Athlete Pose Estimation from Monocular TV Sports Footage

Mykyta Fastovets
University of Surrey
Guildford, UK
mykyta.fastovets@surrey.ac.uk

Jean-Yves Guillemaut
University of Surrey
Guildford, UK
j.guillemaut@surrey.ac.uk

Adrian Hilton
University of Surrey
Guildford, UK
a.hilton@surrey.ac.uk

Abstract

Human pose estimation from monocular video streams is a challenging problem. Much of the work on this problem has focused on developing inference algorithms and probabilistic prior models based on learned measurements. Such algorithms face challenges in generalization beyond the learned dataset. We propose an interactive model-based generative approach for estimating the human pose in 2D from uncalibrated monocular video in unconstrained sports TV footage without any prior learning on motion captured or annotated data. Belief-propagation over a spatio-temporal graph of candidate body part hypotheses is used to estimate a temporally consistent pose between key-frame constraints. Experimental results show that the proposed generative pose estimation framework is capable of estimating pose even in very challenging unconstrained scenarios.

1. Introduction

Pose estimation is an important problem and has received considerable interest in the computer vision community [11]. We consider the problem of estimating human pose in 2D from a single view video stream. There are many potential advantages to being able to analyse monocular video streams including lower equipment cost, lower complexity and higher portability of the camera set-up. Additionally, most of the content available to the user, despite recent advances in 3D film and stereoscopic displays, remains single-view. Thus, for most practical purposes only a single video stream of the scene is available for analysis.

Monocular sports footage is challenging due to the complexity of the backgrounds and the highly articulated human body, engaged in fast paced activities. Additional challenges posed by such data include motion blur, low foreground resolution, occlusions and self-occlusions and rapid configuration changes, camera movement and zoom. The goal of the proposed algorithm is not only to deal with such data, but to do so with sufficient reliability to be suitable for use in a broadcast/production environment and can be summarised as follows: *Given a monocular sequence of images with one or more athletes in each image, estimate the full body 2D limb configuration and joint locations of the athletes in each image.*

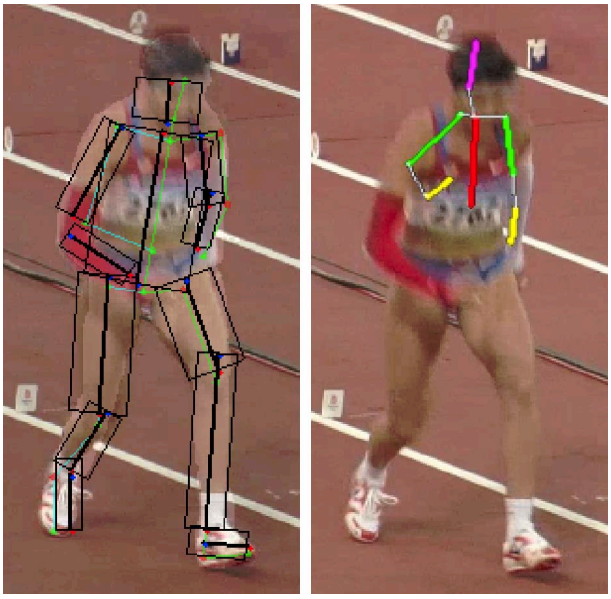

Figure 1. Comparison of the proposed method (left) to an off-the-shelf implementation in [7] (right).

The aim is to provide rapid, yet robust, user-assisted pose estimation and tracking. We opt for a generative model-based approach that requires no offline learning and is generic in the types of motion it is capable of handling but relies on a small amount of human interaction. As the target use scenario is within a production environment with the presence of human operators, complete automation is not critical. Although we focus our experiments on real athletic events footage, no assumptions are made about body size and appearance, type of motion, and the scene.

The algorithm introduces an effective constraint on pose change over time for a full human skeletal model and is capable of estimating pose in unconstrained data requiring

a relatively small number of keyframes.

2. Previous Work

Human pose estimation algorithms in the literature can be separated into two distinct categories - *generative* and *discriminative*. Discriminative approaches are typically data driven and tend to rely on a mapping from image features (silhouettes, salient features etc...) to pose information [5, 1]. The key idea behind discriminative approaches is that the number of typical human poses is far smaller than the number of kinematically possible ones. Discriminative methods learn a model that directly recovers pose estimates from observable image metrics. In other words, the state posterior is modelled directly by learning an image to pose mapping. Once this training process is complete estimating pose is fast but heavily reliant on the training data.

