2D Human Pose Estimation in TV Shows

Vittorio Ferrari[1], Manuel Marín-Jiménez[2], and Andrew Zisserman[3]

[1] ETH Zurich
ferrari@vision.ee.ethz.ch
[2] University of Granada
mjmarin@decsai.ugr.es
[3] University of Oxford
az@robots.ox.ac.uk

Abstract. The goal of this work is fully automatic 2D human pose estimation in unconstrained TV shows and feature films. Direct pose estimation on this uncontrolled material is often too difficult, especially when knowing nothing about the location, scale, pose, and appearance of the person, or even whether there is a person in the frame or not.

We propose an approach that progressively reduces the search space for body parts, to greatly facilitate the task for the pose estimator. Moreover, when video is available, we propose methods for exploiting the temporal continuity of both appearance and pose for improving the estimation based on individual frames.

The method is fully automatic and self-initializing, and explains the spatio-temporal volume covered by a person moving in a shot by soft-labeling every pixel as belonging to a particular body part or to the background. We demonstrate upper-body pose estimation by running our system on four episodes of the TV series *Buffy the vampire slayer* (i.e. three hours of video). Our approach is evaluated quantitatively on several hundred video frames, based on ground-truth annotation of 2D poses[1]. Finally, we present an application to full-body action recognition on the Weizmann dataset.

1 Introduction

Our aim is to detect and estimate 2D human pose in video, i.e. recover a distribution over the spatial configuration of body parts in every frame of a shot. Various pose representations can then be derived, such as a soft-labelling of every pixel as belonging to a particular body part or the background (figure 1b); or the 'stickman' of figure 1c, indicating the location, orientation, and size of body parts. Note, our objective here is not to estimate 3D human pose as in [6,28,31].

We wish to obtain pose estimates in highly challenging uncontrolled imaging conditions, typical of movies and TV shows (figures 10, 11). Achieving this is one of the main contributions of the paper. In this setting, images are often very cluttered, and a person might cover only a small proportion of the image area, as they can appear at any scale. Illumination varies over a diverse palette of lighting conditions, and is often quite dark, resulting in poor image contrast. A person's appearance is unconstrained, as

[1] available at www.robots.ox.ac.uk/~vgg/data/stickmen/index.html

D. Cremers et al. (Eds.): Visual Motion Analysis, LNCS 5604, pp. 128–147, 2009.

a b c

Fig. 1. Objective of this work. (a) Input image. (b) Soft-labelling of pixels to body parts or background. Red indicates torso, green upper arms, blue lower arms and head. Brighter pixels are more likely to belong to a part. Color planes are added up, so that purple indicates overlap between lower-arm and torso, yellow between upper-arm and torso, and so on. (c) Stickman representation of pose, obtained by fitting straight line segments to the segmentations in (b). For enhanced visibility, the lower arms are in yellow and the head is in purple.

they can wear any kind of clothing, including body-tight or loose, short or long sleeves, and any colors/textures. The background is unknown and changes over time, preventing the use of background subtraction techniques [5,9]. Finally, the camera is usually moving, causing motion blur, and multiple people can be present at the same time and can occlude each other during a shot.

Our method covers all poses within the upper-body frontal range. Special attention is given to the arms, as they carry most of the information necessary to distinguish pose. The proposed method supports arms folded over the torso, stretching outwards, pointing forward, etc.

The need for such human centered tracking is evident, with applications ranging from video understanding and search through to surveillance. Indeed 2D human segmentation is often the first step in determining 3D human pose from individual frames [1]. We illustrate the use of the extracted poses with an application to action recognition on the Weizmann dataset.

An earlier version of this work first appeared at [10].

1.1 Approach Overview

We overview the method here for the upper-body case, where there are 6 parts: head, torso, and upper/lower right/left arms (figure 1). Full details are given in section 2. The method is also applicable to full bodies, as demonstrated in section 4.

A recent and successful approach to 2D human tracking in video has been to detect in every frame, so that tracking reduces to associating the detections [24,30]. We adopt this approach where detection in each frame proceeds in three stages, followed by a final stage of transfer and integration of models across frames.

In our case, the task of pose detection is to estimate the parameters of a 2D articulated body model. These parameters are the (x, y) location of each body part, its orientation θ, and its scale. Assuming a single scale factor for the whole person, shared by all body parts, the search space has $6 \times 3 + 1 = 19$ dimensions. Even after taking into account kinematic constraints (e.g. the head must be connected to the torso), there are still a huge number of possible configurations.

Since at the beginning we know nothing about the person's pose, clothing appearance, location and scale in the image, directly searching the whole space is a time consuming and very fragile operation (there are too many image patches that could be an arm or a torso!). Therefore, in our approach the first two stages use a weak model of a person obtained through an upper-body detector generic over pose and appearance. This weak model only determines the approximate location and scale of the person, and roughly where the torso and head should lie. However, it knows nothing about the arms, and therefore very little about pose. The purpose of the weak model is to *progressively reduce the search space* for body parts.

The next stages then switch to a stronger model, i.e. a pictorial structure [9,23,24] describing the spatial configuration of all body parts and their appearance. In the reduced search space, this stronger model has much better chances of inferring detailed body part positions.

1. Human detection and tracking. We start by detecting human upper-bodies in every frame, using a sliding window detection based on Histograms of Oriented Gradients [7], and associate detections over time. Each resulting track connects the detections of a different person. It carves out of the total spatio-temporal volume the smaller subvolume covered by a person moving through the shot. This reduces the search space by setting bounds on the possible (x, y) locations of the body parts and by fixing their scale, thus removing a dimension of the search space entirely.

2. Foreground highlighting. At this stage the search for body parts is only limited by the maximum extent possible for a human of that scale centered on the detected position. We restrict the search area further by exploiting prior knowledge about the structure of the detection window. Relative to it, some areas are very likely to contain part of the person, whereas other areas are very unlikely. This allows the initialization of a GrabCut segmentation [25], which removes part of the background clutter. This stage further constrains the search space by limiting the (x, y) locations to lie within the foreground area determined by GrabCut.

