Pose Estimation and Segmentation of Multiple People in Stereoscopic Movies

Guillaume Seguin, Karteek Alahari, Josef Sivic, Ivan Laptev

Pose Estimation and Segmentation of Multiple People in Stereoscopic Movies

Guillaume Seguin, Karteek Alahari, Josef Sivic, and Ivan Laptev

Abstract—We describe a method to obtain a pixel-wise segmentation and pose estimation of multiple people in stereoscopic videos. This task involves challenges such as dealing with unconstrained stereoscopic video, non-stationary cameras, and complex indoor and outdoor dynamic scenes with multiple people. We cast the problem as a discrete labelling task involving multiple person labels, devise a suitable cost function, and optimize it efficiently. The contributions of our work are two-fold: First, we develop a segmentation model incorporating person detections and learnt articulated pose segmentation masks, as well as colour, motion, and stereo disparity cues. The model also 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 587 annotated human poses, 1158 bounding box annotations and 686 pixel-wise segmentations 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).

Index Terms— Person detection, Pose estimation, Segmentation, 3D data, Stereo movies.

1 Introduction

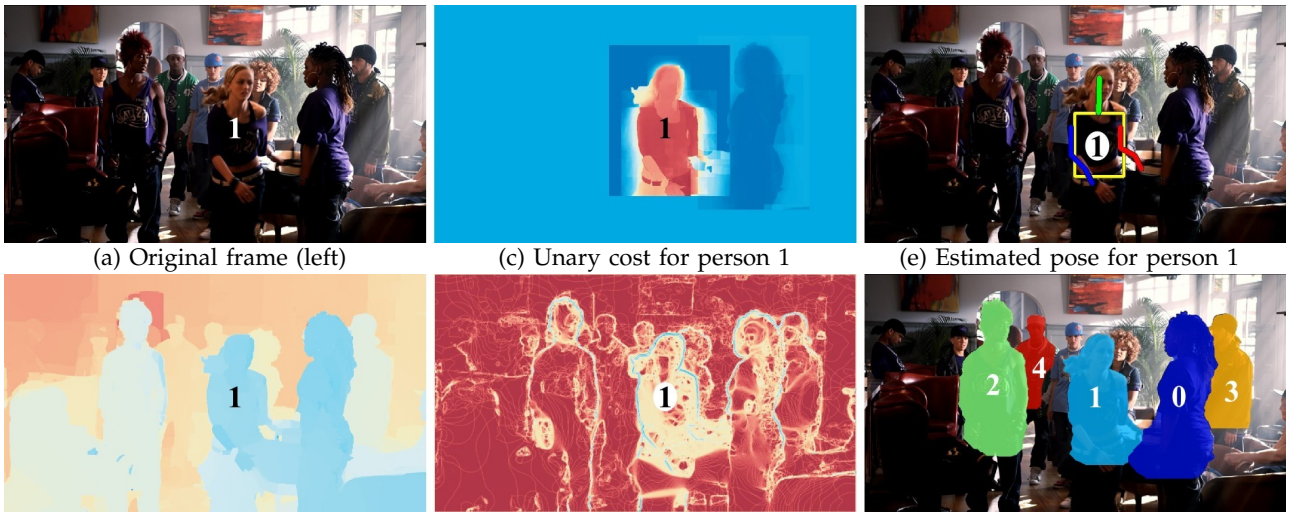
Detecting and segmenting multiple people in a video is a task of great interest in computer vision. We explore this problem in the context of stereoscopic feature length movies, which provide a large amount of readily available video footage of challenging indoor and outdoor dynamic scenes. Our goal 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 [1]. 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 a useful 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 [2], [3], [4], [5]. For instance, a single 90 minute feature length movie can provide more than 150,000 pixel-wise segmented frames. Finally, pose estimation and segmentation of people will also support interactive annotation, editing, and navigation in stereo videos [6], [7], 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 [8], [9], and a plethora of methods to estimate pixel-wise depth from stereo pairs [10], 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(b) 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 our new Inria 3DMovie dataset (Section 5) with challenging realistic dynamic scenes from two stereoscopic feature-length movies "StreetDance" [Giwa and Pasquini, 2010] and "Pina" [Wen-

- *Guillaume Seguin, Josef Sivic and Ivan Laptev are with Inria, WILLOW project-team, Département d'Informatique de l'Ecole Normale Supérieure, ENS/Inria/CNRS UMR 8548, Paris, France.*
- *Karteek Alahari is with the LEAR project-team, Inria Grenoble - Rhône-Alpes, Laboratoire Jean Kuntzmann, Grenoble, France.*

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



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

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

Fig. 1: *Illustration of the steps of our proposed framework on a sample frame (a) 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.**)

ders, 2011] (Section 6). Additionally, we present comparative evaluation of our method on the Humans in Two Views (H2view) dataset [11].

1.1 Related work

The problem of segmenting a stereo video into foreground-background regions has been addressed for a teleconferencing set-up in [12]. 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. Some recent works have also investigated the use of stereo (or depth) signal in tasks such as person detection [13], [14], [15], [16], pose estimation [9], and segmentation [12]. 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 the past few years, significant progress has been made on estimating human poses in still images and videos [17], [18], [19], [20], [21], [22]. For example, the work in [21] has successfully used motion cues to refine the pose estimation results from [19], highlighting the importance of incorporating additional cues. In [22] an improved pose appearance model is used in combination with more expressive body part representations. In addition to these works on human pose estimation, there has been some effort in addressing the problem of joint pose estimation and segmentation [11], [23], [24], [25]. For instance, the model presented in [11] uses disparity cues to perform human pose estimation and segmentation. The dataset used in [11] contains sequences recorded with a stationary camera in an indoor setting, and is limited to cases involving isolated people. In contrast, our new dataset contains multiple people in more challenging indoor and outdoor sequences obtained from non-stationary cameras. To handle such setups, we explicitly model the simultaneous presence of multiple people and their mutual occlusions.

Some recent works [25], [26] have considered the case of multiple people in a scene. The formulation proposed in [25] uses a candidate set of poses for finding a pixel-wise body part labelling of people in the scene. The lack of an occlusion term to model the interaction between multiple people makes this work inapplicable to the cases we consider in our evaluation. A model for joint reasoning about poses of multiple upright people has been proposed in [26]. However, this framework does not provide a segmentation of people in the scene.

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 [27], [28], [29], [30], [31]. For example, Kumar *et al*. [27] 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 or-

dering [28]. In a similar spirit, Yang *et al*. [5] 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 stereoscopic movies. We demonstrate the benefits of the proposed approach on this new challenging data. This paper is an extended version of [32].

