Journal of Pattern Recognition Research 1 (2014) 123-137

Received November 5, 2014. Revised December 11, 2014. Accepted December 7, 2014.

A Body Part Segmentation System for Human Activity Recognition in Videos

Matti Matilainen *matti.matilainen@ee.oulu.fi*
Department of Computer Science and Engineering, University of Oulu PL 4500, 90014 Oulun Yliopisto, Finland

Mark Barnard *mark.barnard@surrey.ac.uk*
Centre for Vision, Speech and Signal Processing University of Surrey, Guildford Surrey, GU2 7XH, UK

Jari Hannuksela *jari.hannuksela@ee.oulu.fi*
Department of Computer Science and Engineering, University of Oulu PL 4500, 90014 Oulun Yliopisto, Finland

Abstract

Identification of body parts is an important first step for many tasks such as action recognition in automatic surveillance systems. In this paper, we present a body part segmentation system for image and video analysis. The proposed system utilises Hidden Markov Models and modified shape context features for statistical modeling of the human body shape. In our solution, we also demonstrate how a general and robust solution can be developed with the synthetically generated training data. The sequences of synthetically generated images are generated using three dimensional rendering and motion capture information. After the training phase, the model is used to segment silhouette images into four body parts; arms, legs, body and head. In experiments, the system is successfully used in body part segmentation, unusual activity detection in surveillance applications and arm swing detection in gait analysis. The advantages of the method include that the same model can be employed without any modifications of parameters after initial training.

Keywords: computer vision, body part recognition, activity recognition, gait analysis, statistical modeling

1. Introduction

In automatic video analysis and surveillance the computing can be performed by sensors almost unnoticeably integrated in the environment. People are tracked, for example, with cameras during their normal day routines. One possible application is monitoring elderly in home to raise alarms when they fall over and need help. Identification of body parts is an important first step for many camera based activity recognition approaches. The problem with these solutions is that they usually have to be trained for each installation location separately. This increases the complexity and cost of the system significantly. Therefore, a robust recognition method that requires neither training nor adjustment of model parameters in a new location is obviously desired. In addition, pattern classification usually requires a lot of labeled training data to be able to generalise enough to recognise patterns it has never seen before. In some cases there is need for a very large training set that requires infeasible manual labeling.

In our body part identification system, we utilise Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM) to segment the body parts from background subtracted silhouette images. The body part identification algorithm gives an estimate of how well the given silhouette corresponds to the ones in the training data. This information is

used at higher level to detect the human activities. To deal with the problem of a lack of labeled training data, we propose the use of artificial videos to generate a large number of examples for human activity recognition. The data is created by animating a pre-labeled 3D model from various viewing angles. Compared to our preliminary early work [34, 35] on this topic, we propose an improved segmentation solution with the new comprehensive test results in this paper. The experiments show that the used shape context features and the statistical method work well in modeling the actions in very noisy videos and under occlusions. The body part segmentation system can detect the arm swings from USF gait recognition database by analysing the ratio of recognised arm and body pixels and the width of leg stride in the frequency domain. We also present the complete exhaustive testing of the system and results on gait analysis.

The rest of the work is structured as follows: Section 2 provides an overview of the related work. Section 3 describes the modified shape context features and histogram bin configurations. In Section 4, we introduce our body part segmentation method. In Section 5, we present the results of the body part segmentation, unusual activity recognition and gait analysis tests. We also describe the motion capture process that was used to create the training database. Finally, in Section 6 conclusions are drawn and future work is presented.

2. Related Work

Model-based approaches, where prior knowledge of the shape of the human body in various poses is used for part identification, are often used for tracking the location of the limbs in videos [5, 24]. Mittal et al. [17] present an algorithm for body part segmentation from silhouette images. Their approach divides the body using negative minimum curvature with the shortest cut being selected when more than one cut is possible. Multiple camera views are used to assemble potential body parts in 3D, new candidate parts are added and if the likelihood of the new configuration increases these parts are retained. Their algorithm performed poorly when the body parts, particularly the arms, were held close to the body.