3. Single-frame parsing. We obtain a first pose estimate based on the *image parsing* technique of Ramanan [23]. The area to be parsed is restricted to the region output of foreground highlighting. Since the person's scale has been fixed by stage 1), no explicit search for body parts over scales is necessary.

In order to reduce the double-counting problems typical of pictorial structures [29], we extend the purely kinematic model of [23] to include *"repulsive" edges* favoring configurations of body parts where the left and right arms are not superimposed.

Both foreground highlighting and parsing stages are run separately for each detection in a track.

4. Spatio-temporal parsing. The appearance of the body parts of a person changes little within a shot. Moreover, the position of body parts changes smoothly over time. We exploit both kinds of temporal continuity in a second pose estimation procedure which (i) uses appearance models integrated from multiple frames where the system is confident about the estimated pose; and (ii) infers over a joint spatio-temporal model of pose, capturing both kinematic/repulsive constraints within a frame, and temporal continuity constraints between frames. As appearance is a powerful cue about the location

of parts, the better appearance models improve results for frames where parsing failed or is inaccurate. At the same time, the spatio-temporal model tightens the posterior distributions over part positions and disambiguates multiple modes hard to resolve based on individual frames.

The spatio-temporal parsing stage runs over an entire track, as a track connects all detections of a person. Multiple persons in the same shot result in separate tracks, for each of which we run spatio-temporal parsing separately.

1.2 Related Works

Our work builds mainly on the *Learning to Parse* approach by Ramanan [23], which provides the pictorial structure inference engine we use in stage 3, and on the *Strike-a-pose* work [24]. The crucial difference to both works is the way the search space of body part configurations is treated. Thanks to the proposed detection and foreground highlighting stages, we avoid the very expensive and fragile search necessary in [23,24]. Moreover, compared to [24], our initial detection stage is *generic over pose*, so we are not limited to cases where the video contains a pre-defined characteristic pose at a specific scale. We also generalize and improve the idea of transferring appearance models of [24]. Rather than using a single frame containing the characteristic pose, we integrate models over multiple frames containing any pose.

Previous use of pictorial structure models have tolerated only limited amounts of background clutter [9,23] and often assume knowledge of the person's scale [23,24] or background subtraction [9]. A few methods operate interactively from regions of interest provided by the user [19].

There are also methods that detect humans using generative models for the entire video sequences, e.g. [14,16]. However, to date these methods have been limited to relatively simple backgrounds and to no occlusion of the person.

Our spatio-temporal model (section 2.4) is most closely related to that of [28,29], but our framework is fully automatic (it does not need any manual initialization or background subtraction).

The work of [20] recovers unusual, challenging body configurations in sports images by combining segmentation and detectors trained to specific body parts, but requires a person centered in the image and occupying most of it.

Finally, very recently two methods [3,12] have been presented for pose estimation of people walking in busy city environments where the camera, multiple people, as well as other objects move simultaneously. Both methods rely heavily on static and dynamic priors specific to walking motion. In contrast, our method makes no assumptions about expected poses, besides the person being upright, and is able to estimate a wide variety of body configurations (figures 10, 11, 12).

2 The Approach in Detail

2.1 Upper-Body Detection and Temporal Association

Upper-body detection. In most shots of movies and TV shows, only the upper-body is visible. To cope with this situation, we have trained an upper-body detector using the

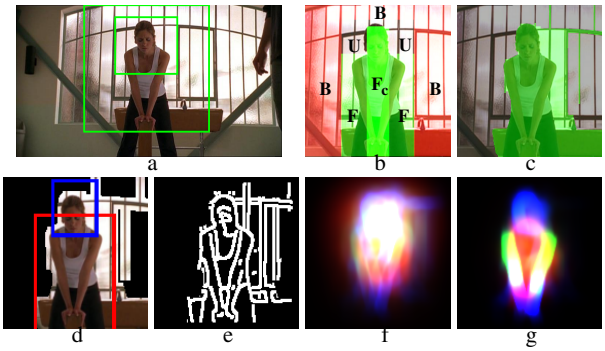


Fig. 2. Overview of the single-frame steps. 1. Upper body detection: The detected person (inner rectangle) and enlarged window where further processing is applied (outer rectangle). **2. Foreground highlighting:** (b) subregions for initializing GrabCut. (c) foreground region output by GrabCut. **3. Parsing:** (d) area to be parsed $\mathcal{F}$ (dilated from (c)) and (e) edges within $\mathcal{F}$. (f) posterior of the part positions $p(l_i|I)$ after the edge-based inference. (g) posterior after the second inference, based on edges and appearance. This visualization is obtained by convolving rectangles representing body parts with their corresponding posterior.

approach of Dalal and Triggs [7], which achieves excellent performance on the related task of full-body pedestrian detection. Image windows are spatially subdivided into tiles and each is described by a Histogram of Oriented Gradients. A sliding-window mechanism then localizes the objects. At each location and scale the window is classified by a linear SVM as containing the object or not. Photometric normalization within multiple overlapping blocks of tiles makes the method particularly robust to lighting variations.

The training data consists of 96 video frames from three movies (*Run Lola run, Pretty woman, Groundhog day*), manually annotated with a bounding-box enclosing an upper-body. The images have been selected to maximize diversity, and include many different actors, with only a few images of each, wearing different clothes and/or in different poses. No images from the test material (shots from *Buffy the Vampire Slayer*) were used for training.

Following Laptev [17], the training set is augmented by perturbing the original examples with small rotations and shears, and by mirroring them horizontally. This improves the generalization ability of the classifier. By presenting it during training with misalignments and variations, it has a better chance of noticing true characteristics of the pattern, as opposed to details specific to individual images. The augmented training set is 12 times larger and contains more than 1000 examples.

We choose an operating point of 90% detection-rate at 0.5 false-positives per image. This per-frame detection-rate translates into an almost perfect per-track detection-rate after temporal association (see below). Although individual detections might be missed, entire tracks are much more robust. Moreover, we remove most false-positives by weeding out tracks shorter than 20 frames.

In practice, this detector works well for viewpoints up to 30 degrees away from straight frontal, and also detects back views (figures 10, 11).

