sum_{(i,k)\in\mathcal{E}^t} \phi^t_{ij}(x_i,x_k), \quad (1)$$

where $\mathcal{V} = \{1,2,\ldots,N\}$ denotes the set of pixels in the video. The pairwise spatial and temporal neighbourhood relations among pixels are represented by the sets $\mathcal{E}$ and $\mathcal{E}^t$ respectively. The temporal neighbourhood relations are obtained from the motion field [25] computed for every frame. The function $\phi_i(x_i;\mathbf{\Theta},\tau)$ is the cost of a pixel i in $\mathcal{V}$ taking a label x_i in $\mathcal{L}$. It is characterized by pose parameters $\mathbf{\Theta} = \{\Theta^0,\Theta^1,\ldots,\Theta^{L-1}\}$ and disparity parameters $\tau = \{\tau^0,\tau^1,\ldots,\tau^{L-1}\}$, where Θ^l and τ^l represent the pose and disparity parameters for a person label l respectively. Note that the pose and disparity parameters vary across time. However, for brevity, we drop this dependency on t in our notation.

The function $\phi_{ij}(x_i,x_j)$ is a spatial smoothness cost of assigning labels x_i and x_j to two neighbouring pixels i and j. Similarly, $\phi^t_{ij}(x_i,x_k)$ is a temporal smoothness cost. Given the parameters $\mathbf{\Theta}$ and τ, minimizing the cost (1) to obtain an optimal labelling $\mathbf{x}^* = \arg\min_{\mathbf{x}} E(\mathbf{x};\mathbf{\Theta},\tau)$, segments the video into regions corresponding to distinct people or background. However, in our problem, we also aim to optimize over the set of pose and disparity parameters. In other words, we address the problem of estimating $\mathbf{x}^*$, the optimal segmentation labels, and $\mathbf{\Theta}^*,\tau^*$, the optimal pose and disparity parameters as: $\{\mathbf{x}^*,\mathbf{\Theta}^*,\tau^*\} = \arg\min_{\mathbf{x},\mathbf{\Theta},\tau} E(\mathbf{x};\mathbf{\Theta},\tau)$, where $E(\mathbf{x};\mathbf{\Theta},\tau)$ is the cost of label assignment $\mathbf{x}$, given the pose and disparity parameters, as defined in (1). Given the difficulty of optimizing E over the joint parameter space, we simplify the problem and first estimate pose parameters $\mathbf{\Theta}$ independently of $\mathbf{x}$ and τ as described in Section 3. Given $\mathbf{\Theta}$, we then solve for $\mathbf{x},\tau$ as:

$$\{\mathbf{x}^*,\tau^*\} = \arg\min_{\mathbf{x},\tau} E(\mathbf{x},\tau;\mathbf{\Theta}). \quad (2)$$

Further details are provided in Section 4. A graphical representation of our model is shown in Figure 2. The remainder of this section defines the unary costs, which are computed independently in every frame, and the spatio-temporal pairwise costs in (1).

Occlusion-based unary costs. Each pixel i takes one of the person or background labels from the label set $\mathcal{L}$. Building on the approach of [40], we define occlusion-based costs corresponding to these labels, $\phi_i(x_i = l;\mathbf{\Theta},\tau), l$ in $\mathcal{L}$, as a function of likelihoods β^l, computed for each label l, as follows:

$$\phi_i(x_i = l;\mathbf{\Theta},\tau) = -\log P(x_i = l|\mathbf{\Theta},\tau), \quad (3)$$

$$\text{where } P(x_i = l|\mathbf{\Theta},\tau) = \beta^l_i \prod_{0\le m<l} (1-\beta^m_i). \quad (4)$$

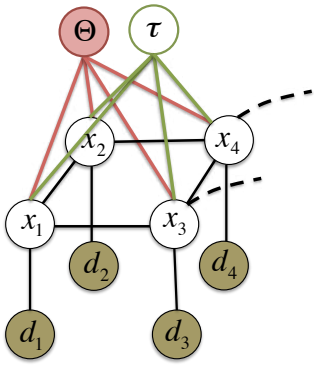


Figure 2. *A graphical illustration of our model, where the observed variables are shaded. Each pixel in the video is represented as a variable d_i in the graph. For clarity, we show 4 pixels from a frame, and 2 of the temporal links (dashed line), which connect pixels in one frame to the next. The person label x_i and disparity parameters τ are inferred given the image features d_i, and the pose parameters Θ.*