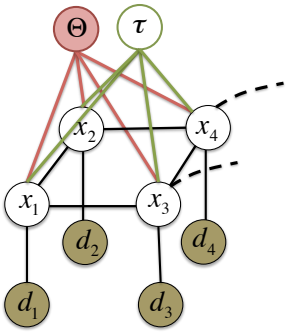


Fig. 2: *A graphical illustration of our model, where the observed variables are shaded. The variable d_i in the graph represents the features computed at each pixel i in the video. For clarity, we show 4 pixels from a frame, and 2 of the temporal links (dashed line), which connect pixels in one frame 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.*

2 The Segmentation Model

We aim to segment stereoscopic video sequences extracted from 3D movies 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:

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

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 [33] 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. The disparity parameters determine the front-to-back ordering of people in the scene, as discussed in more detail in Section 4.1. 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, minimization of the cost (1) to obtain an optimal labelling $\mathbf{x}^* = \arg\min_{\mathbf{x}} E(\mathbf{x}; \mathbf{\Theta}, \tau)$, results in segmentation of 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 $\mathbf{\Theta}^*$, τ^*, the optimal pose and disparity parameters as:

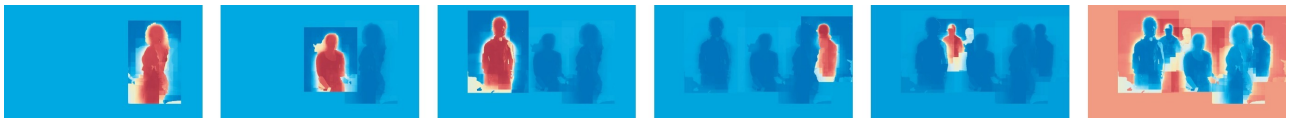
$$\{\mathbf{x}^*, \mathbf{\Theta}^*, \tau^*\} = \arg\min_{\mathbf{x}, \mathbf{\Theta}, \tau} E(\mathbf{x}; \mathbf{\Theta}, \tau), \quad (2)$$

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 $\mathbf{x}$, τ as:

$$\{\mathbf{x}^*, \tau^*\} = \arg\min_{\mathbf{x}, \tau} E(\mathbf{x}, \tau; \mathbf{\Theta}). \quad (3)$$

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,

1. We refer to image regions that correspond to objects other than people as background.



Person 0 | Person 1 | Person 2 | Person 3 | Person 4 | Background

Fig. 3: *Illustration of the occlusion-based unary costs for the example in Figure 1. From left to right we show the unary costs for persons labelled 0 – 4 and the background. The cost for a pixel to take a label (person or background) is denoted by the red (low) - blue (high) spectrum of colours. Here we observe the effect of accumulating the label likelihoods in a front-to-back order. For example, in the illustration for Person 4, a low cost (red) for taking label 4 is observed only for the pixels that are not occluded by the other people in front.* (**Best viewed in colour.**)

which are computed independently in every frame, and the spatio-temporal pairwise costs in (1).

2.1 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 [5], 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:

$$\phi_i(x_i = l; \mathbf{\Theta}, \tau) = -\log P(x_i = l | \mathbf{\Theta}, \tau), \tag{4}$$

$$\text{where} \quad P(x_i = l | \mathbf{\Theta}, \tau) = \beta_i^l \prod_{\{m | \tau^m > \tau^l\}} (1 - \beta_i^m). \tag{5}$$

Here, β_i^l is the likelihood of pixel i taking the person (or background) label l. Note that β_i^l's do not sum to one over the label set for any given pixel. The label likelihood over the entire image β^l is then formed by composing the likelihoods β_i^l, 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 as in (5). This order is determined by the disparity parameters τ we estimate (see Section 4). In other words, to compute the cost of a pixel taking a person label i, we consider all the other person labels that satisfy $\tau^m > \tau^i$, i.e., are in front of person i. This makes sure that pixel i is likely to take label l, if it has sufficiently strong evidence for label l (i.e., β_i^l is high), and also has low evidence for other labels m, which correspond to people in front of person l (i.e., β_i^m is low for all labels with $\tau^m > \tau^l$). Figure 3 shows an illustration of these costs on an example.

2.2 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. Note that 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:

$$\beta_i^l = (1 - \alpha^l)\ \psi_p(\Theta^l) + \alpha^l\ \psi_d(\tau^l), \tag{6}$$

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:

$$\psi_d(d_i; \tau^l, \sigma^l) = \exp\left(-\frac{(d_i - \tau^l)^2}{2(\sigma^l)^2}\right), \tag{7}$$

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 $\beta_i^L = 0.9$ for all the pixels for the background label L. The method for estimating the parameters τ^l and σ^l for person labels (i.e., for all $l \neq L$) is detailed in Section 4.

2.3 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 examples in Figure 12). 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 [34] given by:

$$\phi_{ij}(x_i, x_j) = \begin{cases} \lambda\ (\lambda_1 \exp(\frac{-(d_i - d_j)^2}{2\sigma_c^2}) + \lambda_2 \exp(\frac{-||\mathbf{v}_i - \mathbf{v}_j||_2^2}{2\sigma_v^2}) \\ \quad +\lambda_3 \exp(\frac{-(pb_i - pb_j)^2}{2\sigma_p^2})) & \text{if } x_i \neq x_j, \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

where λ, 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 ℓ_2-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 [35]) of pixels i and j. The temporal smoothness cost $\phi_{ij}^t(x_i, x_k)$ is simply

a difference of Pb feature 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 disparity 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 [19], 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.

3.1 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 [36]. We found this to perform empirically better than using the articulated pose estimator [19] for detecting people, as shown in Section 6.2. 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 [37] computed on images converted to grayscale as appearance features. The disparity features are obtained by computing HOG on disparity maps. This is done by first converting the disparity map into a grayscale image by linearly mapping the disparity range to [0,1]. We then compute HOG on this grayscale image. Our HOG feature representation for disparity maps is similar to that used in [15], [16] for person/pedestrian detection. The intuition is that HOG robustly captures the *changes* in the disparity rather than the actual disparity values, which can vary from scene to scene. Our joint appearance and disparity based detector is then applied to each frame in the video sequence independently. We also compute point tracks, which start at a frame and continue until some later frame, over the entire sequence with the Kanade-Lucas-Tomasi tracker [38]. Point tracks that lie within each detection result are used to fill-in any missing detections by interpolating the location of the bounding box and also to smooth the detections temporally [39].