Temporal association. After applying the upper-body detector to every frame in the shot independently, we associate the resulting bounding-boxes over time by maximizing their temporal continuity. This produces *tracks*, each connecting detections of the same person.

Temporal association is cast as a grouping problem [30], where the elements to be grouped are bounding-boxes. As similarity measure we use the area of the intersection divided by the area of the union (IoU), which subsumes both location and scale information, damped over time. We group detections based on these similarities using the Clique Partitioning algorithm of [11], under the constraint that no two detections from the same frame can be grouped. Essentially, this forms groups maximizing the IoU between nearby time frames.

This algorithm is very rapid, taking less than a second per shot, and is robust to missed detections, because a high IoU attracts bounding-boxes even across a gap of several frames. Moreover, the procedure allows persons to overlap partially or to pass in front of each other, because IoU injects a preference for *continuity scale* in the grouping process, in addition to location, which acts as a disambiguation factor.

In general, the 'detect & associate' paradigm is substantially more robust than regular tracking, as recently demonstrated by several authors [22,30].

2.2 Foreground Highlighting

The location and scale information delivered by an upper-body detection greatly constrains the space of possible body parts. They are now confined to the image area surrounding the detection, and their approximate size is known, as proportional to the detection's scale. However, to accommodate for all possible arm poses we must still explore a sizeable area (figure 2a). Stretching out the arms in any direction forms a large circle centered between the shoulders. In challenging images from TV shows, this area can be highly cluttered, confusing the body part estimator.

Fortunately, we have *prior knowledge* about the structure of the search area. The head lies somewhere in the middle upper-half of the detection window, and the torso is directly underneath it (figure 2b). This is known because the detector has been explicitly trained to respond to such structures. In contrast the arms could be anywhere. We propose to exploit this knowledge to initialize GrabCut [25], by learning initial foreground/background color models from regions where the person is likely to be present/absent. The resulting segmentation removes much of the background clutter, substantially simplifying the later search for body parts (figure 2c).

Let $\mathcal{R}$ be a region of interest obtained by enlarging the detection window as in figure 2a. $\mathcal{R}$ is divided into four subregions F, F_c, B, U (see figure 2b). GrabCut is initialized as follows: the foreground model is learnt from F and F_c (F_c is known to belong to the person, while F contains mostly foreground, but some background as well); and the background model from B (it covers mostly background, but it might also include part of the arms, depending on the pose). Furthermore, the region F_c is clamped as foreground, but grabcut is free to set pixel labels in all other subregions (we have extended the original GrabCut algorithm to enable these operations). The *U* region is neutral and no color model is learnt from it. The setup accurately expresses our prior knowledge and results in a controlled, upper-body-specific segmentation, assisted by as

Fig. 3. Examples of foreground highlighting

much information as we can derive from the previous object detection process. Near the head, B and F_c compete for the U region, with the foreground growing outwards until it meets a background-colored area, resulting in a good head segmentation. Along the sides, the background floods into the initial F to segment the shoulders, while at the same time the arms get labeled as foreground because they are colored more similarly to the initial F than to the initial B (figure 3).

The above procedure is rather conservative, and it often retains parts of the background. The goal is not to achieve a perfect segmentation, but to reduce the amount of background clutter (figure 3). It is more important not to lose body parts, as they cannot be recovered later. To validate this, we have inspected 1584 frames of a *Buffy* episode (i.e. every 10th frame) and only in 71 a body part was lost (4.5%). In contrast to traditional background subtraction, used in many previous works to extract silhouettes [5,9,13], our method does not need to know the background *a priori*, and allows the background to change over time (in video).

2.3 Single-Frame Parsing

Our main goal is to explain the spatio-temporal volume covered by a person moving in a shot. In particular, we want to estimate the 2D pose of the person, as the location, orientation and size of each body part. Ideally, the exact image regions covered by the parts should also be found. For estimating 2D pose in individual video frames, we build on the *image parsing* technique of Ramanan [23]. In the following we first briefly summarize it, and then describe our extensions.

Image parsing [23]. A person is represented as a pictorial structure composed of body parts tied together in a tree-structured conditional random field (figure 5a). Parts, l_i, are oriented patches of fixed size, and their position is parametrized by location and orientation. The posterior of a configuration of parts $L = \{l_i\}$ given an image I can be written as a log-linear model

$$P(L|I) \propto \exp\left(\sum_{(i,j)\in E} \Psi(l_i, l_j) + \sum_i \Phi(l_i)\right) \tag{1}$$

The pairwise potential $\Psi(l_i, l_j)$ corresponds to a spatial prior on the relative position of parts and embeds the kinematic constraints (e.g. the upper arms must be attached to the torso). The unary potential $\Phi(l_i)$ corresponds to the local image evidence for a part in a particular position (likelihood). Since the model structure E is a tree, inference is performed exactly and efficiently by sum-product Belief Propagation [4].

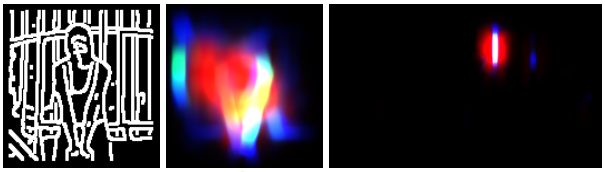


Fig. 4. The *image parsing* pose estimation algorithm of [23] applied to the image in figure 1. (a) All edges inside region $\mathcal{R}$, without filtering them through foreground highlighting. (b) Parsing applied to the whole region $\mathcal{R}$. It achieves a worse estimate than when helped by foreground highlighting, figure 2g, because it is attracted by the bars in the background. (c) Parsing applied directly to the whole image, without reducing the search space to $\mathcal{R}$ based on the initial person detection. It fails entirely.

The key idea of [23] lies in the special treatment of Φ. Since the appearance of neither the parts nor the background is known at the start, only edge features are used. A first inference based on edges delivers soft estimates of body part positions, which are used to build appearance models of the parts (e.g. in figure 2f the torso is in red). Inference in then repeated using an updated Φ incorporating both edges and appearance. The process can be iterated further, but in this paper we stop at this point. The technique is applicable to quite complex images because (i) the appearance of body parts is a powerful cue, *and* (ii) appearance models can be learnt from the image itself through the above two-step process.

