LABELLED SILHOUETTES FOR HUMAN POSE ESTIMATION

Daniel Chen, Clinton Fookes

Image and Video Research Laboratory
Queensland University of Technology
GPO Box 2434, Brisbane, Queensland 4001

ABSTRACT

This paper proposes a new method of using foreground silhouette images for human pose estimation. Labels are introduced to the silhouette images, providing an extra layer of information that can be used in the model fitting process. The pixels in the silhouettes are labelled according to the corresponding body part in the model of the current fit, with the labels propagated into the silhouette of the next frame to be used in the fitting for the next frame. Both single and multi-view implementations are detailed, with results showing performance improvements over only using standard unlabelled silhouettes.

1. INTRODUCTION

The ability to capture the pose of a person leads to a variety of interesting applications such as for computer animation, person identification though the use of gait analysis, action recognition, or as a means of human-computer interactions. Traditionally, this 'motion capturing' has been performed by attaching markers to specific parts of a subject's body. The nature of these makers (eg. LEDs) allow their locations to be easily identified in an image, and given multiple calibrated cameras, enable accurate triangulation of the position in 3D space.

This setup is obtrusive due to the use of markers, limiting the use to prearranged capture sessions. This has prompted the development of marker-less motion capture, using, ideally, only conventional video cameras and computer hardware. This greatly expands the possible applications, such as in surveillance where a possible application would be to identify people through their gait. The continuous rapid increase in computational power has now made this a possibility.

Many different methods have been developed to try to estimate the human pose without the use of markers. Agarwal and Triggs [1] used an example-based approach to recover 3D pose from monocular video. Silhouettes are extracted and a nonlinear regression was used to learn mappings between silhouette shape descriptors to a pose.

Bottom-up approaches to pose estimation attempt to first find individual body parts and then compile them into a human body. Mori *et al*. [2] segmented an image based on edges and applied classifiers to them in an attempt to identify the individual body parts. Ren *et al*. [3] identifies parallel lines from an edge map and then applies pairwise constraints between body parts to assemble these lines into a human body.

Efros *et al*. [4] uses a holistic method to perform action recognition. Based on the detected action, an example pose sequence is then transferred to give a rough estimate of the pose of the subject.

Model based approaches use a model of a human body and attempt to fit this to the observed data. Thome *et al*. [5] skeletonised the silhouette image and then tried to recover the pose based on the configuration of the branches. The skeleton was decomposed into a directed acyclic graph and matched to a graphical model of a human body. Menier *et al*. [6] performs the skeletonisation in 3D on the visual hull created from silhouettes generated from multiple cameras. A 3D skeleton model was then fitted to estimate pose in 3D.

Gavrila and Davis [7] built a volumetric model out of super-quadrics. The model edges are projected into multiple views where they are matched against the chamfer image (distance transform of edge maps).

Deutscher *et al*. [8] uses a simple model comprised of cylinders and produces accurate tracking through the use of annealed particle filtering. The model is matched to both silhouettes and edge images.

It is with this type of approach to pose estimation that the system presented here in this paper will follow. Specifically, it will focus on the silhouette matching problem and try to increase the amount of information that can be used from them. Both single and multi-view implementations will be presented with the results derived from experiments performed on publicly available datasets.

2. SILHOUETTE LABELLING FOR POSE ESTIMATION

The algorithm presented in this paper attempts to estimate the pose of a person by fitting a human body model to that of the silhouette of the person extracted from images. The model is projected into the available camera views where a fitness function is computed to determine how well the model matches the silhouette. Model parameters are tweaked such that this fitness is maximised. Unlike other pose estimation implementations that use this method, the silhouette will not be treated as a binary image. Instead, the silhouette will be labelled based on the body parts in the model that fit it such that this information

Fig. 1. **Body Models** *left* 2D model, side view. *centre* 2D model, front view. *right* 3D model.

can be used to improve the fitting process in subsequent frames.

2.1. Body Model

The models used consists of 12 distinct regions; the head, torso, left and right arms, forearms, thighs, legs and feet. These make up the labels that are transferred to the silhouettes. In the single view case, where the subject's orientation relative to the camera is assumed to be static, either front or side on, 2D boxes with rounded off ends are used to construct model. Attachment points of the limbs to the torso vary between the two cases. This model consists of 15 degrees of freedom (DOF), one for each joint angle along with global rotation, x/y position and scale.