A number of papers concentrate on reconstructing a full 3D pose. Li et al. [15] apply a geometric transform in their method. A visual pattern is represented with certain invariance through this transform before learning is applied. This process is sensitive to noisy contours so the outline is skeletalised and then re-grown to produce a smooth contour. For body part segmentation, a shape space of body contours is created and the limbs are manually segmented. The images are then transformed and matched to images in the shape space and limb segments can then be also matched. Bourdev and Malik [31] utilised Support Vector Machine to find poselets in the images. Poselets are parts of the human body that are tightly clustered in 3D joint configurations and 2D appearances. Agarwal and Triggs [1] recover a set of 54 3D limb angles by regressing against the shape context features for the silhouette image. Once joint angles have been used to reconstruct the pose, the position of the limbs can be inferred. Mori and Malik [18] propose shape context features to match test silhouettes to a set of exemplars. The locations of the body joints are manually labeled on each exemplar and the 2D locations of the joints are copied to the best matching test silhouette. Joint locations are then used to reconstruct the 3D pose and locate the limbs. Ladický [9] et al. present a joint pixel-wise and part-wise model for body part segmentation using colour images. Zuffi et al. [10] introduced a flowing puppet model to segment body parts. Their method requires an initial body pose estimate before the puppet model is used. The puppet model utilises optical flow to fit to the movement of the subject.

The introduction of depth information can make the segmentation of the foreground object much easier compared to a traditional camera. Knoop et al. [32] used stereo cameras and time of flight cameras for tracking human body movements. Grest et al. [33] presented a markerless motion capture system using stereo imaging. Both these approaches lacked real time performance. Recently the Microsoft Kinect sensor has been used in body part segmentation applications [30]. Their implementation runs in real time. The shortcoming with Kinect, and active stereo imaging that uses infrared projectors in general, is that it does not work well in direct sunlight. This reduces the sensors usability in some applications.

In our work, we present a body part segmentation system that works with noisy images and videos. Compared to the existing solutions, we treat the body part classification as a problem in itself rather than solving the more difficult problem of 3D pose reconstruction. In our solution, the outline of a human silhouette is a sequence of pixels and our task is to segment it into a series of body parts. The segmentation results of the sequence is then used to classify a particular human pose.

3. Shape Context Features

In our system, we use silhouette images in order to segment body parts. These were chosen for a number of reasons. Firstly, a binary silhouette image contains no texture or colour information. This gives a degree of robustness to different subjects and environments. Additionally silhouettes are relatively easy to produce, given a method of background subtraction. Silhouettes are, however, prone to artifacts such as shadows caused by poor background subtraction performance. These can adversely affect global shape features such as moments. Therefore, we require a feature that can capture the global object shape whilst being robust to local artifacts in the silhouette shape. We use shape context features, which are a generic set of features used to describe the shape of a silhouette outline.

Shape context features were originally used in shape matching and defining the aligning transformation between two objects by Belognie et al. [3]. Mori and Malik [18] used shape context features to match silhouettes in order to recostruct the 3D pose of the person from single uncalibrated 2D image. Agarwal and Triggs [1] used shape context features also for recovering the 3D pose of the person from a 2D image. Their algorithm recovers the pose by direct nonlinear regression against shape descriptor vectors. Poppe and Poel [21] compared shape context features to Fourier Descriptors and Hu Moments. They recovered human pose from large database of human silhouettes under different view angles, body dimension and noise. The Hu Moments were considerably inferior to the compared methods under all deformations. Körtgen et al. [13] introduced enhanced 3D shape descriptors that work in a similar manner than the 2D shape context features. They showed that the descriptors give good recognition rates even with added noise.

Shape context features robustly encode local shape information in the form of edge histograms. The locality of shape context features overcomes the problem of artifacts or noise in one area of the silhouette affecting the entire descriptor. This property of locality also aids in the recognition of partially occluded shapes. Shape context features are invariant to translation since all measurements are taken with respect to points of the object. Scale invariance can be achieved by normalizing all distances by the mean distance between points. Rotation invariance can be achieved by rotating the coordinate system at each point in a way that the x-axis is aligned with the tangent vector.