Here, β^l_i is the likelihood of pixel i taking the person (or background) label l. The label likelihood β^l is then formed by composing the likelihoods β^l_i, for all pixels $i \in \mathcal{V}$ in the image. In essence, β^l is a soft mask, which captures the likelihood for one person detection. It can be computed using the pose estimate of the person, and image features such as disparity, colour, and motion, as discussed in the following section. To account for the fact that the people in a scene may be occluding each other, we accumulate the label likelihoods in a front-to-back order[2] as in (4). This makes sure that pixel i is likely to take label l, if it has sufficiently strong evidence for label l (i.e., β^l_i is high), and also has low evidence for other labels m, which correspond to people in front (i.e., β^m_i is low for all labels $m < l$).

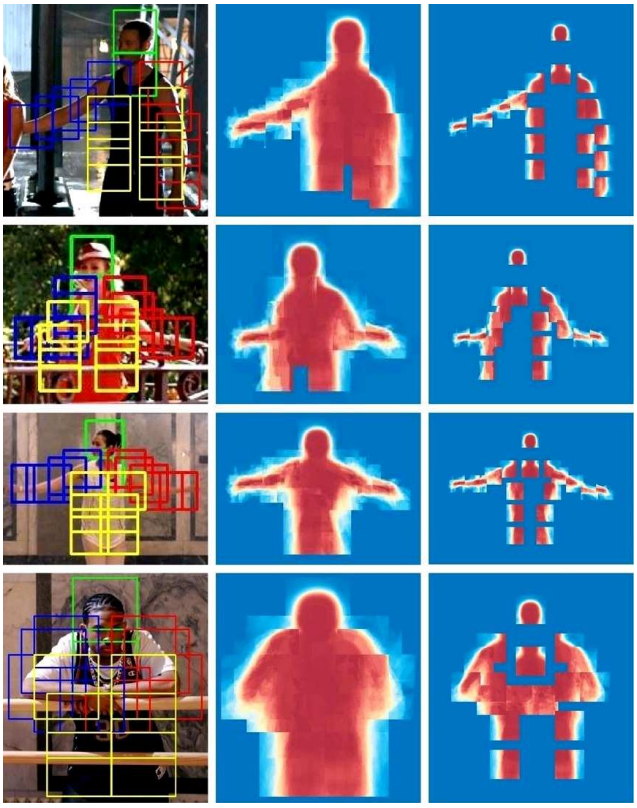
Label likelihood β^l. Given a person detection and its corresponding pose estimate Θ^l, the problem of computing the label likelihood β^l can be viewed as that of segmenting an image into person *vs*. background. However, we do not make a binary decision of assigning pixels to either the person or the background label. This computation is more akin to generating a *soft* likelihood map for each pixel taking a particular person label. We define this using disparity and pose cues as: $\beta^l_i = \alpha^l\psi_p(\Theta^l) + (1-\alpha^l)\psi_d(\tau^l)$, where $\psi_p(\Theta^l)$ is an articulated pose mask described in Section 3, $\psi_d(\tau^l)$ is a disparity likelihood, and α^l is a mixing parameter that controls the relative influence of pose and disparity. The disparity potential is given by:

$$\psi_d(d_i;\tau^l,\sigma^l) = \exp\left(-\frac{(d_i-\tau^l)^2}{2(\sigma^l)^2}\right), \quad (5)$$

where d_i is the disparity value computed at pixel i. The disparity potential is a Gaussian characterized by mean τ^l and standard deviation σ^l, which together with the pose parameter Θ^l determines the model for person l. We set $\beta^L_i = 0.9$ for all the pixels for the background label L.

Smoothness cost. In some cases, the disparity cue used for computing the unary costs may not be very strong or may "leak" into the background (see example in Figure 5).

[2]The order is determined by the disparity parameters τ as discussed in Section 4.