For the multi-view case, a volumetric model consisting of truncated cones and boxes are used. Some joints, such as the shoulder, now can have up to 3 DOF each, brining the total DOF up to 27.

A visualisation of the models can be seen in Figure 1.

2.2. Silhouette Labelling

Assuming a good initial fit of the model, the pixels in the silhouettes of each view are assigned an identifier based on the projection of the model onto the silhouettes. Pixels are given labels corresponding to the closest body part in the model.

As the labels are derived from the current model fitting, further optimisation based on the now labelled silhouette is meaningless. They can, however, can be used to drive the fitting in subsequent frames. In order for this to be useful, the labels need to be transferred onto the silhouette of the next frame.

Optical flow is used to map the pixels in the current frame to the next, transferring the labels along with it. Pixels that fail to fall inside the new silhouette are ignored. Regions of the silhouette that are not given a label are left alone (given a neutral label) to prevent erroneously labelling a part that may be coming out of self occlusion.

Now that the silhouette has been labelled, model fitting can be performed on this new frame. Once the model parameters have been optimised to an acceptable level, the silhouette is re-labelled based on the new fit and the algorithm reiterates. A flowchart of this can be seen in Figure 2. Figure 3 shows example images of each step.

2.3. Model Fitting and Tracking

Pose estimation is achieved by projecting the model into each of the views and optimising the fit of the model with the observation, which in this case would be the labelled silhouettes. Given observation y and model parameters x, optimisation is performed by maximising $f(y|x)$. When dealing with a standard binary silhouette, $f(y|x)$ can be defined as

$$f(y|x) = \frac{1}{N}\Sigma_b, \tag{1}$$

where N is the number of foreground pixels in the silhouette and Σ_s is the number of overlapping pixels between the silhouette and the model (sum of the logical AND of the two).

A means of determining errors in the model fit needs to be formulated, specifically for when parts of the model appear over an area of the silhouette that isn't considered foreground. Matching errors where the model only slightly extend beyond the silhouette can be attributed to segmentation errors and the imperfect modelling of the subject, and thus needs to be tolerated. Should an entire limb stick out far from the silhouette, however, heavy penalties should be applied. To achieve this, a distance transform of the silhouette is computed, where values in the map correspond to the closest distance to the silhouette. Σ_e is established which is the sum of the distance values for the pixels where the model lies outside the silhouette. The fitness function now becomes

$$f(y|x) = \frac{1}{N}\left(\omega_s\Sigma_s + \omega_e\Sigma_e\right), \tag{2}$$

where ω are weight values. As errors needs to be minimised, ω_e requires to be negative.

For labelled silhouettes, Σ_l is introduced which represents the pixels where its label matches its corresponding body part. As label transferring process is not perfect, it would be beneficial to still consider pixels where labels do not match yet are still within the silhouette, thus keeping the Σ_s term.

The final fitness function $f(y|x)$ becomes

$$f(y|x) = \frac{1}{N}\left(\omega_l\Sigma_l + \omega_s\Sigma_s + \omega_e\Sigma_e\right). \tag{3}$$

Matching edge maps can also be used. However, the nature of the video sequences used limits their effectiveness as explained later in Section 3.

For optimisation, the Metropolis-Hastings algorithm is used.

As the optical flow has been calculated, it will be used to perform an initial estimate of the model parameters in the new frame; joint locations follow the flow vectors to their new locations. No motion models or other tracking was used.

3. EXPERIMENTS

Two publically available datasets have been used, the IXMAS dataset from INRIA [9] and one from the Weizmann