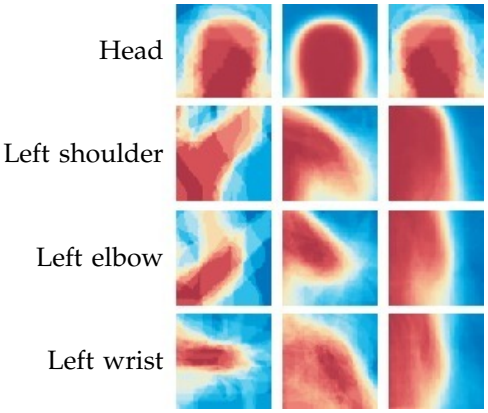
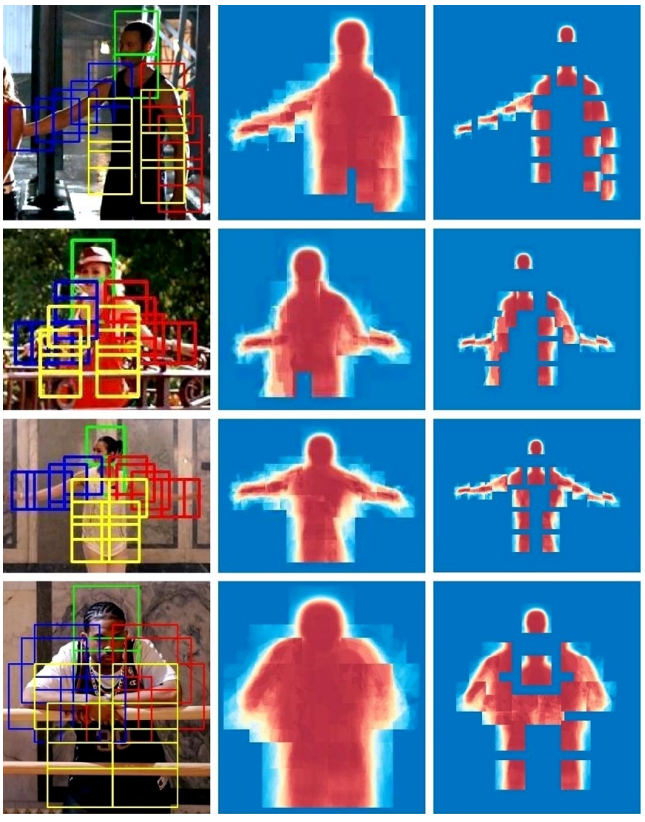


Fig. 4: *Articulated pose masks for three mixture components are shown for some of the body parts. The pose masks for each part capture a different configuration of the pose. For instance, the masks for "Left wrist" show three different locations of the lower arm: stretched out, partially bent over the shoulder, and lying by the torso.* (**Best viewed in colour.**)

3.2 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 (in contrast to [19], [40], which use appearance features only). 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 [40], into the pose estimation framework of [19].

In this framework, the model is represented as a set of K 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. We refer the reader to [19] for more details on the training procedure.



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

Fig. 5: *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.**)

3.3 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 5(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 4.

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 5. 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 (3). 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) [41]. The additional complexity of optimizing over disparity parameters τ further adds to the challenge. Methods like [42] explore joint optimization solutions for such problems. In this paper we propose a two-step strategy, where we first: (i) estimate the optimal disparity parameters τ^* using an approximation to (3), 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.

4.1 Obtaining disparity parameters

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

$$\tau^* = \arg\min_{\{\tau\}} \tilde{E}(\tilde{\mathbf{x}}; \mathbf{\Theta}, \tau), \tag{9}$$

where we approximate the original cost function (1) by only using unary and ignoring the pairwise terms[2] 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 (9) 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

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

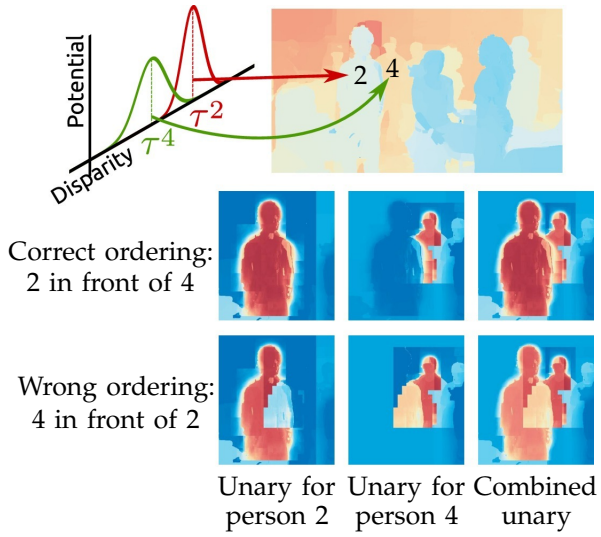


Fig. 6: *The front-to-back ordering of people in a scene is determined by τ^l, the disparity parameter in the potential (7), estimated for each person (shown at the top). The optimal set τ^* is estimated jointly for all the people by solving (9) over a candidate set. Here we show the effect of picking wrong τ^l for two people, which implies wrong ordering (shown at the bottom). This results in poor unary cost functions and a higher overall cost, due to the additional negative evidence in the form of $(1-\beta_i^m)$ as defined in (5). The colours red, yellow and blue in the unary cost figures represent low, medium and high costs respectively. Unaries (here for persons 2 and 4) are combined (third column) by taking their per-pixel minimum, as described in Section 4.1. Note the lower cost (more red) of the combined unary for the correct person ordering.* (**Best viewed in colour.**)

person l. 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. Note that the disparity parameters are estimated jointly for all the people in the scene. We illustrate this on a sample image in Figure 6.

4.2 Person segmentation

With the estimated disparity (and pose) parameters, we compute the unary and smoothness costs, and use the efficient α-expansion algorithm [43] to optimize (1). This assigns every pixel a person or background label from the set $\mathcal{L}$.

5 Inria 3DMovie Dataset

Our new annotated Inria 3DMovie dataset used for evaluation in this paper is available on the project website [44]. Most of this data was extracted from the "StreetDance 3D" [Giwa and Pasquini, 2010] and "Pina" [Wenders, 2011] stereo movies. We chose these movies since they are filmed in true stereoscopic 3D, unlike others where 3D effects are added in post-production and result in inferior disparity estimation. Some of the negative stereo pairs were harvested from Flickr and were originally shot with a Fuji W3 camera. The dataset includes stereo pairs (as jpegs), estimated disparity, (manually annotated) ground truth segmentations, poses and person bounding boxes.

The movie "StreetDance" was split into two parts (roughly in the middle), from which we selected the training and test frames, containing multiple people, respectively. The training set is composed of 520 annotated person bounding boxes, 438 annotated poses and 198 annotated segmentation masks from over 230 stereo pairs. Negative training data is extracted from 247 images with no people, taken from the training part of the movie, and from stereo pairs shot with a Fuji W3 camera.

The test set contains 36 stereo sequences (2727 frame pairs). For quantitative evaluation we provide 638 person bounding boxes and 149 pose annotations in 193 frames, among which a few do not contain any people. Given the cost of manually annotating pixel-wise segmentation, we provide this on a smaller set of 180 frames, containing 686 annotated person segmentations.