Ren *et al*. [15] propose a framework that exploits symmetry of limbs and parallel lines present throughout the human body. A major drawback of this approach is that without knowledge of scale and appearance the part detector is weak resulting in many false detections. In a similar approach [13] Mori *et al*. use superpixel through normalized graph cuts (or NCut) and low-level salient limb and torso features to produce partial candidate locations for most limbs. The parts are then combined into partial body configurations, which are completed by combinatorially searching the space of superpixels to recover full body pose. Srinivasan and Shi [19] propose a method where a subset of salient shapes detected in an image (via NCut segmentation) is combined into a shape similar to that of the human body. Pose recovery is then formulated as a parsing problem.

Part-based discriminative approaches model the human body as a collection of parts, and the problem changes from estimating the entire human pose to estimating pose of every body part from image metrics. The advantage of part-based approaches is that occlusion is easily modelled and efficient global search techniques can be used. In [16], partial configurations, where some body parts are missing, are allowed into the model. Unfortunately the approach may fail to detect a pose due to similarity of appearance between the person and the background scene.

Generative approaches presuppose an explicitly known parametric body model (generally a kinematic tree) and estimate pose by inverse kinematics or numeric optimization of a model-image correspondence metric over the pose variables, using forward-rendering to predict the images. Bottom-up Bayes' rule is used and the state posterior density is modelled using observation likelihood or a cost function. Decision making can proceed even in the case of missing or incomplete data and assumptions as well as any prior information (such as pose constraints) are easily accommodated, but most importantly, there exists a bilateral dependence between the model and the data. Such algorithms generally incorporate a likelihood model capable of discriminating incorrect hypothetical poses from correct ones based on image evidence and an estimation algorithm able to generate possible pose hypotheses from the parametric body model [9, 10]. Model-based tracking is a derived technique which focuses on tracking the pose estimate from one time step to the next, starting from a known initialization based on an approximate dynamical model. Most generative pose estimation frameworks suffer from the fact that the optimisation is prone to hitting local minima, requiring a good initialisation and often failing on complex motions [2].

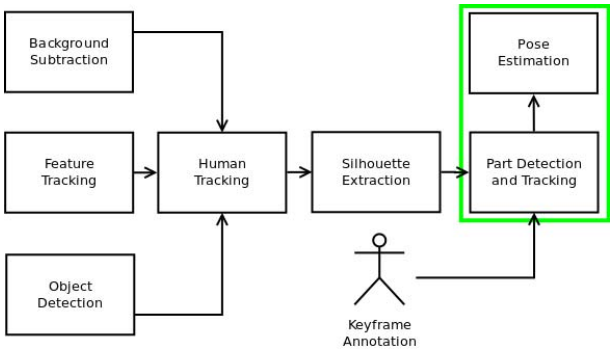

Figure 2. Generic pipeline of a full pose estimation framework. The area in green represents where the proposed algorithm could fit into the pipeline.

Pictorial Structures (PS) is a probabilistic inference for a tree-structured graphical model where the overall cost function for a pose decomposes across edges and nodes of the tree. PS recovers locations scales and orientations of rigid rectangular part templates that represent a body. PS is a well developed detection method and can be used for tracking by detection [6, 3, 4]. A technique to extend existing training data sets is presented in [14]. These methods, however, require a large motion capture and image training dataset. Lan and Huttenlocher [10] use a spatio-temporal model though pictorial structures as states in a Hidden Markov Model(HMM). A recent development [3] introduces the idea that pictorial structures may be a generic multi-purpose solution to pose estimation, detection and tracking. The method is reliant on a strong discriminatively trained appearance model. Ferrari *et al*. [8] propose a temporal link between frames when estimating pose in videos, but the proposed algorithm only works for frontal upper-body poses. A variation on the temporal smoothness term used for estimating 3D pose based on calibrated multi-view ground truth learning data is presented in [18]. Although the smoothness term is similar, it is applied within a three frame moving window rather than globally.