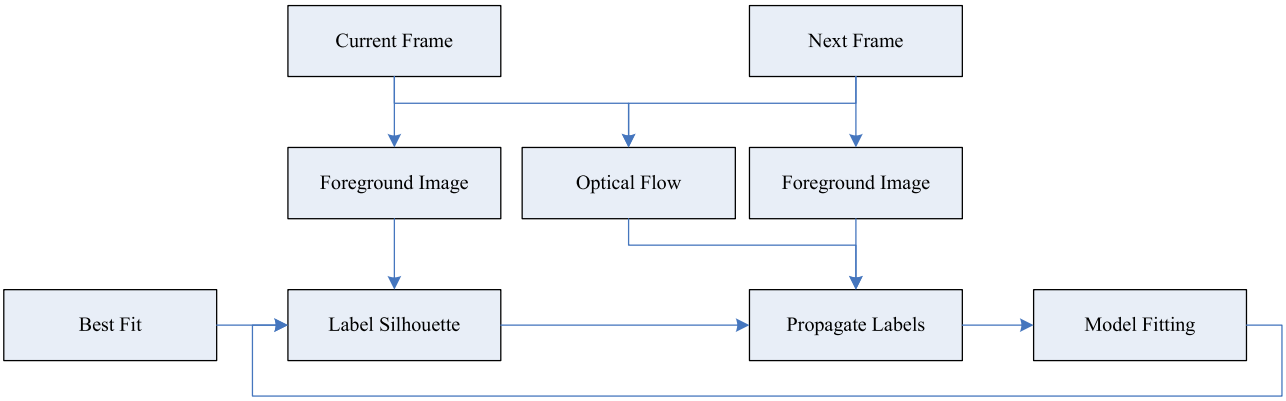
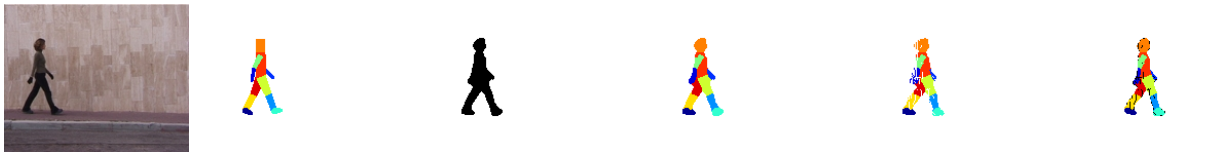


Fig. 2. **Flowchart**



(a) (b) (c) (d) (e) (f)

Fig. 3. **Silhouette Labelling** (a) Original image. (b) Initial fit. (c) Foreground silhouette. (d) Labelled silhouette. (e) Labels propagated based on optical flow. (f) Labels transferred onto silhouette of next frame.

Institute of Science (WIS) [10]. Both datasets are designed for the application of action recognition, with the IXMAS dataset being multi-view consisting of 5 cameras while the WIS dataset is taken from a single camera viewpoint either from the front or side of the subject depending on which action is performed.

The low resolution of the video in the case of the WIS dataset and the generally dark clothing worn by the subjects in the IXMAS dataset prevent clean edge maps to be extracted. As such, edge information will not be used in the matching process.

To test the effectiveness of silhouette labelling for pose estimation, the algorithm was applied to the above datasets.

4. RESULTS AND ANALYSIS

Figure 4 shows an example test sequence. The model was able to fit the silhouettes quite well, though there are problems in distinguishing between the left and right limbs. This can be attributed to the lack of a tracking framework.

Without the use of a robust tracking system, the implementation fails where significant self occlusion occurs. The silhouette labelling system actually performs worse compared to the one without under such conditions. This is likely due to the fact that any errors in the fitting are accumulated to some degree by the labelling process, as subsequent frames are forced somewhat to converge on the flawed fitting in previous frames. During periods of significant self occlusion, these errors cascade, resulting in catastrophic failure. An example of this can be seen in Figure 5.