The appearance models used in [23] are color histograms over the RGB cube discretized into $16 \times 16 \times 16$ bins. We refer to each bin as a *color* c. Each part l_i has foreground and background likelihoods $p(c|fg)$ and $p(c|bg)$. These are learnt from a part-specific soft-assignment of pixels to foreground/background derived from the posterior of the part position $p(l_i|I)$ returned by parsing. The posterior for a pixel to be foreground given its color $p(fg|c)$ is computed using Bayes' rule and used during the next parse.

As in [23], in our implementation we explicitly maintain a 3D binned volume to represent the possible (x, y, θ) positions of each part (discretization: every pixel for (x, y) and 24 steps for θ). This dense representation avoids the sampling needed by particle representations (e.g. [28,29]). The kinematic potential Ψ has a relative location (x, y) and a relative orientation (θ) components. The former gives zero probability if $(l_j - l_i)$ is out of a box-shaped tolerance region around the expected relative location (i.e. the parts *must* be connected [23]). The relative orientation component is a discrete distribution over $\theta_i - \theta_j$. The parameters of Ψ are learned from training data in [23]. The relative orientation prior is nearly uniform, allowing our approach to estimate a variety of poses.

When [23] is run unaided on a highly cluttered image such as figure 1a, without any idea of where the person might be or how large it is, parsing fails entirely (figure 4c). There are simply too many local image structures which could be a limb, a head, or a torso. This is assessed quantitatively in section 3.

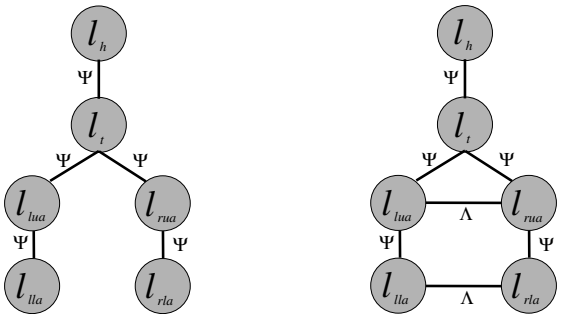


Fig. 5. Single-frame models. Each node represents a body part (h: head, t: torso, left/right upper/lower arms lua, rua, lla, rla). (a) The kinematic tree includes edges between every two body parts which are physically connected in the human body. (b) The repulsive model extends the kinematic tree with edges between opposite-sided arm parts.

We reduce the space explored by parsing based on three sources of information:

- (i) the *location and scale information* supplied by the upper-body detector, is used to define the enlarged search region $\mathcal{R}$. Parsing is run only within $\mathcal{R}$, rescaled to a fixed size, tuned to roughly yield the part sizes expected by the parser. Thanks to the proper use of scale information from detection, we effectively obtain scale-invariant pose estimation, without having to explicitly search for body parts at multiple scales. This significantly reduces ambiguity and false positive detections.
- (ii) *Foreground highlighting*. We further simplify pose estimation by restricting the area to be parsed to the region $\mathcal{F} \subset \mathcal{R}$ output of foreground highlighting (figure 2d). This is realized by removing all edges outside $\mathcal{F}$ and setting all pixels $\mathcal{R} \setminus \mathcal{F}$ to black. This causes the image evidence $\Phi(l_i)$ to go to $-\infty$ for $l_i \notin \mathcal{F}$, and hence it is equivalent to constraining the space of possible poses.
- (iii) *Head and torso constraints*. A final assistance is given by mildly constraining the (x, y) location of the head and torso based on our prior knowledge about the spatial structure of $\mathcal{R}$ (see section 2.2). The constraints come in the form of broad subregions $\mathcal{H}, \mathcal{T} \in \mathcal{R}$ where the head and torso must lie, and are realized by setting $\Phi(l_{head}), \Phi(l_{torso})$ to $-\infty$ for $l_i \notin \mathcal{H}, \mathcal{T}$ (figure 2d). These constraints directly reflect our prior knowledge from the detection process and therefore do not limit the range of poses covered by the parser (e.g. for the arms).

All the above aids to pose estimation are made possible from the initial generic upper-body detection. Foreground highlighting and location constraints can only be automated when building on a detection window. The combined effect of these improvements is a vastly more powerful parser, capable of estimating 2D pose in a highly cluttered image, even when the person occupies only a small portion of it. Moreover, parsing is now faster, as it searches only an image region, supports persons at multiple scales, and multiple persons at the same time, as each detection is parsed separately.

Repulsive model. A well-known problem with pictorial structure models evaluated as trees such as the one above, is that different body parts can take on similar (x, y, θ)

states, and therefore cover the same image pixels. Typically this happens for the left and right lower (or upper) arms, when the image likelihood for one is substantially better than the likelihood for the other. It is a consequence of the assumed independence of the left and right arms in the tree. This is referred to as the *double-counting problem* and was also noticed by other authors [29]. One solution, adopted in previous work, is to explicitly model limb occlusion by introducing layers into the model [2,15,29], though the graphical model is then no longer a tree.

Here, in order to alleviate the double-counting problem we add to the kinematic tree model two *repulsive edges* (figure 5b). The first edge connects the left upper arm (lua) to the right upper arm (rua), while the second edge connects the left lower arm (lla) to the right lower arm (rla). The posterior of a configuration of parts in the extended model becomes

$$P(L|I) \propto \exp\left(\sum_{(i,j)\in E} \Psi(l_i, l_j) + \sum_i \Phi(l_i) + \Lambda(l_{lua}, l_{rua}) + \Lambda(l_{lla}, l_{rla})\right) \tag{2}$$

The repulsive prior $\Lambda(l_i, l_j)$ gives a penalty when parts l_i and l_j overlap, and no penalty when they don't

$$\Lambda(l_i, l_j) = \begin{cases} w_\Lambda & \text{if } |l_i - l_j| \leq t_\Lambda \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Therefore, the extended model prefers configurations of body parts where the left and right arms are not superimposed. It is important to notice that this new model does not forbid configurations with overlapping left/right arms, but is *biased against them*. If the image evidence in their favor is strong enough, inference will return configurations with overlapping arms (figure 10-c2). This properly reflects our prior knowledge that, in the majority of images, the arms don't occlude each other.