6 Experiments

We first detail our method for extracting disparity maps from stereo videos (Section 6.1) and report results for person detection (Section 6.2), pose estimation (Section 6.3), and segmentation (Section 6.4). Next, in Section 6.5, we investigate the sensitivity of our algorithm to its main parameters and in Section 6.6, we analyze the robustness of our approach by replacing its components with ground truth results. Finally, in Section 6.7 we evaluate the segmentation accuracy of our method on the H2view dataset [11].

6.1 Disparity estimation

We chose to work directly with disparity instead of depth (similarly to [16]), since this avoids: (i) explicitly estimating the parameters of the stereo rig (focal length, baseline), (ii) problems when dealing with infinite depth and amplifying errors at small disparities. We estimate the disparity for each frame independently. A joint estimation of motion and disparity from video is also possible [45]. 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 to the same horizontal scan-line. We use the method of Ayvaci *et al.* [46] 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 [7].

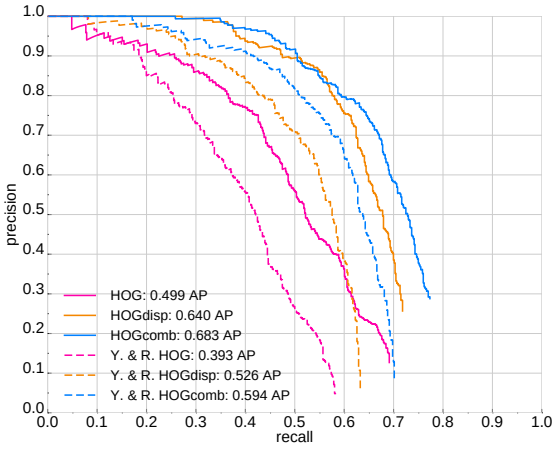


Fig. 7: *Precision-recall curves for person detection based on Yang and Ramanan [19] (*Y. & R.*; dashed lines) and Felzenszwalb* et al. *[36] (solid lines) methods. For both methods we report the performance of the appearance (*HOG*) and disparity (*HOGdisp*) based detectors, as well as the jointly trained appearance and disparity based detector (*HOGcomb*).* HOGcomb*, the detector based on [36] performs significantly better than the other models. (***Best viewed in colour.***)*

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 [33].

We use the horizontal component of the estimated disparity field in our formulation. 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 [46].

6.2 Person detection

We trained our person detection and pose estimation (evaluated in Section 6.3) methods on annotated sequences from the training part of the movie "StreetDance". This trained model is applied on our test set, as well as the 7 test video sequences from the H2view dataset [11] (Section 6.7).

For person detection, we report 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 features, using the concatenation of the two features (HOGcomb). We evaluated them on standard metrics from the PASCAL VOC development kit 2011 [2]. Precision-recall curves are shown in Figure 7, 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.

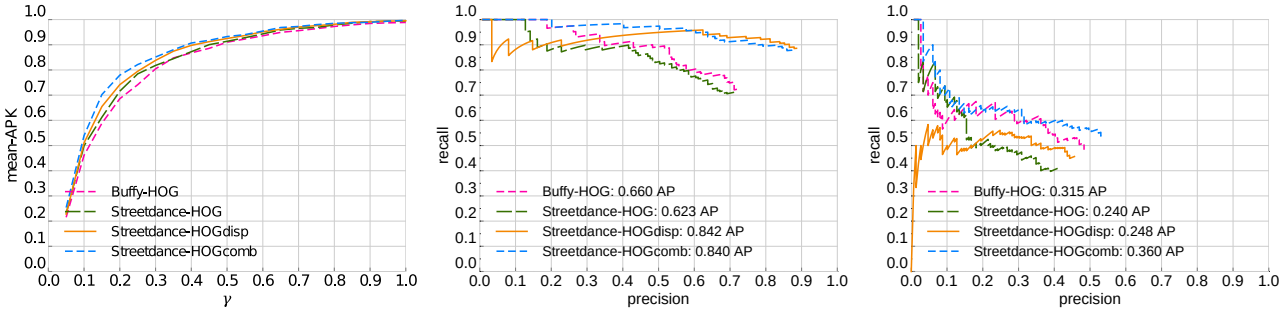
TABLE 1: *Evaluating pose estimation. We report global APK scores as well as scores for all five body parts, as in [47]. We also evaluate the upper-body model from [19] trained on the Buffy dataset. The combination of appearance and disparity features (*HOGcomb*) outperforms the individual estimators (*HOG, HOGdisp*). Note that these scores are the average of the left and the right body parts, while those in Figure 8(b,c) show the scores for the left elbow and wrist only.*

	[19]	HOG	HOGdisp	HOGcomb
Head	0.976	0.983	**0.993**	0.986
Shoulders	0.935	0.931	0.947	**0.969**
Elbows	0.658	0.665	0.759	**0.784**
Wrists	0.298	0.294	0.297	**0.400**
Hips	0.563	0.705	0.714	**0.757**
Global	0.686	0.716	0.742	**0.779**

As discussed in Section 3.1, the three variants above – HOG, HOGdisp, HOGcomb – are based on the deformable part-based person detector [36]. We found this to perform empirically better than the person detection component in [19]; see Figure 7. This is likely due to [19] relying on accurate detection of all individual body parts (e.g., elbows, wrists, which are challenging to detect) to predict the location of the person, whereas [36] uses a more holistic person model. In other words, [36] is more robust to body parts being occluded or poorly detected.

6.3 Pose estimation

Pose estimation is typically evaluated using the percentage of correctly estimated body parts (PCP) score [19], [48]. A body part is deemed to be 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. However, as argued in [47], a relaxed version of this definition has often been used in place of the original one, making it hard to compare published results. Furthermore, PCP requires matching the ground truth poses with the estimated ones, but there is no specification of how this matching should be done. Lastly, this measure uses the ground truth length of each part to set the radius within which the part is deemed to be correctly detected. This results in a foreshortening bias, where shorter limbs (which have a shorter radius) are penalized more severely than longer limbs. Thus, we follow [47] and use their average precision of keypoints (APK) measure instead. In contrast to PCP, which evaluates the correctness of a part (connected to two joints/keypoints), APK measures the correctness of each keypoint. To overcome the foreshortening bias, the APK measure is based on the size of the ground truth person bounding box, rather than the individual parts. More precisely, a keypoint is considered to be correctly estimated if it lies within a radius given by the largest side of the ground truth pose bounding box, scaled by γ. Since the person detections are



(a) Mean-APK (b) Left elbow (c) Left wrist

Fig. 8: *Pose estimation results.* Buffy-HOG *is the upper-body model from [19], and* Streetdance- *corresponds to our models trained on appearance or/and disparity features extracted from the 3D movie Streetdance. (a) Mean-APK curves, which are produced by varying the γ threshold. (b) & (c) Precision-recall curves for left elbow and left wrist respectively. Using disparity cues improves the recall of the pose estimator for elbows, and combining them with appearance cues shows a good initial precision performance. Estimating the wrist position remains a challenge, and the overall performance for this part is similar to [19].* (**Best viewed in colour.**)

evaluated separately (Section 6.2), we use APK to only measure the pose estimation accuracy by considering the pose with the highest automatically obtained confidence score for each person detected.