Ignoring the computation for the optical flow, the sil-

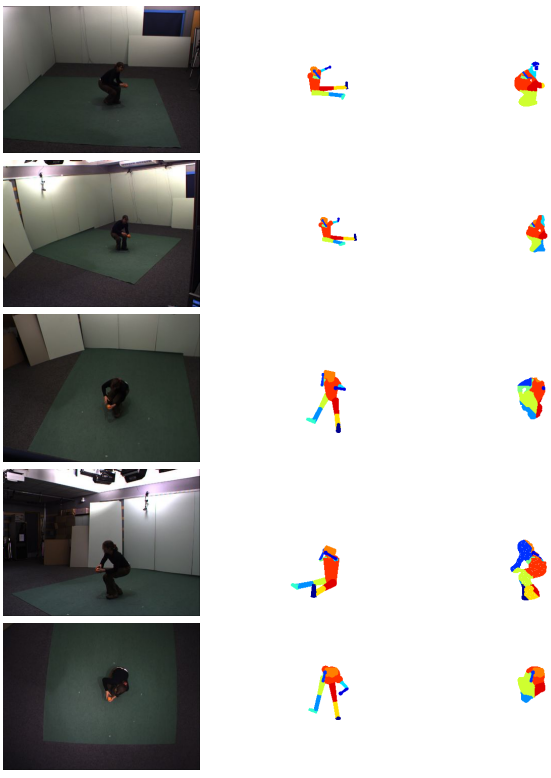


Fig. 5. **Example Failure**

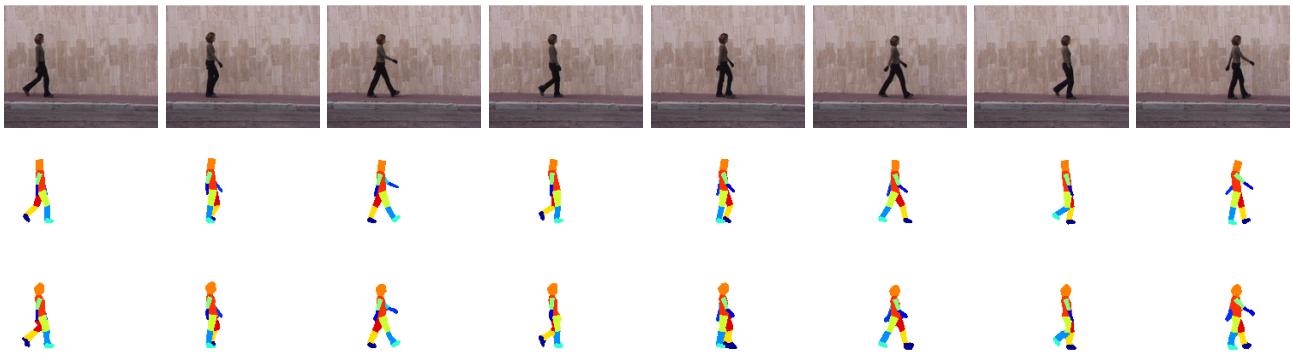


Fig. 4. **Example Sequence**

houette labelling approach adds only small amount of extra processing time compared to the unlabelled implementation. In the test setup used, the increase is about 12%. Under non occluding conditions, the labelled silhouette requires less iterations of the optimisation routine to achieve a similar fit. As a result, the new algorithm is able to converge faster towards the optimal model fit.

5. CONCLUSIONS AND FUTURE WORK

It has been shown that labelling silhouettes for model matching to estimate human pose provide some benefits over normal silhouette methods. It is able to converge towards the optimal fit faster, however, is also more prone to errors.

The use of a robust tracking framework is essential to a successful pose estimation system. It is intended that this algorithm be integrated into a particle filter based tracking framework, such as the one in [8]. Its support of multiple hypotheses gives the system the opportunity to recover from error. However, due to probabilistic nature of the tracker, the current method of 'hard' labelling the silhouette cannot be used. A more fuzzy approach to silhouette labelling will need to be devised to incorporate it into such a system.

6. REFERENCES

[1] A. Agarwal and B. Triggs, "Recovering 3d human pose from monocular images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 1, pp. 44–58, 2006.

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

[3] X. Ren, A. C. Berg, and J. Malik, "Recovering human body configurations using pairwise constraints between parts," in *Proceedings International Conference on Computer Vision*, vol. 1, 2005, pp. 824–831.

[4] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Proceedings International Conference on Computer Vision*, vol. 2, 2003, pp. 726–733.

[5] N. Thome, D. Merad, and S. Miguet, "Human body part labeling and tracking using graph matching theory," in *Proceedings International Conference on Video and Signal Based Surveillance*, 2006.

[6] C. Menier, E. Boyer, and B. Raffin, "3d skeleton-based body pose recovery," in *Proceedings International Symposium on 3D Data Processing, Visualization, and Transmission*, 2006, pp. 389–396.

[7] D. M. Gavrila and L. S. Davis, "Towards 3-D model-based tracking and recognition of human movement," *International Workshop on Face and Gesture Recognition*, pp. 272–277, 1995.

[8] J. Deutscher, A. Blake, and I. Reid, "Articulated body motion capture by annealed particle filtering," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2000, pp. 126–133.

[9] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer Vision and Image Understanding*, vol. 104, pp. 249–257, 2006.

[10] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Proceedings International Conference on Computer Vision*, vol. 2, 2005, pp. 1395–1402.