While there exist multiple pose estimation techniques capable of extracting human pose from monocular image

streams, existing approaches do not provide a suitable solution to the problem of general pose estimation in sports footage to the fast and extreme motion captured with moving and zooming cameras. Most algorithms focus on recovering pose from single images and/or do not make full use of the temporal constraint on limb motion [3, 17, 14, 2]. Available off-the-shelf single image methods such as [7] have trouble coping with the difficulty of the data, even when the human is detected. Figure 1 shows a side-by-side comparison of the output of the algorithm developed in [7] obtained using their online demo tool and the proposed method.

We propose a novel generative approach to pose estimation from sports video to avoid loss of generality and bypass the requirement for motion captured training data of specific athlete motions. The focus of the framework is on pose estimation with the assumption that a human is detected and an approximate foreground mask is available. Silhouette extraction is a well studied problem and there is a wide variety of methods that deal with background subtraction, for example [20]. The diagram in Figure 2 shows where our algorithm fits within a complete pose estimation framework.

The contributions of this work are twofold. First, we propose a temporal smoothness term applied to the pictorial structures model and show that this term makes a significant impact on the outcome of the pose estimation problem. This temporal smoothness constraint should be useful in most existing pose estimation algorithms.

Second, we propose a framework for estimating full body pose in truly challenging circumstances by effectively moving the learning stage of the algorithm online. Generally, for a sequence of around 200 images our algorithm will require 20 to 30 keyframes (depending on difficulty of the motion) to produce high quality results. The keyframes provide a motion prior and local part appearance models.

3. Methodology

The input into the algorithm is an image sequence along with a corresponding sequence of masks and several user-created keyframes. The keyframes are used for building appearance models of body parts as well as to provide a body part location prior through motion interpolation. The process of pose estimation consists of two steps: body part detection and pose recovery.

We employ a human skeletal model consisting of $b = 13$ body parts (root (R), head (H), neck (N), right and left femur(RF, LF), tibia, (RT, LT), metatarsal (RM, LM), humerus (RH, LH) and radius (RR, LR) as depicted in Figure 3.

Let $\mathbf{x}_t$ represent the state of the skeleton at time instant t. Let $\mathbf{p}_i$ represent the state (x, y, θ, s) of part i as per the standard PS model [6]. Then $\mathbf{x}_t = \{\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_b\}_t$.

Given a sequence of n frames the problem of estimating human pose at each time instance can be written as:

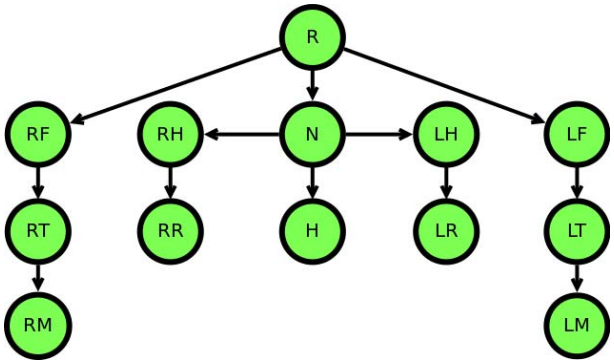

Figure 3. Graph representation of the human kinematic skeletal model consisting of 13 body parts.

$$\mathbf{X} = \mathbf{argmin}_{\mathbf{x}} \alpha \sum_{t=1}^{n} \mathbf{D}(\mathbf{x}_t) + (1-\alpha) \sum_{t=1}^{n} \mathbf{S}(\mathbf{x}_t, \mathbf{x}_{t-1}) \quad (1)$$

where $\mathbf{X}$ represents the temporal sequence of poses $\{\mathbf{x}_1, \mathbf{x}_2 ... \mathbf{x}_t\}$, $\mathbf{D}(\mathbf{x}_t)$ represents the data term and $\mathbf{S}(\mathbf{x}_t, \mathbf{x}_{t-1},)$ the smoothness term at time instance t. The data term evaluates the cost of a state $\mathbf{x}_t$ with respect to image observations, while the smoothness term evaluates the temporal continuity of the sequence. Sections 3.1 and 3.2 explain in more detail how these terms are computed. In order to solve the problem we first represent it as a factor graph and solve the resulting graph using an implementation of generalised Belief Propagation [12].