To create shape context features a silhouette edge is sampled at regular intervals. For each sampled edge point the distance and direction to all other edge points that fall under

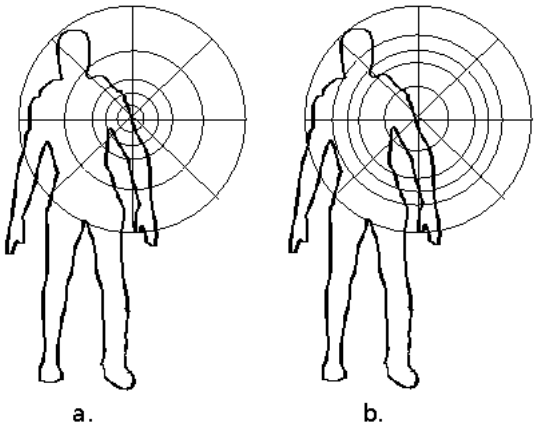

Fig. 1: (a) Typical shape context feature extraction with log radial bins (b) our proposed use of weighted bins.

the maximum distance are calculated. The distances and directions are stored in a log-polar histogram. Figure 1(a) illustrates the original log-polar histogram that has 8 angular and 5 radial bins.

This bin configuration was used to increase the locality of the feature, however, we found that in the presence of noisy silhouettes this locality places too much emphasis on local pixels, thus causing recognition errors. We, therefore, propose a weighting on the radial bin size as shown in the Figure 1(b). Instead of the original logarithmic spacing we use a weighted spacing of the radial bins to place emphasis on pixels in the middle distance as opposed to very close pixels. The spacing is linear, with the spacing of the middle third of the bins being half that of the inner and outer bins.

The weighted spacing of the bins is defined as

$$d = \begin{cases} \frac{R}{2N} & \text{if } \frac{R}{3} < r < \frac{2R}{3} \\ \frac{R}{N} & \text{else} \end{cases} \tag{1}$$

where N is the number of radial shells with equal spacing, R is the radius of the descriptor, r is the distance from the center of the descriptor and d is the width of the radial bins. When the bin configuration is weighted this way, the emphasis is placed on the pixels in the middle distance [2].

4. Body Part Segmentation

In pattern recognition the aim is to classify data based on either *a priori* knowledge or on statistical information extracted from the data. In our case, the data to be classified are measurements, defining points in an appropriate multidimensional feature space. A point in this feature space is defined as $X = \{x_1, x_2, ..., x_D\}$, where D is the dimension of the feature space. Statistical pattern recognition is the process of creating a probability density function $p(X|c_k)$ in the feature space of X corresponding to a particular class or label c_k that has an *a priori* probability of $P(c_k)$. The task of classification based on statistical modelling is that given a data vector X and a set of N classes $(c_1, c_2, ..., c_k, ..., c_N)$ we wish to assign X to one particular class c_k. The selected class is given by

$$c_k^* = \underset{c_k}{\text{argmax}} P(c_k|\mathbf{X}). \tag{2}$$

One key question is how do we create the probability density function $p(X|c_k)$ for the classes we wish to recognise. In the fields of image and video analysis this is further complicated by the large dimension of the feature space and the often high level semantic concepts encoded in the classes $(c_1, c_2, ..., c_k, ..., c_N)$. This process of going from a large dimensional low-level feature space to high-level semantic concepts is often described as the semantic gap. There are currently three main approaches (not necessarily disjoint) to the problem of bridging this gap. The first, often described as a bottom-up approach is to produce more sophisticated and specific features in order to reduce the dimension of the eventual feature space used for classification. In this case a relatively simple classifier may then be used to extract high level semantic information from these features. The second approach, known as top-down, involves the use of prior knowledge of the problem domain in order to design a solution. This depends on the prior knowledge being accurate for the feature space one is attempting to model. The final approach is to use more complex classifiers trained on a set of labeled generic low-level features in order to classify high level semantic concepts.