Since the extended graphical model has loops, we perform approximate inference with sum-product Loopy Belief Propagation, which in practice delivers a good estimate of the posterior marginals and is computationally efficient. The weight w_Λ and the threshold t_Λ are manually set and kept fixed for all experiments of this paper.

Figure 6 shows a typical case where the purely kinematic model delivers a posterior whose mode puts both lower arms on the left side, while the extended model yields the correct pose, thanks to the repulsive edges.

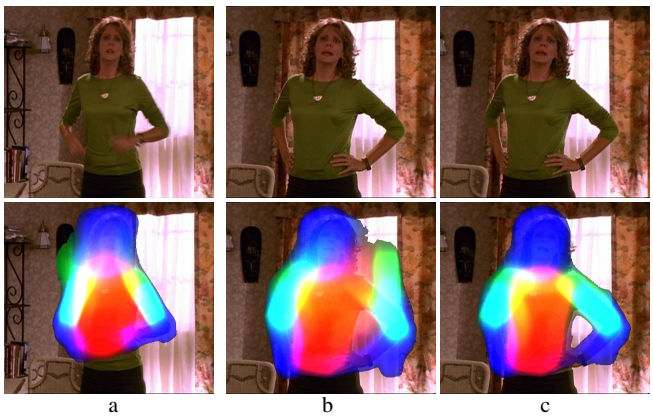
2.4 Spatio-Temporal Parsing

Parsing treats each frame independently, ignoring the temporal dimension of video. However, all detections in a track cover the same person, and people wear the same clothes throughout a shot. As a consequence, the appearance of body parts is quite stable over a track. In addition to this continuity of appearance, video offers also continuity of geometry: the position of body parts changes smoothly between subsequent frames.

In this section, we exploit the continuity of appearance for improving pose estimations in particularly difficult frames, and the continuity of geometry for disambiguating multiple modes in the positions of body parts, which are hard to resolve based on individual frames.

a b c

Fig. 6. Impact of repulsive model. (a) Original image. (b) Pose estimated by the kinematic tree model. As in previous figures, the visualization is obtained by convolving rectangles representing body parts with their corresponding posterior probability over (x, y, θ). The right (in the image) upper arm (green) has two equally probable modes, one at the correct position, and one on the left upper arm. For the right lower arm (blue) instead, nearly all of the probability mass is on the left side, while only very little is on the correct position. This double-counting phenomenon visibly affects the estimation. (c) Pose estimated after extending the model with repulsive edges. The position of the right lower arm is now correctly estimated.



a b c

Fig. 7. Impact of transferring appearance models. (a) one of several frames with low TPE after single-frame parsing, from which integrated appearance models are learnt (top). The pose estimate is quite clear (bottom). (b) a frame with high TPE. The system is uncertain whether the right arm lies on the window or at its actual position (bottom). (c) Parsing the frame in (b) while using the learned integrated appearance models. The right arm ambiguity is now resolved (bottom), as the system acquires from other frames the knowledge that white is a color occurring on the background only.

Learning integrated appearance models. The idea is to find the subset of frames where the system is confident of having found the correct pose, integrate their appearance models, and use them to parse the whole track again (figure 7). This improves pose estimation in frames where parsing has either failed or is inaccurate, because appearance is a strong cue about the location of parts.

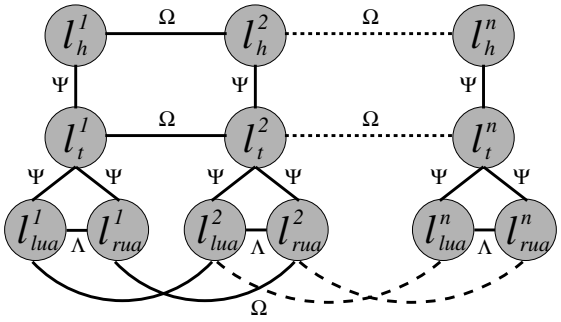


Fig. 8. Spatio-temporal model. For clarity, only head (l_h), torso (l_t), and left/right upper arms (l_{lua}, l_{rua}) are drawn.

Frames where parsing infers a highly confident configuration of body parts provide good reference appearance models (figure 7a). The measure of confidence used here is the entropy of the posterior of the part positions $p(l_i|I)$, accumulated over all parts $L = \{l_i\}$ to give the *total pose entropy* TPE:

$$TPE(L|I) = -\sum_i \left(\sum_{x,y,\theta} p(l_i = \{x, y, \theta\}|I) \cdot \log p(l_i = \{x, y, \theta\}|I) \right) \tag{4}$$

Rather than simply selecting the single frame with the lowest TPE, we learn models by *integrating* over all frames with a similar low TPE. It can be shown [26] that the distribution minimizing the total KL-divergence to a set of distributions is their average. Hence, we integrate the foreground and background likelihoods $\{p_r(c|fg)\}, \{p_r(c|bg)\}$ from the reference frames r by averaging them. The integrated posteriors $p_i(fg|c)$ are then obtained by applying Bayes' rule.

The integrated models are richer, in that $p_i(fg|c)$ is nonzero for a broader range of colors, so they generalize to a larger number of frames. Moreover, they are more accurate, because estimated over a wider support. Examples of the benefits brought by using the learned integrated appearance models to re-parse frames are shown in figure 7 and by the difference between figure 2g (purely single-frame) and figure 1b (re-parsing).