3.1. Smoothness Term

The smoothness term consists of two distinct components: an inter-limb distance cost $\mathbf{J}_{\mathbf{x}_t}$, representing joint connectivity, and a temporal distance term $\mathbf{T}_{\mathbf{x}_t}$, representing smoothness of motion across time:

$$\mathbf{S}(\mathbf{x}_t, \mathbf{x}_{t-1}) = \beta \mathbf{J}(\mathbf{x}_t) + (1-\beta)\mathbf{T}(\mathbf{x}_t, \mathbf{x}_{t-1}) \quad (2)$$

The joint distance cost for skeleton $\mathbf{x}_t$ is given by:

$$\mathbf{J}(\mathbf{x}_t) = \sum_{k=1}^{b} dist(\mathbf{b}_k, par(\mathbf{b}_k)) \quad (3)$$

where the $dist()$ function evaluates the distance between corresponding joints and par gives the hierarchical parent of bone $\mathbf{b}_k$. The root bone has no parent. This creates a soft requirement on inter-part connectivity. The assumption here is based on the physical reality that adjacent limbs should be connected. This term alone, however, is not likely to lead to desirable solutions due to frame to frame jitter.

In order to reduce jitter between frames, a temporal smoothness term is introduced:

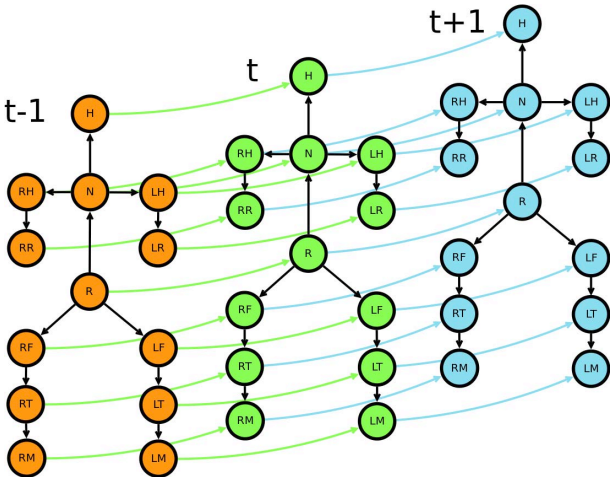

Figure 4. Representation of temporal and inter-limb links for frames at time instances $t-1$, t and $t+1$. Coloured links represent the temporal constraint, while those in black represent the joint connectivity constraint.

$$\mathbf{T}(\mathbf{x}_t, \mathbf{x}_{t-1}) = d(\mathbf{x}_t, \mathbf{x}_{t-1}) \tag{4}$$

The temporal term is effectively the Euclidean distance between the joint locations at t and $t-1$. Figure 4 illustrates the complete graph including the temporal term for three consecutive frames.

3.2. Data Term

The data term is defined simply as the sum of support scores $\mathbf{S}(k_{\mathbf{x}_t})$ assigned to each generated part candidate for configuration $\mathbf{x}_t$:

$$\mathbf{D}(\mathbf{x}_t) = \sum_{k=1}^{b} \mathbf{S}(k_{\mathbf{x}_t}) \tag{5}$$

To generate part candidates for part k a search region centred around the midpoint between the two joints of the interpolation for that part is created. At each pixel within the search region and the foreground mask a rectangular template $\mathbf{R} = \{p_0, p_1, ..., p_q\}$ containing q pixels is centred with varying orientations and scales. Each instance of the template receives a support score $\mathbf{S}$ based on the strength of the support region as detailed in Equation 6.

$$\mathbf{S}(k) = \frac{\sum_{i=0}^{n} \mathbf{d}(p_i)_k}{q} \tag{6}$$

To obtain the descriptor $\mathbf{d}$ in Equation 6, an 8^3 RGB colour histogram for each body part using information from the two closest user-created keyframes is first built. The histogram is sampled from image data at 8 bits per colour channel and is normalised by the total number of samples. The histogram is accumulated across all regions within the foreground in a defined area around each bone. This normalised 3D colour histogram serves as an appearance model for body parts. Figure 5 is an example of regions from which the histogram model is built while Figure 6 shows a sample histogram of a foot model.