The first two approaches mentioned above suffer from the fact that they become tailored to a particular instance of the recognition problem to be solved and so have difficulty generalising to other instances of the problem. In the last approach, classifying low-level features, a problem is created by the lack of labeled training data, particularly in the area of image and video analysis.

As mentioned above, one approach is create a model that corresponds to the probability density function $p(X|c_k)$. The parameters of this model are determined by using a set of labeled training data that represents the distribution for the particular class c_k. How well this model approximates the actual distribution is governed, in part, by how closely the distribution of the training set matches the actual distribution. Clearly the more examples in the training set the closer the distribution should be to the real distribution of the features. However, we must also take into account the variance of the distribution and attempt to match this in the training set. This is particularly an issue in images and video material where there are many sources of variability such as, illumination, viewing angle or colour. Indeed much of the effort in image and video analysis has focused on the development of features and descriptors that are robust to these variations.

Our goal is to demonstrate that using artificially generated training data we can produce a robust segmentation of the human outline taken from binary silhouette images. This segmentation should be robust to changes in a viewing angle, occlusions of the body and noise in the body silhouette outline. To demonstrate this we first train models on synthetic data with labelled body part segments viewed in a range of different body configurations and from a range of different viewing angles. Using a validation set also comprised of synthetic data sequences we are able to set all the hyper-parameters of the model. In the testing phase real video sequences recorded from various sources are used to test the performance of the model trained. We also show that the same model, with no changes to any parameters, can give good results on data collected from different sources in different environments with different subjects. While we focus on body part segmentation in human silhouettes in this work, this technique could be applied to any silhouette that can be decomposed into meaningful sub-shapes, such as vehicles or animals.

Synthetic data has been used effectively in other fields where the amount of training data is limited. Varga and Bunke [26] proposed a perturbation model for creating synthetic training data for handwritten text recognition. While Heisele and Blanz [11] used morphable

models to create more training data from a face database that had only a small number of images per person.

These labeled features derived from the training data are used to train a GMM for each body part. The GMMs form the states of a HMM. We can consider each silhouette outline as a sequence of shapes corresponding to body parts. Using a HMM we can constrain the shape recognition by taking into account the transitions between shapes using Viterbi decoding [27].

Instead of determining these transitions from a priori domain knowledge, that may be flawed or not related directly to the data set we are dealing with, we determine the transitions from the training data itself using the labeled synthetic data. This gives us a distribution of the transition probabilities between body parts within the data. We train a four state fully connected HMM using the labeled synthetic data. One state in the model for each body part, in our case Head, Body, Arms and Legs.

In order to cope with uncertainty in the recognition process, we used a ratio of the likelihood between the most likely class and the second most likely class for each pixel on the silhouette outline as in,

$$\mathcal{L} = log(L_M) - log(L_S) \tag{3}$$

where L_M is the likelihood of the most likely class and L_S is the likelihood of the second most likely class for a particular pixel. Any pixel where the likelihood falls below a threshold is discarded as the uncertainty on the pixel is considered too high. The threshold value was determined by validation and it was set to 20. In later experiments we will show how a global measure of uncertainty over all pixels can be used as a way to detect unusual poses.

5. Experimental Results

In this section we describe experiments in order to demonstrate that models trained purely on synthetically generated data can be used to accurately and robustly segment a human silhouette into body parts. We also discuss the method we used to obtain the training data.

5.1 Data Collection

The synthetic data used for our experiments was created through motion capture. The hardware used for creating the motion capture data was three Sony DFW-710 FireWire cameras and a Unibrain 800 Mbps FireWire board. We used a PC that had a 2.4GHz Intel Core 2 Duo processor and 2GB of RAM. The video stream from each camera was captured at 15 gray scale frames per second at 800x600 resolution.