(a) Estimated pose (b) Pose mask (c) Per-mixture masks

Figure 3. *Estimated poses and masks on sample frames. Given a pose estimate (a), we compute a pose-specific mask (b) using per-mixture part masks learnt from manually segmented training data. In (c) we show a scaled version of the masks, doubling the actual distances between part masks. This visually explains how each per-mixture mask is contributing to the final mask. In (b,c), the cost for a pixel to take a person label is denoted by the red (low) - blue (high) spectrum of colours.* (**Best viewed in colour.**)

We introduce colour and motion features into the cost function (1), as part of the smoothness cost, to alleviate such issues. The smoothness cost, $\phi_{ij}(x_i, x_j)$, of assigning labels x_i and x_j to two neighbouring pixels i and j takes the form of a generalized Potts model [7] given by:

$$\phi_{ij}(x_i, x_j) = \begin{cases} \lambda_1 \exp(\frac{-(d_i-d_j)^2}{2\sigma_c^2}) + \lambda_2 \exp(\frac{-||\mathbf{v}_i-\mathbf{v}_j||_2^2}{2\sigma_v^2}) \\ +\lambda_3 \exp(\frac{-(pb_i-pb_j)^2}{2\sigma_p^2}) & \text{if } x_i \neq x_j, \\ 0 & \text{otherwise,} \end{cases} \quad (6)$$

where λ_1, λ_2, λ_3, σ_c, σ_v and σ_p are parameters of the model. The function $(d_i - d_j)^2$ measures the difference in disparity between pixels i and j. The motion vector at pixel i is denoted by $\mathbf{v}_i \in \mathbb{R}^2$, and $||\mathbf{v}_i - \mathbf{v}_j||_2$ is the norm of the motion vector difference of pixels i and j. The function $(pb_i - pb_j)^2$ measures the difference of colour features (Pb feature values [4]) of pixels i and j. The temporal smoothness cost $\phi_{ij}^t(x_i, x_k)$ is simply a difference of Pb features values for two pixels i and k connected temporally by the motion vector $\mathbf{v}_i$.

Thus far we have discussed the model given person detections, their pose and disparity parameters. In what follows, we will describe our method for detecting people, their poses, and the likelihood computed from them (Section 3). We then provide details of the inference scheme for determining the parameters and the pixel-wise segmentation (Section 4).

3. Estimating an Articulated Pose Mask

The aim here is to obtain an articulated pose segmentation mask for each person in the image, which can act as a strong cue to guide the pixel-wise labelling. We wish to capture the articulation of the human pose as well as the likely shape and width of the individual limbs, torso, and head in the specific pose. We build here on the state-of-the-art pose estimator of Yang and Ramanan [41], and extend it in the following two directions. First, we incorporate disparity as input to take advantage of the available stereo signal. Second, we augment the output to provide an articulated pose-specific soft-segmentation mask learnt from manually annotated training data.

Person detection and tracking. We obtain candidate bounding boxes of individual people and track them throughout the video. Detections are obtained from the deformable part-based person detector [14]. We found this to perform empirically better than using the articulated pose estimator [41] for detecting people. To benefit from the stereo signal, we trained a joint appearance and disparity model by concatenating appearance and disparity features into one representation. We use HOG [10] computed on images converted to grayscale, similar to [41], as appearance features. The disparity features are obtained by computing HOG on disparity maps. Our HOG feature representation for disparity maps is similar to that used in [32, 35] for person/pedestrian detection. We track the person detections computed in each frame of the video, and interpolate to fill-in any missing detections, similar to [13]. The detections are also smoothed temporally.

Pose estimation from appearance and disparity. We estimate the pose of the person within each person detection bounding box. We restrict our pose estimation models to upper body poses, which are more commonly found in movie data. Again, to benefit from the stereo video, we extract both appearance and disparity features in the frame. The advantage is that some edges that are barely visible in the image, e.g., between people in similar clothing, can be more pronounced in the disparity map. We use HOG features for both appearance and disparity, as described above for person detection. We introduce specific mixtures for handling occlusion, as in [11], into the pose estimation framework of [41].

The model is represented as a set of parts, where a part refers to a patch centered on a joint or on an interpolated point on a line connecting two joints. For example, we have

one part for an elbow, one for a wrist, and two parts between the elbow and the wrist, spread uniformly along the arm length. We use a model with 18 parts. The set of parts includes 10 annotated joints, *head, neck, 2 shoulders, 2 elbows, 2 wrists, 2 hips*, together with 8 interpolated parts. Further, each part is characterized by a set of mixtures. The mixture components for an elbow part, for example, can be interpreted as capturing different appearances of the elbow as the pose varies, including occlusions by other limbs or people, that are explicitly labelled in the training data. We learn up to eight mixture components, among which one or two are dedicated to handle occlusions, for each part.