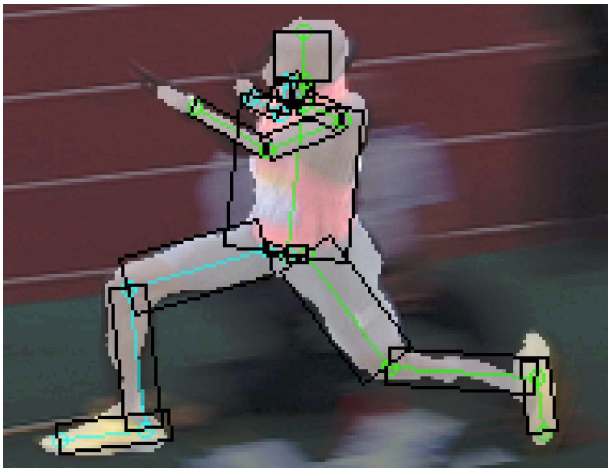

Figure 5. Image with mask overlay and keyframe skeleton with part regions that part models are built from.

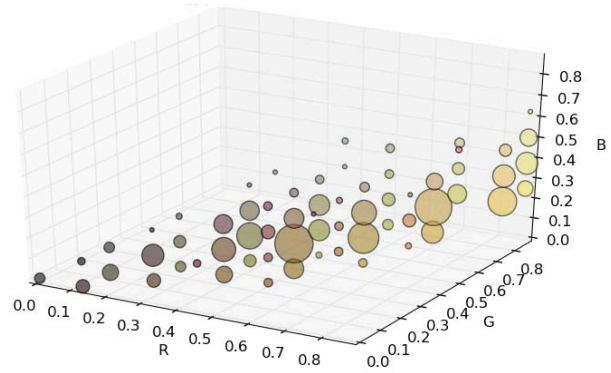

Figure 6. Sample foot histogram model. Blob size represents frequency and blob colour represents the actual bin colour.

The histogram models provide k colour distribution functions, one for each body part. A k-dimensional vector is built at every pixel of the foreground region with each element giving the likelihood of the pixel belonging to a given part based on colour using the colour distribution functions previously obtained. This vector is then normalised by the maximum of its elements, creating k-dimensional descriptor $\mathbf{d}$ assigned to the pixel p(x, y) where each element of the descriptor is the probability of the pixel belonging to the corresponding body part.

Additionally, a 3D motion interpolation is computed using the two closest keyframes. The interpolation limits the search region and provides initial orientation and scale estimates for body part candidate generation. Interpolation is used to limit the search space and make the problem tractable on commodity hardware. The interpolation is

computed using the manually annotated keyframes based on locations of joints in 3D space. Joint angles are represented in quaternions and we use linear interpolation to estimate the body state at a time instance.

4. Results and Evaluation

The framework is evaluated on three sports sequences with different camera angles, zoom and motion from two different sports - triple jump and hurdles. Due to the fast motion of the athlete's body many frames of suffer from motion blur effects. Table 1 summarises the data used to evaluate the framework and Figures 12, 13, 14, show sample results from each sequence. The size of the athlete (height in pixels) differs in every sequence, ranging from under 100 to over 500.

Table 1. Summary of sequences used for framework evaluation.

Name	Sport	Resolution	Frames	Keyframes
hurdlesSD	hurdles	720x576	76	17
triplejumpHD	triple jump	1920x1080	198	42
triplejumpSD	triple jump	720x288	89	12

We use the interpolation between keyframes as the motion prior and assume that it gives a reasonable initial estimate. In order to help reduce the search space we do not vary scale and vary limb orientation by 90° in each direction in increments of 10°. We also limit the search region for body parts to twice the interpolated limb length. This strikes a good balance between accuracy and computability allows us to obtain results for 10 images (8 frames plus two keyframes) at speeds of around 15 seconds per frame on a commodity laptop computer, depending on foreground region size.

The error of a body part to ground truth is defined as the square of Euclidean distance to ground truth body part joints and is given by:

$$\mathbf{E_k} = \frac{d(j_0, g_0)^2 + d(j_1, g_1)^2}{2} \tag{7}$$

where j_0 and j_1 are estimated joint positions for the body part k and g_0 and g_1 are the ground truth positions. The graphs in Figures 9, 10, 11, show a comparison of RMS error at each time instance between 3D interpolation and our method for each of the three sequences.The ground truth is obtained by manual annotation of the sequences. The manual annotation of the sequence includes labelling of limbs as occluded to the point where even a human operator cannot reliably determine the location of the limb. Limbs labelled in this way are removed from error analysis as their true location is unknown.