In Figure 8, we present mean APK curves, where we vary γ between 0 and 1, and plot APK curves for left elbow and left wrist for $\gamma = 0.2$, similar to [47]. The APK scores for all the parts are given in Table 1. The jointly trained pose estimator (HOGcomb) outperforms the individual estimators. We observe that the head and shoulder body parts are localized with high accuracy. Furthermore, combining appearance and disparity cues improves the localization of lower arms (elbows and wrists) by at least 7%.

6.4 Segmenting multiple people

In our experiments we used the following parameter values: $\lambda = 1.0$, $\lambda_1 = 6.3$, $\lambda_2 = 6$, $\lambda_3 = 2.7$, $\sigma_c^2 = 0.025$, $\sigma_v^2 = 0.01$, $\sigma_p^2 = 0.025$, which were set empirically, and fixed for the evaluation. A quantitative evaluation of the segmentation model using ground truth annotations is shown in Table 2. In this evaluation we compare three variants of our approach and two baseline methods. The first one ("No mask, single frame") 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 ("Uni mask, single frame") 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 [5]. 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 ("Pose mask, single frame") incorporates the articulated pose mask, described in Section 3. Our complete model ("Proposed") introduces temporal smoothness across frames.

TABLE 2: *Evaluation of pixel-wise person segmentation on our Inria 3DMovie dataset. We used precision, recall and intersection* vs. *union scores to compare the methods. Our method ("*Proposed*"), which uses disparity, colour, and motion features, along with pose likelihoods and temporal terms shows the best performance. We also show results of variants of our approach and two baseline methods.*

Method	Precision	Recall	Int. vs Union
Proposed	**0.869**	**0.915**	**0.804**
Variants of our method:			
No mask, single frame	0.525	0.371	0.278
Uni mask, single frame	0.783	0.641	0.544
Pose mask, single frame	0.849	0.905	0.779
Baselines:			
Colour only	0.778	0.769	0.630
[48]	0.762	0.853	0.662

For the "Colour only" baseline, we used a colour-based model for the unary costs without the disparity potential. These costs were computed from colour histograms for each label [34]. In other words, each label is associated with a histogram computed from a region in the image, and the unary cost of a pixel is a function of the likelihood of the pixel, given its colour, taking this label. 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., "No mask, single frame", 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 [48], which computes the pose of a person in a scene. We evaluated the (monocular) person *vs.* background segmentation performed as part of this formulation on our dataset.

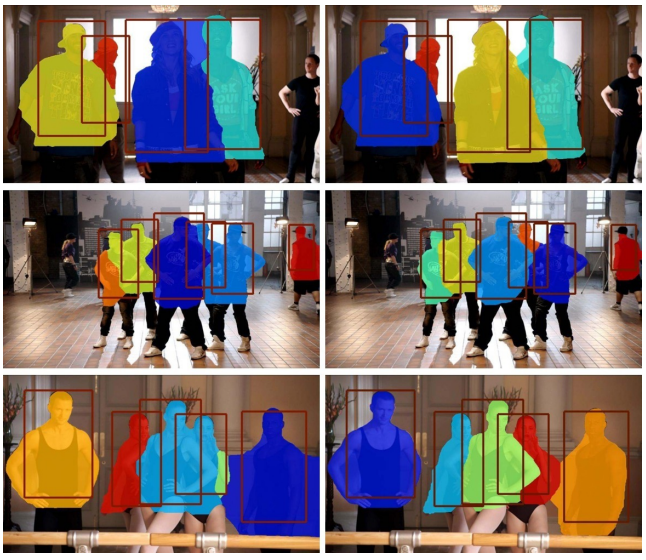
We used the precision, recall, and intersection *vs.* union [2] measures to evaluate our segmentation results. From Table 2, our method "Proposed" shows the best performance. The poor performance of the "Colour only" method, despite a reasonable initial-

(a) Original image (b) Segmentation result

Fig. 9: *Qualitative results on images from the movies "StreetDance" and "Pina". 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 very difficult examples. For instance, in Row 1, the segmentation for the people in the 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.**)

ization for the histograms, is perhaps an indication of the difficulty of our dataset. From Figures 1 and 9 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.

We then evaluated the benefits of the temporal smoothness terms in (1). Performing segmentation



(a) (b)

Fig. 10: *Comparison of segmentation performed: (a) individually on each frame; and (b) temporally on video. We overlay the result of our person detector on each image. We observe that the temporal consistency term reduces leaking (Row 1, rightmost person). It also helps segment more people in the scene accurately (Rows 2 and 3).* (**Best viewed in colour.**)

temporally shows a 2% increase in the intersection *vs*. union score (Table 2). We also observe that it reduces flickering artifacts, produces more consistent segments and reduces leaking in the segmentation, as shown in Figure 10 and the video results [44]. Other methods [49] to propagate segmentations from a few key frames of the video onto others can also be used.

Results on a few sample frames for the "Proposed" method are shown in Figure 9. The influence of the articulated pose mask is analyzed in Figure 11. Another component of our model – the smoothness terms based on colour, motion, and depth – are analyzed in Figure 12.

The success of our approach depends on the quality of detections. Here, we operated in the high-precision mode, at the expense of missing difficult examples, e.g., heavily occluded people. Other prominent failure modes of our method are: (i) challenging poses, which are very different from the training data; and (ii) cases where the disparity signal is noisy for people far away from the camera (e.g., Figure 9, row 1).