Articulated pose mask ψ_p. The output of the pose estimator is the location of the individual parts in the frame as shown in Figure 3(a). To obtain a pose-specific mask we learn an average mask for each mixture component for each part. This is achieved by applying the trained pose-estimator on a training set of people with manually provided pixel-wise segmentations. All training masks, where mixture component c of part k is detected, are then rescaled to a canonical size and averaged together to obtain the mean mask $m_{kc}(i)$. The value at pixel i in the mean mask counts the relative frequency that this pixel belongs to the person. An illustration of masks for individual parts and mixture components is shown in Figure 3(c).

At test time, given an estimated pose with an instantiated mixture component c^* for a part k, the likelihood for the person, $\psi_p(\Theta, i)$ at pixel i, is obtained by laying out and composing the articulated masks m_{kc^*} for all the parts. If, at pixel i, multiple masks overlap, we take the maximum as $\psi_p(\Theta, i) = \max_k m_{kc^*}(i)$. We found that taking the max was beneficial for person segmentation targeted in this paper as it suppresses internal edges between body parts, such as a hand positioned in front of the torso. An illustration of the articulated pose masks for various examples is shown in Figure 3. Note how the part masks can capture fine variations in the shape and style of the pose.

4. Inference

In the previous section we have outlined how we compute the pose parameters Θ^l and the corresponding articulated pose mask for each person l. Poses are estimated independently for each person and fixed throughout the rest of the inference procedure described next. The aim is to compute the optimal disparity parameters τ^* and pixel labels $\mathbf{x}^*$ given the pose parameters Θ, as described by the minimization problem (2). It is well known that minimizing multi-label functions such as $E(\mathbf{x}; \mathbf{\Theta}, \tau)$, which corresponds to the segmentation problem, given the pose and disparity parameters, is in itself NP-hard (for the type of smoothness cost we use) [6]. The additional complexity of optimizing over disparity parameters τ further adds to the challenge. We propose a two-step strategy, where we first: (i) estimate the optimal disparity parameters τ^* using an approximation to (2), without the pairwise terms; and then (ii) obtain the pixel labels $\mathbf{x}^*$ with the estimated (and now fixed) parameters τ^* by minimizing the full cost (1). These two steps are detailed below.

Obtaining disparity parameters. The estimation of the set of disparity parameters τ for all the people in a frame can be succinctly written as:

$$\tau^* = \arg\min_{\{\tau\}} \tilde{E}(\tilde{\mathbf{x}}; \mathbf{\Theta}, \tau), \tag{7}$$

where we further approximate the original cost function (1) by only using unary and ignoring the pairwise terms[3] as $\tilde{E}(\mathbf{x}; \mathbf{\Theta}, \tau) = \sum_{i \in \mathcal{V}} \phi_i(x_i; \mathbf{\Theta}, \tau)$. Note that for this modified cost function, the optimal pixel labelling $\tilde{\mathbf{x}}$ for a given τ can be obtained independently for each pixel as $\tilde{x}_i = \arg\min_{m \in \mathcal{L}} \tilde{E}(x_i = m, \mathbf{\Theta}, \tau)$. Further, the disparity parameter τ is inversely related to depth, and determines the front-to-back order of people in a frame. Thus, this minimization problem (7) explores various combinations of the relative order of people in a frame by optimizing over $\{\tau\}$. The set of possible disparity parameter values for each person can still be large, and exploring the exponentially many combinations for all the people in the frame may not be feasible. To address this issue, we obtain and optimize over a small set of (up to 3) candidates $\{\tau^l\}$, for each person l.[4] Note that the disparity parameters are estimated jointly for all the people in the scene.

Person segmentation. With the estimated disparity (and pose) parameters, we compute the unary and smoothness costs, and use the efficient α-expansion algorithm [8] to optimize (1). This assigns every pixel a person or background label from the set $\mathcal{L}$.

5. Experiments

In this section we detail our method for extracting disparity maps from stereo videos, and report results for person detection, pose estimation, and segmentation. We present results on challenging sequences, involving multiple people, extracted from two stereoscopic movies, "StreetDance" and "Pina". Our new annotated Inria 3DMovie Dataset used for evaluation in this paper is available on the project website [3]. We also compare our method with [30] on the H2view dataset.