Five subjects were used to create the synthetic training set: 3 male and 2 female. Each subject was recorded performing two action sequences SEQ1 and SEQ2. In the first action sequence (SEQ1) the subject walks around the room and sits on a chair. In the second sequence (SEQ2) the subject walks around the room, falls over and then gets up and then walks around the room again. Each of these sequences was approximately 20 seconds long, although this time did vary between individuals.

A 3D human figure was animated using the motion capture data. Each sequence was rendered from 12 different viewing angles, six from straight forward and six from the side with different offsets. In order to label the body parts we coloured the initial skin used in the models with separate colours for each part. Figure 2 shows examples of the labeled figure rendered from twelve different angles.

One of the main advantages of generating synthetic training data is that the skin used in the 3D rendering of the action sequences can be labeled according to body parts, as shown

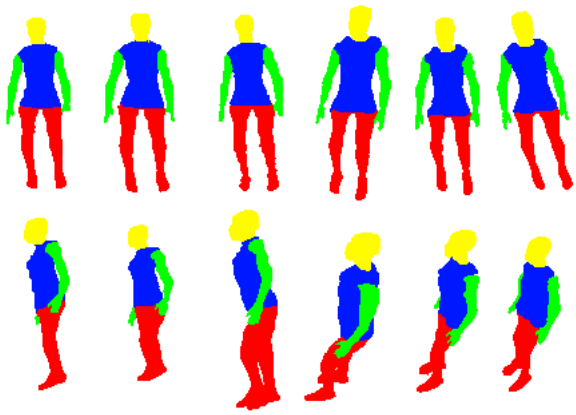

Fig. 2: A figure with labeled body parts shown rendered from twelve different viewing angles.

in Figure 2. This labeling needs only to be performed manually once, then this pre-labeled skin can be applied to different skeletons. These skeletons, with the attached skin, can then be animated using the motion capture data. So instead of, tediously, manually labeling each frame, we can label a single frame and this labeling is then carried through for all subsequent frames in the training set as the figure adopts different poses.

We rendered approximately 50000 frames of labeled video from 12 different viewing angles. Compared to other publicly available databases this represents a significant amount of labeled data. For example, 8000 frames of the University of South Florida's gait database [20] has been hand labeled, this consists of people walking viewed from one angle. However, this data has been primarily used to test previously trained models and not to train the models.

For the experiments we divided the data into training and validation sets, the test sets being provided by real sequences from a variety of sources. The training set is composed of three subjects: two male and one female, each performing SEQ1 and SEQ2. These sequences were rendered from six different viewing angles with different offsets (three viewing angles from the front and three viewing angles from the side). In total we had 36 training sequences. In addition to this data collected by our group we included 8 sequences from the CMU motion capture database `http://mocap.cs.cmu.edu`. This data featured subjects performing various actions with the arms and upper body. The motion capture data was downloaded from the CMU site and then rendered by us using the same labeled skin as the previous sequences.

Similarly a validation set was created using two subjects: one male and one female. The sequences were rendered as in the training set. Having a labeled validation set allowed us to se the hyper-parameters of the model using a measure of performance. In our case there are two critical hyper-parameters: the number of Gaussians in the GMMs (we assume here each GMM has the same number of mixtures) and the likelihood ratio threshold for discarding pixels. We used a simple pixel based recognition rate and the model configuration with the highest recognition rate was selected. The number of Gaussians in each state was set at 75 and the likelihood ratio threshold was set to 20.