Spatio-temporal inference. We extend the single-frame person model to include dependencies between body parts over time (figure 8). The extended model has a node for every body part in every frame of a continuous temporal window (11 frames in our experiments). The posterior of all configurations of parts $\{L^t\} = \{l_i^t\}$ given all frames $\{I^t\}$ can be written as

$$P(\{L^t\}|\{I^t\}) \propto$$
$$\exp\left(\sum_{t,i} \left(\sum_{j|(i,j)\in E} \Psi(l_i^t, l_j^t) + \Phi(l_i^t) + \Omega(l_i^t, l_i^{t+1}) + \Lambda(l_{lua}^t, l_{rua}^t) + \Lambda(l_{lla}^t, l_{rla}^t) \right) \right) \tag{5}$$

In addition to the *kinematic/repulsive* dependencies Ψ, Λ between different parts in a single frame, there are *temporal* dependencies Ω between nodes representing the same

part in subsequent frames. As a temporal prior $\Omega(l_i^t, l_i^{t+1})$ we use a simple box-shaped distribution limiting the difference in the $l_i^t = (x, y, \theta)$ position of a body part between frames. We use the integrated appearance models to obtain a better image likelihood Φ. Approximate inference in the spatio-temporal model with loops is carried out with sum-product Loopy Belief Propagation.

The spatio-temporal inference is a batch process treating all frames in the temporal window simultaneously, as opposed to traditional tracking, where only past frames can influence estimation in the current frame. The inference procedure outputs the full marginal posteriors $p(l_i^t|\{I^t\})$, defining the probability of every possible (x, y, θ) body part position in every frame. This is better than a single MAP solution [24], as any remaining ambiguity and uncertainty is visible in the full posteriors (e.g. due to blurry images, or tubular background structures colored like the person's arms). Finally, our joint spatio-temporal inference is better than simply smoothing the single-frame posteriors over time, as the kinematic dependencies within a frame and temporal dependencies between frames *simultaneously help each other*.

Thanks to the proposed joint spatio-temporal inference, the final pose estimates are tighter and more accurate than single-frame ones. As a typical effect, multiple modes in the positions of body parts are disambiguated, because modes not consistently recurring over time are attenuated by the temporal prior Ω (figure 9). Moreover, the estimated poses are now more temporally continuous, which is useful for estimating the motion of body parts for action recognition.

3 Upper-Body Pose Estimation Results

We have applied our pose estimation technique to episodes 2,4,5 and 6 of season five of *Buffy the vampire slayer*, for a total of more than 70000 video frames over about 1000 shots.

The examples in figures 10 and 11 show that the proposed method meets the challenges set in the introduction. It successfully recovers the configuration of body parts in spite of extensive clutter, persons of different size, dark lighting and low contrast (c3, f2, f3). Moreover, the persons wear all kinds of clothing, e.g. ranging from long sleeves to sleeveless (b3, a3, h3), and this is achieved despite the fact that their appearance is unknown *a priori* and was reconstructed by the algorithm. The system correctly estimated a wide variety of poses, including arms folded over the body (c1, c2, g1), stretched out (e3, f3), and at rest (e1, f1). The viewpoint doesn't have to be exactly frontal, as the system tolerates up to about 30 degrees of out-of-plane rotation (c1). Persons seen from the back are also covered, as the upper-body detector finds them and we don't rely on skin-color segmentation (e1, f1). Finally, the method deals with multiple persons in the same image and delivers a separate pose estimate for each (h2, note how the pose of each person is estimated independently).

We quantitatively assess these results on 69 shots divided equally among three of the episodes. We have annotated the ground-truth pose for four frames spread roughly evenly throughout each shot, by marking each body part by one line segment [20] (figure 10a). Frames were picked where the person is visible at least to the waist and the arms fit inside the image. This was the sole selection criterion. In terms of imaging

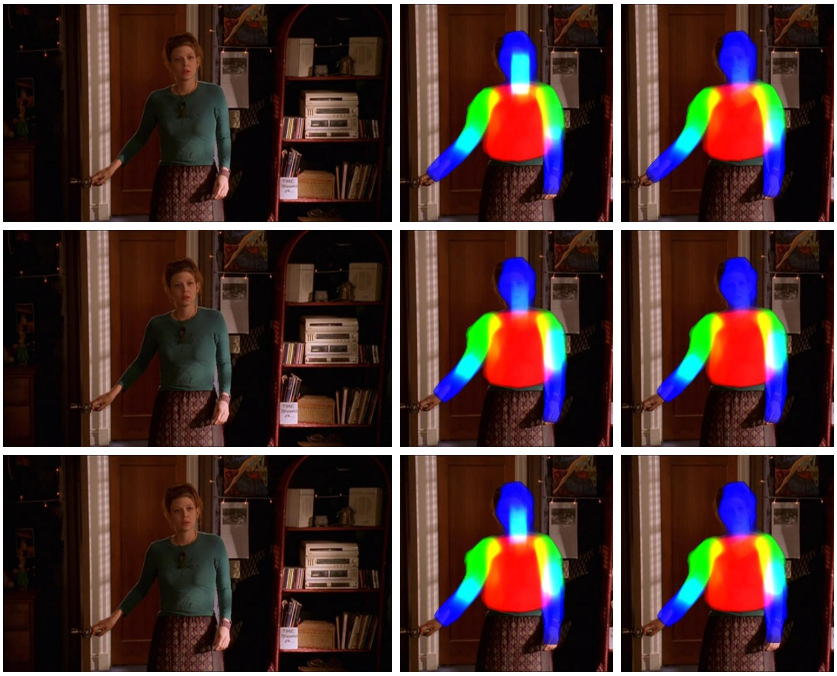


Fig. 9. Impact of spatio-temporal parsing. (a) Three subsequent video frames. (b) Pose estimated by single-frame parsing after transferring appearance models, but without dependencies between body parts over time. In the first and third frames, the right upper arm has a strong second mode on the face, but not in the second frame. (c) Pose estimated by the complete spatio-temporal model. The spurious mode has been largely eliminated.

conditions, shots of all degrees of difficulty have been included. A body part returned by the algorithm is considered correct if its segment endpoints lie within 50% of the length of the ground-truth segment from their annotated location.