6.5 Sensitivity to parameters

In this section we experimentally investigate the sensitivity of the proposed algorithm to its main parameters. The parameter α^l in (6) moderates the relative weight of the pose mask and the disparity cues for person label l. We used one single $\alpha = 0.45$ for all the labels in the results discussed thus far. In Figure 13(a) we show the influence of varying α on the segmentation score. We observe that using

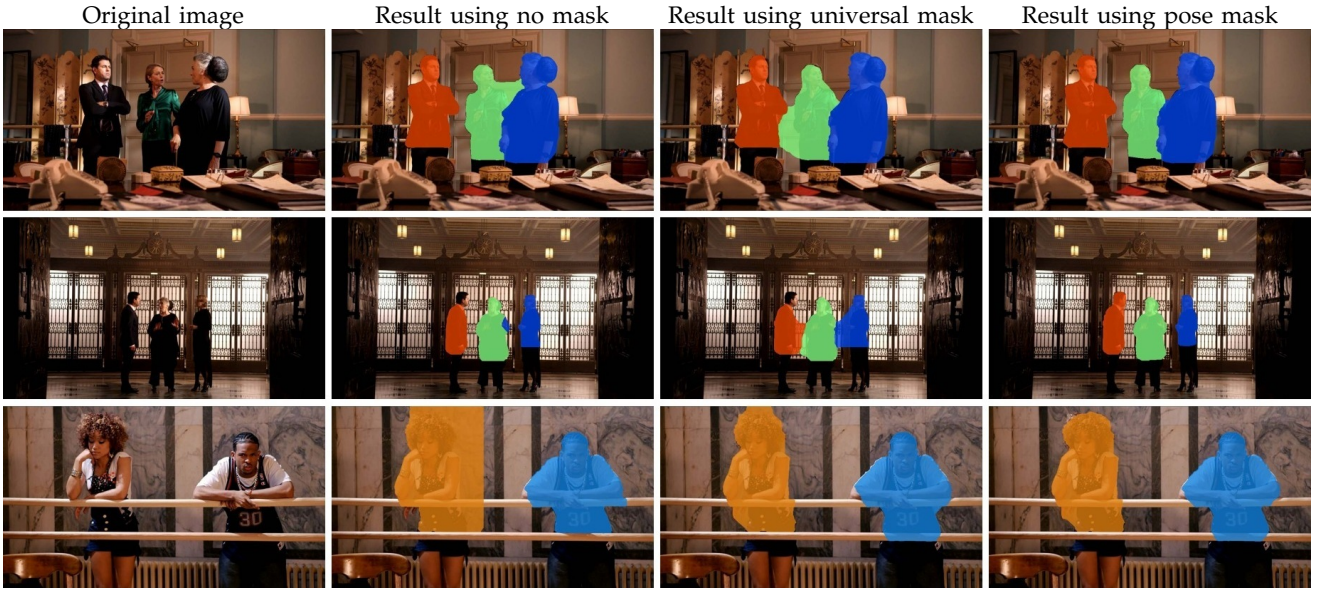
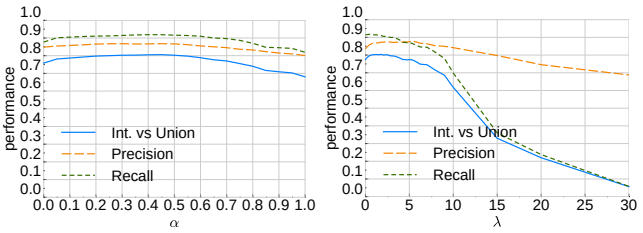


Fig. 11: *Benefits of the articulated pose mask. (a) Left input image. (b) Segmentation result using no mask. In this case, the disparity-based likelihoods are not combined with any pose prior. (c) Segmentation result using a single universal pose mask. The disparity-based likelihood is combined with a potential computed from the universal mask.(d) Segmentation result using articulated pose-specific masks; see Section 3.3. We observe that using a mask improves the segmentation, and the pose-specific masks show the best performance.* (**Best viewed in colour.**)

Fig. 12: *Influence of the motion, colour and disparity sensitive smoothness cost on segmentation results. (a) Left input image. (b) Illustration of the spatial smoothness cost. Red denotes high cost, and the yellow to blue range of colours denotes low cost. (c) Segmentation result using no smoothness cost. (d) Segmentation result using the smoothness cost. Using this pairwise term reduces person segments leaking into the background.* (**Best viewed in colour.**)

no pose cues (i.e., $\alpha = 1.0$) shows a lower average performance than giving equal importance to pose and disparity cues on the entire dataset. However, we note that increasing the influence of the disparity term segments articulated poses more accurately, as shown in Figure 14, at the expense of reduced precision in other situations, such as scenes with multiple people who are close to each other and at similar depth where pose estimates help. We use $\alpha = 0.45$ so that the pose and disparity terms have nearly equal influence and avoid a bias towards one of the extremes.

We also analyzed the influence of the parameters λ_1, λ_2 and λ_3 in the pairwise term (8). The segmentation score was fairly robust to changing these parameters. For instance, disabling any of the three terms still leads to a reasonable performance, and varying the



(a) (b)

Fig. 13: *(a) Influence of the parameter α, specifying the relative weight of the pose mask and disparity cues. All the results in the paper are produced with $\alpha = 0.45$. Using only disparity cues ($\alpha = 1.0$) leads to worse overall performance than using a combination of pose and disparity cues. (b) Influence of the overall weight λ of the pairwise terms. We use $\lambda = 1.0$ in all the experiments.*

relative influence of each term showed only minor variations in the segmentation quality. In contrast, changing the overall influence of the pairwise term, λ in (8), shows first a slight increase in the segmentation score but putting too much weight on the pairwise terms reduces the segmentation score as shown in Figure 13(b).

6.6 Analysis with ground truth components

We further analyze the robustness of our approach by replacing its components with ground truth results. In particular, we use ground truth person detections, pose estimates and disparity parameters. The ground truth disparity parameters are mean and standard deviation computed with the disparity values of all the pixels within each ground truth person segmentation mask. The analysis is performed on individual frames, where ground truth annotations are available, i.e., using the method "Pose mask, single frame" (see Section 6.4) without any temporal smoothing. The results are summarized in Table 3 and demonstrate that using the noisy disparity and pose estimates (rows 1-3) results in only a moderate loss in the segmentation accuracy compared to the segmentation with their ground truth values (row 4). Please note that the segmentation results in Tables 2 and 3 are not directly comparable, since all results in Table 3 are based on the full set of ground truth person detections.

6.7 H2view dataset

The H2view dataset [11] 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 [11]. As our method models only the upper body, we cropped the ground truth, our

(a) $\alpha = 0.2$ (b) $\alpha = 0.45$ (c) $\alpha = 0.8$

Fig. 14: *Qualitative influence of the mixing parameter α, specifying the relative weight of the pose mask and disparity cues. Note that putting more weight on the disparity cues (increasing α) results in a better segmentation of people with articulated poses (Row 1), but performs worse when multiple people at a similar depth are close to each other (Row 3).* **(Best viewed in colour.)**

TABLE 3: *Evaluation of pixel-wise person segmentation on our Inria 3DMovie dataset using ground truth components. We show results using ground truth detection (Det.), ground truth pose masks (Pose) and ground truth disparity parameters τ (Disp.). Using the noisy estimated pose and disparity parameters (rows 1-3) results in only a moderate loss in the segmentation accuracy compared to the segmentation with their ground truth values (row 4).*

Method	Precision	Recall	Int. vs Union
Variants with ground truth:			
Det.	0.862	0.864	0.759
Det. + Disp.	0.872	0.884	0.782
Det. + Pose	0.869	0.908	0.799
Det. + Pose + Disp.	**0.892**	**0.929**	**0.835**

results, and those from [11] 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 (see Table 4). Qualitative results on frames from different sequences in the H2view dataset are shown in Figure 15. Our segmentation produces cleaner, and more human-like shapes, compared to the seed-based segmentation from [11].