The performance of this model was then tested in three different sets of experiments. The first is a test of the body part segmentation performance of the model under different

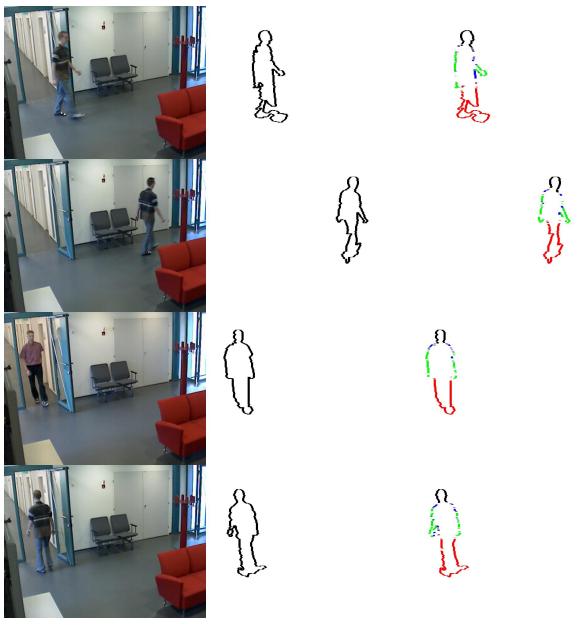

Fig. 3: Subjects in different poses from various angles, demonstrating the robustness of the body part segmentation to viewing angle.

viewing angles, with noisy or bad silhouettes and also under occlusions. The second set of experiments aims to classify unusual poses through a confidence measure taken over the whole model. The third set of experiments uses sequences from the USF gait recognition database to test the model's performance in detecting both leg and arm motion from the low resolution pre-segmented silhouettes provided in the database.

For these experiments the test set was composed of 13 sequences of people walking, standing, sitting and falling over, from different viewing angles and in different environments. These sequences were all recorded with single un-calibrated camera. Eleven of the sequences were recorded at the University of Oulu, while the remaining two sequences were taken from the CMU MoCap database.

One important point to stress before we move on to look at the actual experiments, is that a single model was used for all experiments in this paper. Once the parameters and hyper-parameters of the model have been set on the training set and validation set respectively, they are not altered.

5.2 Body Part Segmentation Test

In this set of experiments we tested the performance of the proposed body part recognition method. The tests were run on sequences containing people performing normal human activities. We tested the performance of the proposed method on noisy, poorly segmented silhouettes, from a variety of different viewing angles. The test sequences also contained partially occluded silhouettes. The aim is to classify each pixel on the silhouette outline into one of four body part classes, Head, Body, Arms and Legs. The results of performing body part segmentation on silhouettes in various poses and from different viewing angles is shown in Figure 3. The images feature three sections, the first is the original frame from the video sequence, the second is the silhouette outline obtained from using simple background subtraction and the third is the silhouette outline segmented into body parts. These results show that our method is robust to any changes in the horizontal viewing angle.

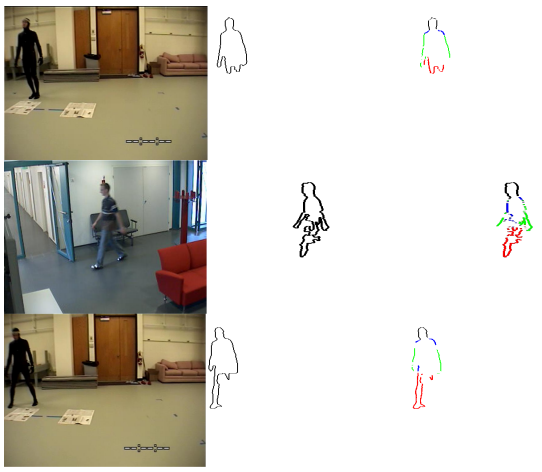

Fig. 4: A selection of noisy, badly segmented silhouettes, showing our methods robustness to noisy silhouette outlines.

One significant problem in dealing with silhouette data is that it can be very noisy. This noise is caused by shadows, changes in lighting conditions or the background matching the foreground we wish to segment. The noise causes changes in the shape of the silhouette. This causes changes in features that are computed globally over the whole silhouette. In Figure 4 we show the performance of our model in noisy conditions. In each of the images the legs have been poorly segmented by the background subtraction method. It can be seen that this does not affect the overall recognition of the remaining body parts.

Another source of error in silhouette segmentation is occlusions. This is particularly true in the case of surveillance and monitoring applications where the subject might enter or leave a room. The Figure 5 shows frames where the subject leaves or enters the scene. From these frames we can see that the limbs are segmented correctly even though there is occlusion present. Also the number of rejected pixels increases when the uncertainty of the model increases.