Disparity estimation. We estimate the disparity for each frame independently. A joint estimation of motion and disparity from video is also possible [38]. We assume that the stereo pair is approximately rectified, i.e., for a particular pixel in view 1 the corresponding pixel in view 2 lies close

[3] We note that this is a reasonable approximation, as τ only directly affects the unary cost ϕ_i in (1).

[4] Using a thresholded pose mask, we compute mean disparity μ^l of all the pixels within, and set $\{\tau^l\} = \{\mu^l, \mu^l \pm \sigma^l\}$. The parameter σ^l is set according to a linear decreasing function of μ^l.

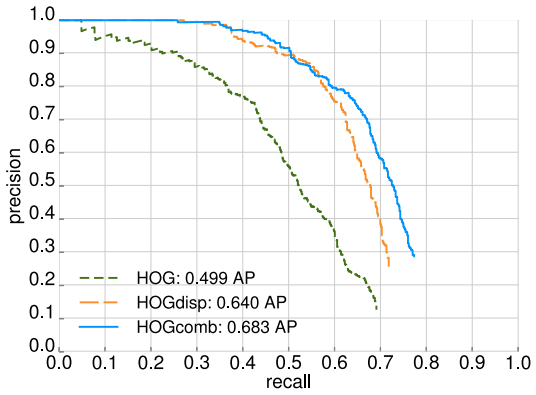


Figure 4. *Precision-recall curves for person detection using appearance* (HOG) *and disparity* (HOGdisp) *based detectors, as well as the jointly trained appearance & disparity based detector* (HOGcomb)*. Note that the detectors using disparity cues have an almost perfect precision until around 35% recall.*

to the same horizontal scan-line. We use the method of Ayvaci *et al*. [5] for estimating disparities. It performs a search in a 2D window, and thus can deal with small vertical displacements. Such an approach alleviates the need to rectify the stereo pairs, which is in itself a challenging task in the context of stereo movies. This is partly due to the fact that, in stereo movies, parameters of the camera rig, such as the focal length, baseline or verging angle can change across shots and even during a shot [22]. The 2D search also helps to compensate for some unmodelled effects, e.g., due to radial distortion. Furthermore, the ability to handle occlusions explicitly resulted in better disparity maps than other methods, such as [25].

We use the horizontal component of the estimated disparity field in our formulation. We follow [35] and work with disparity values directly rather than depth to avoid problems with infinite depth, and amplifying errors at small disparities. Estimating the dense disparity field for a single stereo pair of 960×540 pixels takes approximately 30 seconds on a modern GPU using the implementation from [5].

Datasets. We trained our person detection and pose estimation methods on an annotated dataset from the feature-length movie "StreetDance". We evaluated our methods on video clips from two movies, namely "StreetDance" and "Pina". The movie "StreetDance" was split into two parts (roughly in the middle), from which we selected the training and test frames, respectively. The training set is composed of 520 annotated person bounding boxes and poses from 265 frames. Negative training data is extracted from 247 images with no people, taken from the training part of the movie, and from stereo pairs shot using the Fuji W3 camera, which were harvested from Flickr.

The test set for evaluating person detection has 638 person bounding boxes in 193 frames, among which a few do not contain any people. Given the cost of annotating poses and pixel-wise segmentation, we evaluated them on a smaller subset of 180 frames, containing 464 annotated person segmentations and poses.

Person detection and pose estimation. We report person detection and pose estimation results for models trained using: (i) standard HOG extracted from grayscale images (HOG), (ii) HOG extracted from disparity maps (HOGdisp), and (iii) joint appearance and disparity based model, using the concatenation of the two features (HOGcomb). First, we compare the three person detection models using standard metrics from the PASCAL VOC development kit 2011 [1]. Precision-recall curves are shown in Figure 4, with corresponding average precision (AP) values. It shows that the disparity-based detector (HOGdisp) improves over the appearance-based detector (HOG). Combining the two representations (HOGcomb) further increases person detection performance. Pose estimation is evaluated using the standard percentage of correctly estimated body parts (PCP) score [12]. A body part is deemed correct if the two joints it links are within a given radius of their ground truth position, where the radius is a percentage of the ground truth length of the part. The PCP values for PCP-threshold 0.5 are reported in Table 1. The jointly trained pose estimator (HOGcomb) outperforms the individual estimators. We observe that the head and the torso body parts are localized with high accuracy. Furthermore, combining appearance and disparity cues improves the lower arm localization by about 4%.