An extension of our method for full body segmentation can be envisaged by expanding the bounding boxes (in which we perform the segmentation) vertically. Since our articulated pose mask does not capture the lower limbs, we only used depth cues in this setting. Although this led to some leaking in the segmentation result (due to the noisy disparity signal close to the ground), our method achieves an overall segmentation performance similar to [11] (see Table 4).

Computation time: On a 960×540 frame it 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 imple-

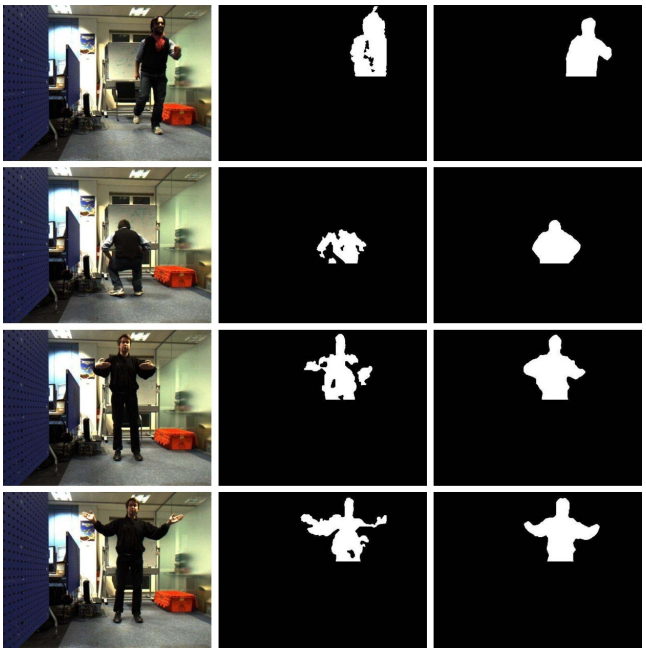
TABLE 4: *Evaluation of pixel-wise person segmentation on the H2view dataset. Our method for segmenting upper bodies shows about 9% improvement in int.* vs. *union score over [11]. Note that our method for full body segmentation only uses upper body pose mask.*

Method	Precision	Recall	Int. vs Union
Upper body segmentation:			
[11]	0.848	0.841	0.735
Proposed	**0.940**	**0.871**	**0.825**
Full body segmentation:			
[11]	0.796	**0.832**	0.692
Proposed	**0.880**	0.789	**0.706**

mentation. The time for segmentation is 6s per frame for the H2view dataset, which contains 512×384 frame sequences of a single person.

7 Discussion

We have developed a model for segmentation of people in stereoscopic movies. The model explicitly represents occlusions, incorporates person detections, pose estimates, and recovers 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, provides a strong shape prior for segmentation. The developed representation presents a building block for modelling and recognition of human actions and interactions in 3D movies.



(a) (b) (c)

Fig. 15: *Qualitative results on images from the H2view dataset. (a) The original image, (b) result from [11] (upper body only), and (c) our result, are shown in each row. Note that our approach shows better performance, including cases with challenging poses (Row 2). Some of the finer details in the segmentation could be improved further, e.g. hands.* (**Best viewed in colour.**)

References

[1] B. Yao and L. Fei-Fei, "Modeling mutual context of object and human pose in human-object interaction activities," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2010, pp. 17–24.

[2] "http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2011."

[3] 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, Proc. Int'l Conf. Computer Vision*, 2011, pp. 1127–1133.

[4] J. C. Niebles, B. Han, and L. Fei-Fei, "Efficient extraction of human motion volumes by tracking," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2010, pp. 655–662.

[5] Y. Yang, S. Hallman, D. Ramanan, and C. Fowlkes, "Layered object models for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1731–1743, 2011.

[6] D. B. Goldman, C. Gonterman, B. Curless, D. Salesin, and S. M. Seitz, "Video object annotation, navigation, and composition," in *Proc. User Interface Software and Technology*. ACM, 2008, pp. 3–12.

[7] S. Koppal, C. Zitnick, M. Cohen, S. Kang, B. Ressler, and A. Colburn, "A viewer-centric editor for 3D movies," *Computer Graphics and Applications*, vol. 31, no. 1, pp. 20–35, 2011.

[8] X. Ren, L. Bo, and D. Fox, "RGB-(D) scene labeling: Features and algorithms," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2012, pp. 2759–2766.

[9] 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 *Proc. Conf. Computer Vision and Pattern Recognition*, 2011, pp. 1297–1304.

[10] "http://vision.middlebury.edu/stereo/," 2013.

[11] G. Sheasby, J. Valentin, N. Crook, and P. H. S. Torr, "A robust stereo prior for human segmentation," in *Proc. Asian Conf. Computer Vision*, 2012, pp. 94–107.

[12] V. Kolmogorov, A. Criminisi, A. Blake, G. Cross, and C. Rother, "Bi-layer segmentation of binocular stereo video," in *Proc. Conf. Computer Vision and Pattern Recognition*, vol. 2, 2005, pp. 407–414.

[13] C. Keller, M. Enzweiler, M. Rohrbach, D. Llorca, C. Schnorr, and D. Gavrila, "The benefits of dense stereo for pedestrian detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 4, pp. 1096–1106, 2011.

[14] K. Schindler, A. Ess, B. Leibe, and L. Van Gool, "Automatic detection and tracking of pedestrians from a moving stereo rig," *ISPRS J. Photogrammetry and Remote Sensing*, vol. 65, no. 6, pp. 523–537, 2010.

[15] L. Spinello and K. O. Arras, "People detection in RGB-D data," in *Intelligent Robots and Systems*, 2011, pp. 3838–3843.

[16] S. Walk, K. Schindler, and B. Schiele, "Disparity statistics for pedestrian detection: Combining appearance, motion and stereo," in *Proc. European Conf. Computer Vision*. Springer, 2010, pp. 182–195.

[17] S. Johnson and M. Everingham, "Learning effective human pose estimation from inaccurate annotation," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2011, pp. 1465–1472.

[18] B. Sapp, D. Weiss, and B. Taskar, "Parsing human motion with stretchable models," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2011, pp. 1281–1288.

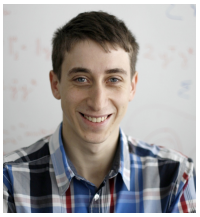
[19] Y. Yang and D. Ramanan, "Articulated pose estimation using flexible mixtures of parts," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2011, pp. 1385–1392.