A more detailed description of the experiments can be found in [34].

5.3 Unusual Activity Recognition Test

The detection of unusual events in video has recently become a very active area of investigation in computer vision [19, 28, 29]. In this section we present the experiments to demonstrate the use of our model in detecting unusual poses. Here we define an unusual pose as any pose other than sitting, walking and standing. Postulating a surveillance application, the most usual behaviour is actually composed of sitting, walking and standing and any other pose should therefore be considered unusual. To this end the training data was primarily composed of subjects in these three poses. Of course, the training set could be composed of any set of poses one wishes to recognise.

It has been noted in previous work [28] that the class of unusual pose or behaviour is poorly defined. So any learning based system faces the problem of recognising a class for which the probability density function in the features is very complex. Indeed we noted in our own data collection that when subjects were asked to "fall", they all adopted a different pose in the act of falling. Therefore we do not explicitly model the unusual poses, we instead define an unusual pose as a pose that the model has not seen before.

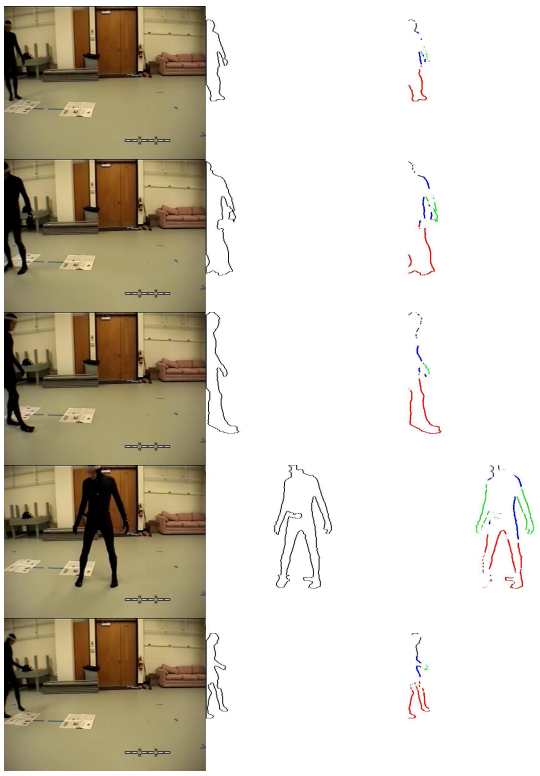

Fig. 5: A subject under different types of occlusion. These results show the performance of our algorithm under various degrees of occlusion.

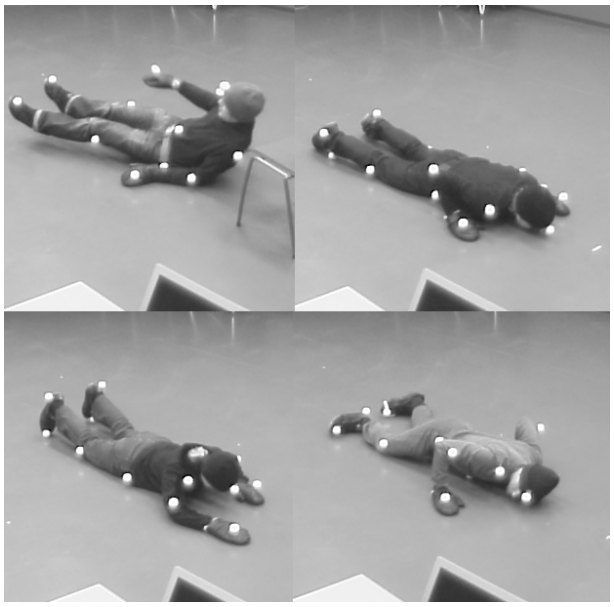

Fig. 6: Images of four subjects having fallen, demonstrating that the matching of a particular pose to the act of falling is not possible.