The parameters used for obtaining the results were experimentally derived using a short section of the triplejumpSD sequence. Testing has shown that values of $\alpha = 0.2$ and $\beta = 0.7$ (Equations 1 and 2) produce good results. Graphs in figures 7 and 8 show our findings for the triplejumpSD sequence. To determine the optimal value of α the sequence was solved using the same set of candidate parts with $\beta = 0$ at intervals of 0.05 from 0 to 1. Similarly, to determine the value of β, the value of α was fixed to 0.2 and the same candidates were used with β being varied in the same range with identical frequency. This experiment also serves to confirm that the temporal term has a quantifiable impact on the quality of the results.

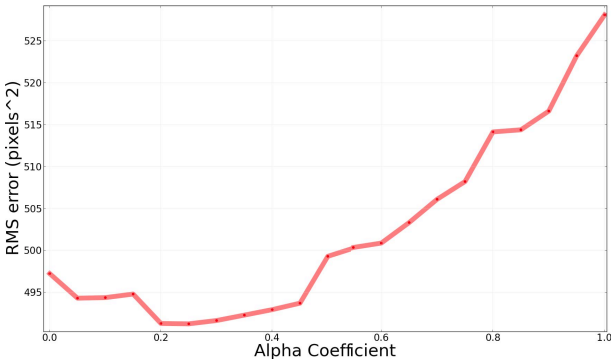

Figure 7. Error to ground truth of our algorithm for the trijumpSD sequence at different α values.

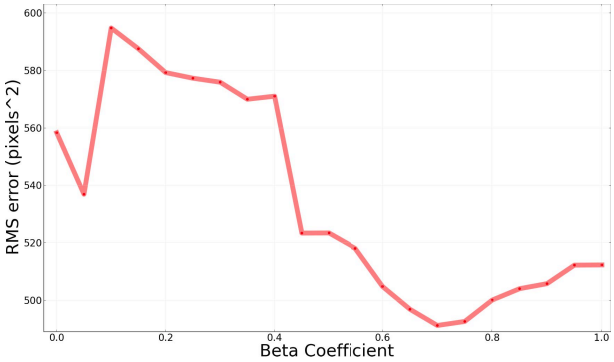

Figure 8. Error to ground truth of our algorithm for the trijumpSD sequence at different β values.

Analysis of these sequences indicates that qualitatively our method is almost always better than pure interpolation. The significance of improvement can depend on how good interpolation is for a particular sequence. Since the interpolation we compare against is done on a 3D skeleton it is fairly good at coping with normal human motion and is guaranteed connectivity and temporal consistence as long as it is across one gait cycle. Currently, we make use of interpolation as a motion prior and search space limiter with the assumption that the interpolation provides a reasonable initial estimate of pose. Thus, even a seemingly small improvement in distance often has significant visual impact with regard to accuracy of body part locations. Furthermore the problem of the interpolation straying significantly from the ground truth, preventing detection, and potentially drag-

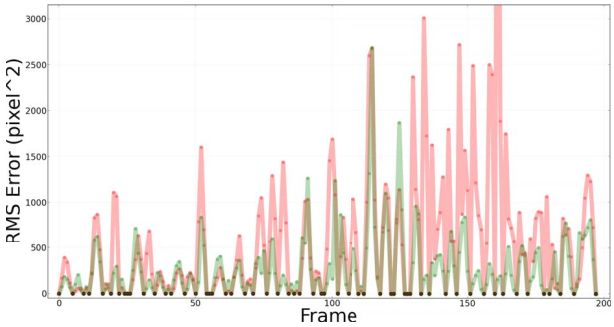

Figure 9. Error to ground truth of our algorithm (green) and motion interpolation (red) for the triplejumpHD sequence.

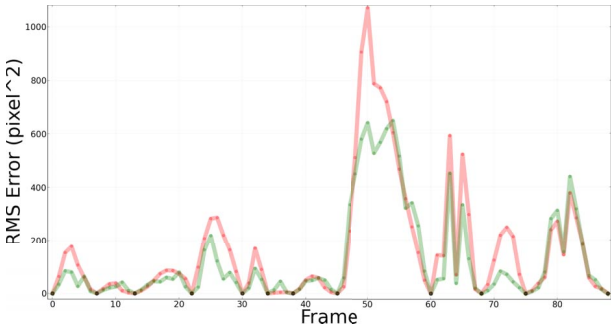

Figure 10. Error to ground truth of our algorithm (green) and motion interpolation (red) for the triplejumpSD sequence.