Segmenting multiple people. In our experiments, we used the following parameter values: $\lambda_1 = 6.3$, $\lambda_2 = 6$, $\lambda_3 = 2.7$, $\sigma_c^2 = 0.025$, $\sigma_v^2 = 0.01$, $\sigma_p^2 = 0.025$, which were set empirically, and fixed for the evaluation. A quantitative evaluation of the segmentation model is shown in Table 2. Sample video sequence results are available on the project webpage [3]. In Table 2 we compare four variants of our approach and two baseline methods. The first one (Proposed + no mask) refers to the case where the label likelihood $\beta_i^l = \psi_d$, i.e., there is no influence of pose on the segmentation. In other words, this method uses disparity features, but not the pose information. The second method (Proposed + uni mask) incorporates a person location likelihood, which is computed by averaging ground truth segmentations of people from the training data (after rescaling them to a standard size) into a single non-articulated "universal" person mask – an approach inspired by the successful use of such masks in the past [40]. We use this as the *person* likelihood ψ_p, and combine it with disparity likelihood ψ_d, as explained in Section 2. The third variant (Proposed + pose mask) incorporates the articulated pose mask, described in Section 3. Our complete model (Proposed + pose mask + temporal) introduces temporal smoothness across frames.

For the "Colour only" baseline, we used a colour-based model for the unary costs without the disparity potential.

	[41]	HOG	HOGdisp	HOGcomb
Head & Torso	0.989	0.989	0.991	**0.998**
Upper arms	0.839	0.856	0.869	**0.889**
Lower arms	0.518	0.559	0.535	**0.594**
Global	0.782	0.802	0.799	**0.827**

Table 1. *Evaluating pose estimation. We report global* PCP *scores as and individual values for three types of body parts, as in [41]. We also evaluate the upper-body model from [41] trained on the Buffy dataset. The numbers in bold indicate the best performance. The combination of appearance and disparity features* (HOGcomb) *outperforms the individual estimators* (HOG, HOGdisp).

These costs were computed from colour histograms for each label, similar to [7]. The success of this model certainly depends on the regions used for computing the histograms. We used the result obtained by segmenting in the disparity space, i.e., "Proposed + no mask", as these regions. We believe that this provides a reasonable estimate for the label potentials. The background histogram was computed with bounding boxes harvested from regions with no person detections. Another baseline we compared with, is derived from the recent work of [12], which computes the poses of multiple people in a scene. We use the (monocular) person *vs*. background segmentation performed as part of this formulation on our dataset.

Intersection *vs*. union measure [1] is used to evaluate our segmentation results. From Table 2, the method "Proposed + pose mask + temporal" performs better than the others. The poor performance of the *Colour only* method, despite a reasonable initialization for the histograms, is perhaps an indication of the difficulty of our dataset. From Figures 1 and 5 we note that the person *vs*. background distinction is not very marked in the colour feature space. Furthermore, these images appear to be captured under challenging lighting conditions. The temporal smoothness terms in (1) reduce flickering artifacts in the segmentation, as shown in our video results [3]. Other methods [9] to propagate segmentations from a few key frames of the video onto others can also be used.

Results on a few sample frames for our temporal model are shown in Figure 5. On a 960×540 frame the method takes about 13s to detect and track people, 8s to estimate the pose of each person, and 30s per frame to perform the segmentation with our non-optimized Matlab implementation. Naturally, the success of our approach depends on the quality of person detections. Here, we operate in a high-precision mode, at the expense of missing difficult examples, e.g., heavily occluded people. The most prominent failure modes of our method are: (i) challenging poses very different from training data; and (ii) cases where the disparity signal is noisy for people far away from the camera (e.g., Figure 5, row 1).

H2view dataset. The H2view dataset [30] was acquired using a static stereo rig, in combination with a Kinect active sensor. Ground truth poses and segmentations are available for 7 test video sequences, with a total of 1598 annotated frames. It is, however, restricted to a single person setup and hence has no inter-person occlusions. We tested our model (trained on 3D movies) directly on this dataset, without any further tuning, and analyzed the segmentation quality using the evaluation code from [30]. As our method models only the upper body, we cropped the ground truth, our results, and those from [30] just above the hips, and considered only upper body (rather than full body) segmentation. Our method achieves a segmentation overlap score of 0.825 compared to their 0.735. The time for segmentation is 6s per frame for this dataset, which contains 512×384 frame sequences of a single person.