The initial detector found an upper-body in 88% of the $69 \times 4 = 276$ annotated frames, and this places an upper bound on performance. Table 1 shows the percentage of the $243 \times 6 = 1458$ body parts in these frames which have been correctly estimated by several versions of our system. Our best result is 62.6%. The image parser of [23] using software supplied by the author, and run unaided directly on the image, achieves only 9.6%, thus highlighting the great challenge posed by this data, and the substantial improvements brought by our techniques. Helping [23] by constraining it by the location and scale delivered by the initial human detection causes performance to jump to 41.2% (section 2.1). Adding foreground highlighting further raises it to 57.9% (section 2.2). These results confirm that both search space reduction stages we proposed (starting from a detection and foreground highlighting) contribute considerably to the quality of the results. Transferring appearance models increases performance moderately to 59.4% (section 2.4). The improvement appears relatively small because in many cases the localization refinements are too fine to be captured by our coarse

Fig. 10. Pose estimation results I. (a1) example ground-truth 'stickman' annotation. All other subfigures are the output of the proposed method, with body part segmentations overlaid. For illustration, in (a2) we also overlay the stickman derived by our method. The color coding is as follows: head = purple, torso = red, upper arms = green, lower arms = yellow. In (c2) a pose with crossed arms is correctly estimated: the repulsive model does not prevent our system from dealing with these cases.

evaluation measure. On the other hand, this suggests that the proposed approach performs well also on static images. Extending the purely kinematic model of [23] with repulsive priors brings a last visible improvement to 62.6%, thanks to alleviating the double-counting problem (section 2.3).

Somewhat surprisingly, including temporal priors does not improve our evaluation score (section 2.4). This is due to two reasons. The first is that many cases where the estimated poses become more temporally continuous do not result in a better score. Our measure does not prize temporal smoothness, it only looks at the position of body parts in individual frames. The second reason is that temporal integration occasionally worsens the estimated poses. Due to the temporal prior, the posterior probability of a (x, y, θ) state in a frame depends also on nearby states in neighboring frames. Therefore, if a body part is 'missing' at its correct position in a frame, i.e. the unary

Fig. 11. Pose estimation results II. More example pose estimations. A fair sample of failures are also included, e.g. (f1) is missed as a detection, and the wrong pose is obtained in (e3) (rear person). Notice how the leftmost person in (h2) is largely occluded.

potential gives it near-zero probability, the posterior probability of the correct position in neighboring frames is decreased by the temporal prior (i.e. it propagates the miss over time; the same behavior also eliminates incorrect modes). A potential solution is to model occluded/missing body parts by extending the state space with a replica for each state, labeled as occluded/missing. This replica would have a low, but non-zero probability. In this fashion, the joint probability of a configuration of body parts including such a state would also be non-zero. This is a topic of our current research.

4 Application: Action Recognition on the Weizmann Dataset

Determining human pose is often a first step to action recognition. For example [18] classifies frames by their pose as a pre-filter to spatio-temporal action detection, and [27] specifies actions by pose queries.

In this section we apply the extracted pose representation to the task of action recognition on the Weizmann dataset [5], which includes nine actions performed by nine different subjects who are fully visible in all frames (including their legs). Following

Table 1. Percentage of correctly estimated body parts by various versions of our method

Method	Performance
Ramanan NIPS 2006 [23]	9.6%
+ detection	41.2%
+ foreground highlighting	57.9%
+ appearance transfer	59.4%
+ repulsive model	**62.6%**
+ complete spatio-temporal model	61.7%

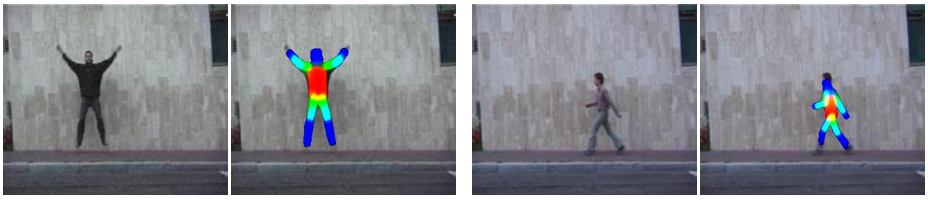


Fig. 12. Pose estimation on the Weizmann dataset [5]. Our methods performs well also on full bodies, and handles a variety of poses (left: a jumping-jack; right: walking).

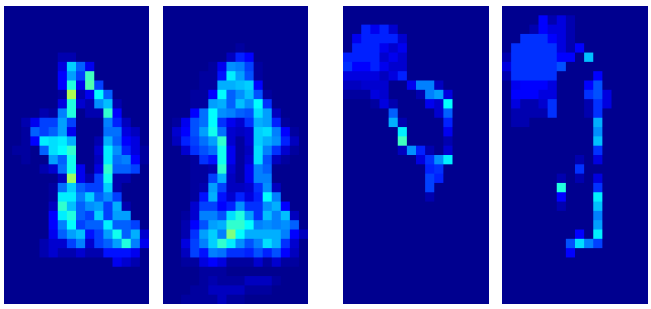


Fig. 13. Action descriptor. Accumulated motion differences for two subjects walking (left) and waving with one hand (right). For illustration, the difference images in this figure are computed from all body parts. Our descriptor instead, is a concatenation of difference images for each body part groups, which provides more discriminative power.

the standard leave-one-out evaluation protocol [5,13,21], we train on eight subjects and test on the remaining one. The experiment is then repeated by changing the test subject and recognition rates are averaged.

Here, we replace the upper-body detector by the standard HOG based pedestrian detector of Dalal and Triggs [7], and employ a full-body pictorial structure including also upper and lower legs (figure 12). Our action descriptor is inspired by motion history images [8] and is obtained as follows. First, for each frame we derive a soft-segmentation from the posteriors of the part positions $p(l_i|I)$ delivered by our pose estimator. Next,

we subtract these soft-segmentations between pairs of subsequent frames, and accumulate the differences over the whole sequence. The resulting accumulated difference image is then subsampled to a 16x32 grid (figure 13). The final descriptor is obtained by computing a separate difference image for each of the four body part groups (torso, arms, legs, and head) and concatenating them. The descriptor is informative because it encodes how much motion each body part group performs, and at which position relative to the coordinate frame of the detection window. It is robust because differences are accumulated over many frames, limiting the impact of incorrect pose estimates in a few frames. For each action, we train a one-vs-all linear SVM on this descriptor, and use them to classify the sequences of the test subjects.