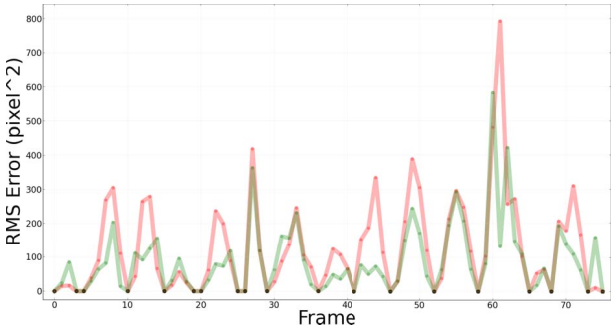

Figure 11. Error to ground truth of our algorithm (green) and motion interpolation (red) for the hurdlesSD sequence.

ging the optimisation in the wrong direction is left for future work but has a definite impact on the results.

The error differences between interpolation and our method vary for the different sequences and sections of sequences. For example, the graph for the triplejumpHD sequence 9 has a much more significant error gap than both triplejumpSD and hurdlesSD. Large gaps can occur when the interpolation has strayed 90 degrees of rotation away from ground truth (at the limit of our used range) but has stayed close enough to contain the image body parts within the search region.

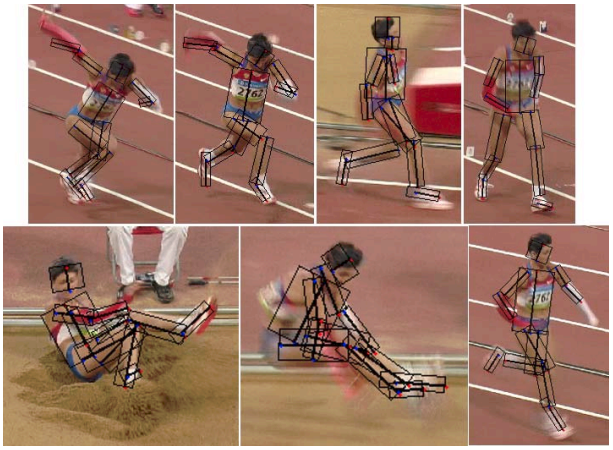

Figure 12. Sample solutions from the trijumpHD sequence. Black marks at zero error indicate keyframes.

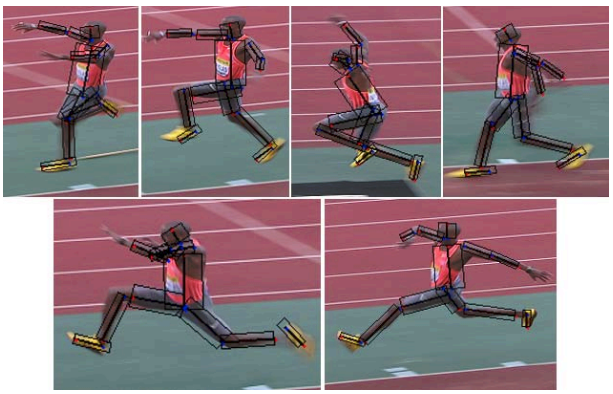

Figure 13. Sample solutions from the trijumpSD sequence. Black marks at zero error indicate keyframes.

5. Conclusion and Future Work

We have presented a method for recovering human pose in challenging sports scenarios using only a single view but requiring some human interaction. The proposed framework is generic in terms of types of motion and pose the athlete could take. The algorithm has been tested on three challenging sports sequences of different sports. Quantitative and qualitative analyses have shown that our method provides a significant improvement to using interpolation and is capable of recovering pose even in the most challenging of conditions.

Future work will focus on improving the appearance model as a stepping stone to weakening the reliance of the algorithm on interpolation and reducing the solution search space by generating higher quality body part candidates. Since our algorithm is capable of recovering joint locations another avenue for exploration is recovery of pose in 3D, which would allow for additional kinematic constraints to be added to the optimisation.