Method	Int. vs Union
Proposed + no mask	0.278
Proposed + uni mask	0.543
Proposed + pose mask	0.779
Proposed + pose mask + temporal	**0.802**
Baselines:	
Colour only	0.630
Eichner *et al*. [12]	0.662

Table 2. *Evaluation of pixel-wise person segmentation. We used the intersection vs. union score, which is the overall percentage of pixels correctly classified. Our method, which uses disparity, colour, and motion features, along with pose likelihoods performs better than the others, notably 14% compared to the baseline [12].*

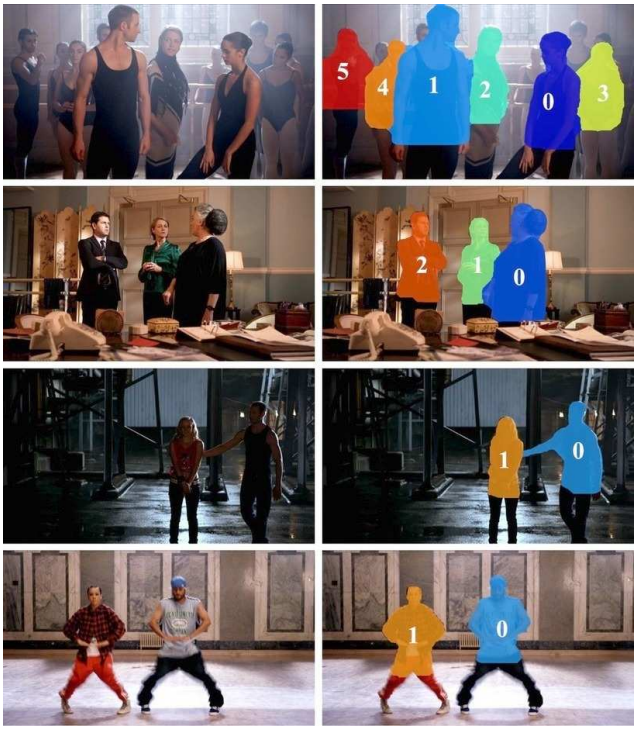
6. Discussion

We have developed a model for segmentation of people in stereoscopic movies. The model explicitly represents occlusions, incorporates person detections, pose estimates, and can recover the depth ordering of people in the scene. The results suggest that disparity estimates from stereo video, while noisy, can serve as a strong cue for localizing and segmenting people. The results also demonstrate that a person's pose, incorporated in the form of an articulated pose mask, can provide a strong shape prior for segmentation. The developed representation presents a building block for modelling and recognition of human actions and interactions in 3D films.

Acknowledgements. The authors would like to thank Jean Ponce for helpful suggestions. This work is partly supported by the Quaero Programme, funded by OSEO, the MSR-INRIA laboratory, ERC grant Activia, Google and the EIT ICT Labs.

References

[1] http://pascallin.ecs.soton.ac.uk/challenges/voc/voc2011, 2011.
[2] http://vision.middlebury.edu/stereo/, 2013.
[3] http://www.di.ens.fr/willow/research/stereoseg, 2013.
[4] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *PAMI*, 2011.



(a) Original image (b) Segmentation result

Figure 5. *Qualitative results on images from the movie "StreetDance". Each row shows the original image and the corresponding segmentation. Rows 1 and 2 demonstrate successful handling of occlusion between several people. The method can also handle non-trivial poses, as shown by Rows 3 and 4. The segmentation results are generally accurate, although some inaccuracies still remain on difficult examples. For instance, in Row 1, the segmentation is leaking into background for persons 3 and 5, due to the weak disparity cue for these people far away from the camera. The numbers denote the front (low values) to back (high values) ordering of people.* (**Best viewed in colour.**)

[5] A. Ayvaci, M. Raptis, and S. Soatto. Sparse occlusion detection with optical flow. *IJCV*, 97(3), 2012.

[6] E. Boros and P. L. Hammer. Pseudo-boolean optimization. *Discrete Applied Mathematics*, 2002.

[7] Y. Boykov and M.-P. Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In *ICCV*, 2001.

[8] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 23(11):1222–1239, 2001.

[9] I. Budvytis, V. Badrinarayanan, and R. Cipolla. Semi-supervised video segmentation. In *CVPR*, 2011.