[20] L. Pishchulin, A. Jain, M. Andriluka, T. Thormahlen, and B. Schiele, "Articulated people detection and pose estimation: Reshaping the future," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2012.

[21] K. Fragkiadaki, H. Hu, and J. Shi, "Pose from flow and flow from pose," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2013.

[22] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Strong appearance and expressive spatial models for human pose estimation," in *Proc. Int'l Conf. Computer Vision*, 2013.

[23] P. Kohli, J. Rihan, M. Bray, and P. Torr, "Simultaneous segmentation and pose estimation of humans using dynamic graph cuts," *Int'l J. Computer Vision*, vol. 79, no. 3, pp. 285–298, 2008.
[24] H. Wang and D. Koller, "Multi-level inference by relaxed dual decomposition for human pose segmentation," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2011, pp. 2433–2440.
[25] L. Ladicky, P. H. S. Torr, and A. Zisserman, "Human pose estimation using a joint pixel-wise and part-wise formulation," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2013.
[26] M. Eichner and V. Ferrari, "We are family: Joint pose estimation of multiple persons," in *Proc. European Conf. Computer Vision*, 2010.
[27] M. P. Kumar, P. H. S. Torr, and A. Zisserman, "Learning layered motion segmentations of video," in *Proc. Int'l Conf. Computer Vision*, vol. 1, 2005, pp. 33–40.
[28] D. Sun, E. Sudderth, and M. Black, "Layered image motion with explicit occlusions, temporal consistency, and depth ordering," in *Proc. Neural Information Processing Systems*, 2010, pp. 2226–2234.
[29] P. H. S. Torr, R. Szeliski, and P. Anandan, "An integrated bayesian approach to layer extraction from image sequences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 297–303, 2001.
[30] J. Y. A. Wang and E. H. Adelson, "Representing moving images with layers," *IEEE Trans. Image Process.*, vol. 3, no. 5, pp. 625–638, 1994.
[31] J. Xiao and M. Shah, "Motion layer extraction in the presence of occlusion using graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1644–1659, 2005.
[32] K. Alahari, G. Seguin, J. Sivic, and I. Laptev, "Pose estimation and segmentation of people in 3D movies," in *Proc. Int'l Conf. Computer Vision*, 2013.
[33] C. Liu, "Beyond pixels: Exploring new representations and applications for motion analysis," Ph.D. dissertation, Massachusetts Institute of Technology, 2009.
[34] Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images," in *Proc. Int'l Conf. Computer Vision*, vol. 1, 2001, pp. 105–112.
[35] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, 2011.
[36] P. F. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, 2010.
[37] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. Conf. Computer Vision and Pattern Recognition*, vol. 1, 2005, pp. 886–893.
[38] J. Shi and C. Tomasi, "Good features to track," in *Proc. Conf. Computer Vision and Pattern Recognition*, 1994, pp. 593 – 600.
[39] M. Everingham, J. Sivic, and A. Zisserman, ""Hello! My name is... Buffy" – Automatic naming of characters in TV video," in *Proc. British Machine Vision Conf.*, vol. I, 2006, pp. 899–908.
[40] C. Desai and D. Ramanan, "Detecting actions, poses, and objects with relational phraselets," in *Proc. European Conf. Computer Vision*, 2012, pp. 158–172.
[41] E. Boros and P. L. Hammer, "Pseudo-boolean optimization," *Discrete Applied Mathematics*, 2002.
[42] H. Isack and Y. Boykov, "Energy-based geometric multi-model fitting," *Int'l J. Computer Vision*, vol. 97, no. 2, pp. 123–147, 2012.
[43] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, 2001.
[44] "http://www.di.ens.fr/willow/research/stereoseg," 2014.
[45] A. Wedel, C. Rabe, T. Vaudrey, T. Brox, U. Franke, and D. Cremers, "Efficient dense scene flow from sparse or dense stereo data," in *Proc. European Conf. Computer Vision*, 2008, pp. 739–751.
[46] A. Ayvaci, M. Raptis, and S. Soatto, "Sparse occlusion detection with optical flow," *Int'l J. Computer Vision*, vol. 97, no. 3, pp. 322–338, May 2012.
[47] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2013.
[48] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari, "2D articulated human pose estimation and retrieval in (almost) unconstrained still images," *Int'l J. Computer Vision*, vol. 99, no. 2, pp. 190–214, 2012.
[49] I. Budvytis, V. Badrinarayanan, and R. Cipolla, "Semi-supervised video segmentation," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2011, pp. 2257–2264.



Guillaume Seguin received the M.S. degree in Computer Science in École Normale Supérieure (ENS), in Paris in 2011. He is currently a PhD student in the research team WILLOW at ENS under the supervision of Josef Sivic and Ivan Laptev. His research interests include computer vision, machine learning and robotics.

Karteek Alahari is a researcher in the LEAR team at Inria Grenoble - Rhône-Alpes. He received the BTech (with Honours) and MS degrees in computer science from IIIT Hyderabad in 2004 and 2005 respectively, and the PhD degree from Oxford Brookes University, where he worked in the Brookes Vision Group, in 2010. Before joining the LEAR team in 2013, he was a postdoctoral fellow in the WILLOW team at Inria Paris - Rocquencourt and École Normale Supérieure. He has been an associate member of the Visual Geometry Group at the University of Oxford since 2006.

Josef Sivic received a degree from the Czech Technical University, Prague, in 2002 and PhD from the University of Oxford in 2006. His thesis dealing with efficient visual search of images and videos was awarded the British Machine Vision Association 2007 Sullivan Thesis Prize and was short listed for the British Computer Society 2007 Distinguished Dissertation Award. His research interests include visual search and object recognition applied to large image and video collections. After spending six months as a postdoctoral researcher in the Computer Science and Artificial Intelligence Laboratory at the Massachusetts Institute of Technology, he currently holds a permanent position as an INRIA researcher at the Departement d'Informatique, Ecole Normale Superieure, Paris. He has published over 40 scientific publications and serves as an Associate Editor for the International Journal of Computer Vision. He has been awarded an ERC Starting grant in 2013.

Ivan Laptev is an INRIA research director at the Departement d'Informatique, École Normale Superieure, Paris, France. He has received his Habilitation degree from École Normale Suprieure (ENS) in 2013 and his PhD degree in Computer Science from the Royal Institute of Technology (KTH) in 2004. He was a research assistant at the Technical University of Munich (TUM) during 1998-1999. He has joined INRIA as a postdoc in 2004 and became a full-time INRIA researcher in 2005. Ivan's main research interests include visual recognition of human actions, objects and interactions. He has published over 50 papers at international conferences and journals of computer vision and machine learning. He serves as an associate editor of IJCV, TPAMI and IVC journals. Ivan was awarded ERC Starting Grant in 2012.