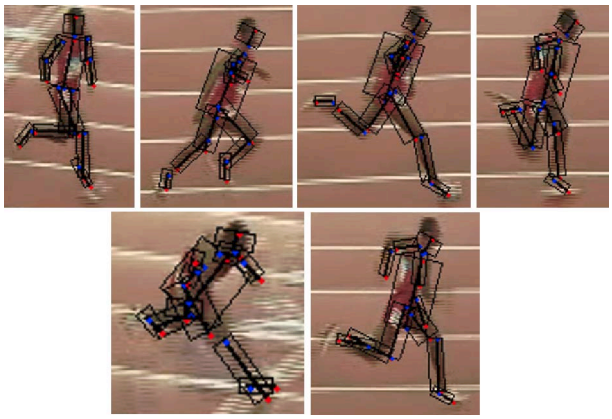

Figure 14. Sample solutions from the hurdlesSD sequence. Black marks at zero error indicate keyframes.

Acknowledgements: The authors would like to thank BBC R&D, Graham Thomas and Robert Dawes for supporting this research and providing data.

References

[1] A. Agarwal and B. Triggs. Recovering 3d human pose from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence. Proceedings.*, 28(1):44–58, 2006. 2

[2] P. Agarwal, S. Kumar, J. Ryde, J. Corso, and V. Krovi. Estimating human dynamics on-the-fly using monocular video for pose estimation. In *Proceedings of Robotics: Science and Systems*, Sydney, Australia, July 2012. 2, 3

[3] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 2, 3

[4] M. Andriluka, S. Roth, and B. Schiele. Monocular 3d pose estimation and tracking by detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, San Francisco, USA, 06/2010 2010. 2

[5] A. Bissacco, M.-H. Yang, and S. Soatto. Fast human pose estimation using appearance and motion via multi-dimensional boosting regression. In *IEEE Conference on Computer Vision and Pattern Recognition.*, pages 1–8, June 2007. 2

[6] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *International Journal on Computer Vision*, 61(1):55–79, January 2005. 2, 3

[7] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *Computer Vision and Pattern Recognition, 2008.IEEE Conference on*, pages 1–8, 2008. 1, 3

[8] V. Ferrari, M. Marn-jimnez, and A. Zisserman. 2d human pose estimation in tv shows. In *In Dagstuhl post-proceedings*, 2009. 2

[9] H. Jiang. Human pose estimation using consistent max-covering. In *International Conference on Computer Vision*, 2009. 2

[10] X. Lan and D. Huttenlocher. A unified spatio-temporal articulated model for tracking. In *IEEE Conference on Computer Vision and Pattern Recognition. Proceedings.*, volume 1, pages I–722–I–729 Vol.1, June-2 July 2004. 2

[11] T. B. Moeslund, A. Hilton, V. Krüger, and L. Sigal, editors. *Visual Analysis of Humans - Looking at People*. Springer, 2011. 1

[12] J. M. Mooij. libDAI: A free and open source C++ library for discrete approximate inference in graphical models. *Journal of Machine Learning Research*, 11:2169–2173, Aug. 2010. 3

[13] G. Mori, X. Ren, A. A. Efros, and J. Malik. Recovering human body configurations: Combining segmentation and recognition. In *IEEE Conference on Computer Vision and Pattern Recognition. Proceedings.*, pages 326–333, 2004. 2

[14] L. Pishchulin, A. Jain, M. Andriluka, T. Thormaehlen, and B. Schiele. Articulated people detection and pose estimation: Reshaping the future. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 1–8, Providence, United States, June 2012. IEEE, IEEE. 2, 3

[15] X. Ren, A. C. Berg, and J. Malik. Recovering human body configurations using pairwise constraints between parts. volume 1, pages 824–831 Vol. 1, October 2005. 2

[16] T. J. Roberts, S. J. McKenna, and I. W. Ricketts. Human pose estimation using partial configurations and probabilistic regions. *International Journal on Computer Vision. Proceedings.*, 73(3):285–306, 2007. 2

[17] B. Sapp, D. Weiss, and B. Taskar. Parsing human motion with stretchable models. In *CVPR*, 2011. 3

[18] L. Sigal, S. Bhatia, S. Roth, M. J. Black, and M. Isard. Tracking loose-limbed people, 2004. 2

[19] P. Srinivasan and J. Shi. Bottom-up recognition and parsing of the human body. In *Computer Vision and Pattern Recognition. Proceedings.*, pages 1–8, June 2007. 2

[20] T. Zhao and R. Nevatia. Bayesian human segmentation in crowded situations. In *IEEE Conference on Computer Vision and Pattern Recognition. Proceedings.*, volume 2, pages II–459–66 vol.2, June 2003. 3