[10] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[11] C. Desai and D. Ramanan. Detecting actions, poses, and objects with relational phraselets. In *ECCV*, 2012.

[12] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari. 2d articulated human pose estimation and retrieval in (almost) unconstrained still images. *IJCV*, 2012.

[13] M. Everingham, J. Sivic, and A. Zisserman. "Hello! My name is... Buffy" – automatic naming of characters in TV video. In *BMVC*, 2006.

[14] P. F. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32(9), 2010.

[15] K. Fragkiadaki, H. Hu, and J. Shi. Pose from flow and flow from pose. In *CVPR*, 2013.

[16] D. Goldman, C. Gonterman, B. Curless, D. Salesin, and S. Seitz. Video annotation, navigation, and composition. In *UIST*, 2008.

[17] V. Gulshan, V. Lempitsky, and A. Zisserman. Humanising grabcut: Learning to segment humans using the kinect. In *IEEE Workshop on Consumer Depth Cameras for Computer Vision, ICCV*, 2011.

[18] S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR*, 2011.

[19] C. Keller, M. Enzweiler, M. Rohrbach, D. Llorca, C. Schnorr, and D. Gavrila. The benefits of dense stereo for pedestrian detection. *IEEE Trans. Intelligent Transportation Systems*, 2011.

[20] P. Kohli, J. Rihan, M. Bray, and P. Torr. Simultaneous segmentation and pose estimation of humans using dynamic graph cuts. *IJCV*, 2008.

[21] V. Kolmogorov, A. Criminisi, A. Blake, G. Cross, and C. Rother. Bi-layer segmentation of binocular stereo video. In *CVPR*, 2005.

[22] S. Koppal, C. Zitnick, M. Cohen, S. Kang, B. Ressler, and A. Colburn. A viewer-centric editor for 3d movies. *Computer Graphics and Applications*, 2011.

[23] M. P. Kumar, P. H. S. Torr, and A. Zisserman. Learning layered motion segmentations of video. In *ICCV*, 2005.

[24] L. Ladicky, P. H. S. Torr, and A. Zisserman. Human pose estimation using a joint pixel-wise and part-wise formulation. In *CVPR*, 2013.

[25] C. Liu. *Beyond Pixels: Exploring New Representations and Applications for Motion Analysis*. PhD thesis, MIT, 2009.

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

[27] X. Ren, L. Bo, and D. Fox. Rgb-(d) scene labeling: Features and algorithms. In *CVPR*, 2012.

[28] B. Sapp, D. Weiss, and B. Taskar. Parsing human motion with stretchable models. In *CVPR*, 2011.

[29] K. Schindler, A. Ess, B. Leibe, and L. Van Gool. Automatic detection and tracking of pedestrians from a moving stereo rig. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65(6):523–537, 2010.

[30] G. Sheasby, J. Valentin, N. Crook, and P. H. S. Torr. A robust stereo prior for human segmentation. In *ACCV*, 2012.

[31] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *CVPR*, 2011.

[32] L. Spinello and K. O. Arras. People detection in rgb-d data. In *IROS*, 2011.

[33] D. Sun, E. Sudderth, and M. Black. Layered image motion with explicit occlusions, temporal consistency, and depth ordering. In *NIPS*, 2010.

[34] P. H. S. Torr, R. Szeliski, and P. Anandan. An integrated bayesian approach to layer extraction from image sequences. *PAMI*, 2001.

[35] S. Walk, K. Schindler, and B. Schiele. Disparity statistics for pedestrian detection: Combining appearance, motion and stereo. In *ECCV*, 2010.

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

[37] J. Y. A. Wang and E. H. Adelson. Representing moving images with layers. *IEEE Trans. Image Processing*, 3(5):625–638, 1994.

[38] A. Wedel, C. Rabe, T. Vaudrey, T. Brox, U. Franke, and D. Cremers. Efficient dense scene flow from sparse or dense stereo data. In *ECCV*, 2008.

[39] J. Xiao and M. Shah. Motion layer extraction in the presence of occlusion using graph cuts. *PAMI*, 2005.

[40] Y. Yang, S. Hallman, D. Ramanan, and C. Fowlkes. Layered object models for image segmentation. *PAMI*, 2011.

[41] Y. Yang and D. Ramanan. Articulated pose estimation using flexible mixtures of parts. In *CVPR*, 2011.

[42] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010.