Although previous works using background subtraction achieve perfect results on this dataset [5,13], the only work we are aware of tackling the task without *any* static background assumption[2] only obtains 73% recognition rate [21]. While operating in the same conditions, our method achieves the significantly higher rate of 88%. These results demonstrate the suitability of our technique to full body pose estimation.

5 Appraisal and Future Work

We have demonstrated automated upper body pose estimation on extremely challenging video material – the principal objective of this work.

The numerous works defining action descriptors based on body outlines [5,13] could benefit from our technique, as it provides outlines without resorting to traditional background segmentation, requiring a known and static background.

Of course, further improvements are possible. For example the body part segmentations could be improved by a further application of GrabCut initialized from the current segmentations. Another possible extention is to explicitly model dependencies between multiple people, e.g. to prevent the same body part from being assigned to different people.

The upper body detector and tracking software is available at `www.robots.ox.ac.uk/~vgg/research/pose_estimation/index.html`

Acknowledgements. We thank Pawan Kumar for helpful discussions on probabilistic models, Varun Gulshan for the extended GrabCut, and Deva Ramanan for the image parsing code [23]. We are grateful for support from EU Project CLASS, the Swiss National Science Foundation, the Royal Academy of Engineering, Microsoft, and the Spanish Ministry of Education and Science (grant FPU AP2003-2405, project TIN2005-01665).

References

1. Agarwal, A., Triggs, B.: 3d human pose from silhouettes by relevance vector regression. In: CVPR (2004)
2. Agarwal, A., Triggs, B.: Tracking articulated motion using a mixture of autoregressive models. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3023, pp. 54–65. Springer, Heidelberg (2004)

[2] The recent work of [32] proceeds without background subtraction at test time, but uses it for training.

3. Andriluka, M., Roth, S., Schiele, B.: People-tracking-by-detection and people-detection-by-tracking. In: CVPR (2008)
4. Bishop, C.: Pattern recognition and machine learning. Springer, Heidelberg (2006)
5. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: ICCV (2005)
6. Bray, M., Kohli, P., Torr, P.: Posecut: Simultaneous segmentation and 3d pose estimation of humans using dynamic graph-cuts. In: ECCV (2006)
7. Dalal, N., Triggs, B.: Histogram of Oriented Gradients for Human Detection. In: CVPR, vol. 2, pp. 886–893 (2005)
8. Davis, J., Bobick, A.: The representation and recognition of action using temporal templates. In: CVPR (1997)
9. Felzenszwalb, P., Huttenlocher, D.: Pictorial structures for object recognition. IJCV 61(1) (2005)
10. Ferrari, V., Marin-Jimenez, M., Zisserman, A.: Progressive search space reduction for human pose estimation. In: CVPR (June 2008)
11. Ferrari, V., Tuytelaars, T., Van Gool, L.: Real-time affine region tracking and coplanar grouping. In: CVPR (2001)
12. Gammeter, S., Ess, A., Jaeggli, T., Schindler, K., Van Gool, L.: Articulated multi-body tracking under egomotion. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 816–830. Springer, Heidelberg (2008)
13. Ikizler, N., Duygulu, P.: Human action recognition using distribution of oriented rectangular patches. In: ICCV workshop on Human Motion Understanding (2007)
14. Jojic, N., Winn, J., Zitnick, L.: Escaping local minima through hierarchical model selection: Automatic object discovery, segmentation, and tracking in video. In: CVPR (2006)
15. Kumar, M.P., Torr, P.H.S., Zisserman, A.: Learning layered pictorial structures from video. In: ICVGIP, pp. 148–153 (2004)
16. Kumar, M.P., Torr, P.H.S., Zisserman, A.: Learning layered motion segmentations of video. In: ICCV (2005)
17. Laptev, I.: Improvements of object detection using boosted histograms. In: BMVC (2006)
18. Laptev, I., Perez, P.: Retrieving actions in movies. In: ICCV (2007)
19. Lin, Z., Davis, L., Doermann, D., DeMenthon, D.: An interactive approach to pose-assisted and appearance-based segmentation of humans. In: ICCV workshop on Interactive Computer Vision (2007)
20. Mori, G., Ren, X., Efros, A., Malik, J.: Recovering human body configurations: Combining segmentation and recognition. In: CVPR (2004)
21. Niebles, J., Fei-Fei, L.: A hierarchical model model of shape and appearance for human action classification. In: CVPR (2007)
22. Ozuysal, M., Lepetit, V., Fleuret, F., Fua, P.: Feature harvesting for tracking-by-detection. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3953, pp. 592–605. Springer, Heidelberg (2006)
23. Ramanan, D.: Learning to parse images of articulated bodies. In: NIPS (2006)
24. Ramanan, D., Forsyth, D.A., Zisserman, A.: Strike a pose: Tracking people by finding stylized poses. In: CVPR, vol. 1, pp. 271–278 (2005)
25. Rother, C., Kolmogorov, V., Blake, A.: Grabcut: interactive foreground extraction using iterated graph cuts 23(3), 309–314 (2004)
26. Schroff, F., Criminisi, A., Zisserman, A.: Single-histogram class models for image segmentation. In: Kalra, P.K., Peleg, S. (eds.) ICVGIP 2006. LNCS, vol. 4338, pp. 82–93. Springer, Heidelberg (2006)
27. Shechtman, E., Irani, M.: Matching local self-similarities across images and videos. In: CVPR (2007)

28. Sigal, L., Bhatia, S., Roth., S., Black, M., Isard, M.: Tracking loose-limbed people. In: CVPR (2004)
29. Sigal, L., Black, M.J.: Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. In: CVPR, vol. 2, pp. 2041–2048 (2006)
30. Sivic, J., Everingham, M., Zisserman, A.: Person spotting: video shot retrieval for face sets. In: Leow, W.-K., Lew, M., Chua, T.-S., Ma, W.-Y., Chaisorn, L., Bakker, E.M. (eds.) CIVR 2005. LNCS, vol. 3568, pp. 226–236. Springer, Heidelberg (2005)
31. Sminchisescu, C., Triggs, B.: Estimating articulated human motion with covariance scaled sampling. In: IJRR (2003)
32. Thurau, C., Hlavac, V.: Pose primitive based human action recognition in videos or still images. In: CVPR (2008)