In order to recognise these unseen poses we define a global confidence measure based on the likelihood ratio in Equation 3 given by

$$\mathcal{G} = \frac{1}{T}\sum_{t=1}^{T}\mathcal{L}_t. \tag{4}$$

When we plot this global confidence measure, $\mathcal{G}$, for a video sequence, the difference in the confidence measures between a normal and an unusual poses can clearly be seen when the subject falls and lies on the floor. When the subject sits down the confidence does drop but it maintains a higher level than seen in the falling sequences. The method is described in more detail in [35]. The measure was tested on all 13 test sequences, 8 of which contained unusual poses, such as falling, tripping or crouching. All these poses were detected by the system.

5.4 Gait Analysis Test

In our final set of experiments we look at the problem of gait analysis. This has become a very active area of research within the computer vision community in recent years with the emphasis being on biometric applications [12,14]. This is based on the fact that each persons way of walking, or gait, is unique. In these experiments we are using the database developed for the HumanID Gait Challenge Problem by the University of South Florida [23]. The baseline algorithm for this challenge uses silhouettes produced by background subtraction. These silhouettes are low resolution, 128x88 pixels, and very noisy. Previous work has shown that results can be improved through working to directly reduce the noise in this silhouette data [16]. We randomly selected 20 sequences of silhouettes from the USF database, each sequence was between 200 to 250 frames long. In these experiments we demonstrate the ability of our model to segment these silhouettes into our four main body parts of Head, Body, Legs and Arms.

As all subjects were recorded from a similar viewing angle and performing the same action, walking, the segmentation of the head and legs was relatively straightforward despite the poor quality of the silhouettes. This can be seen in Figures 7 and 8. The segmentation of the arms from the body, however, is difficult due to the low resolution and noisy nature of this data. The motion of the arms, does however, form an important feature in gait recognition [4]. We can see in Figure 7 that in mid-stride when the legs are closest together the arms appear to be at the side of the body, so more body pixels are recognised and also when the legs are far apart as shown in Figure 8 there are more arm pixels recognised. This arm motion with the arms swinging with the same pace with the legs corresponds to the normal walking motion.

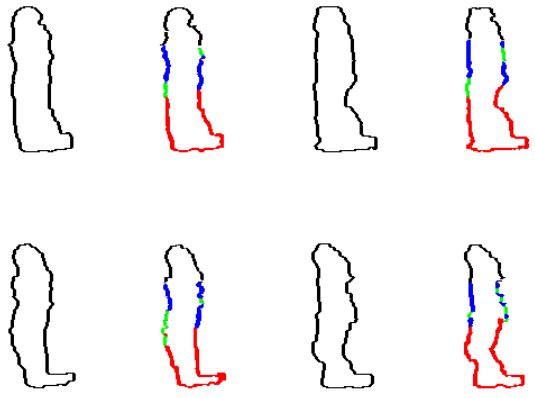

Fig. 7: Segmented silhouettes from the USF gait recognition database. These frames show the stride when the legs are close together and very few arm pixels are being recognised.

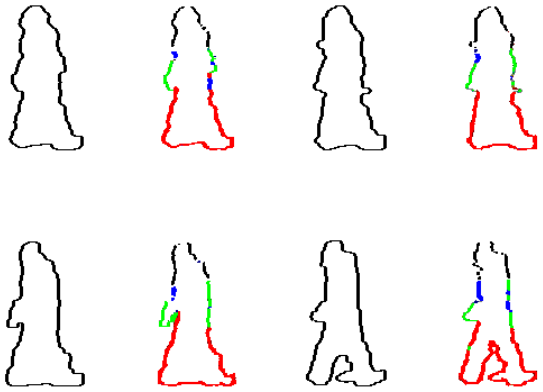

Fig. 8: Segmented silhouettes from the USF gait recognition database. These frames show the stride when the legs are apart and more arm pixels are being recognised.

It is, however, hard to quantify this just from the inspection of the segmented silhouettes. In Figures 9 and 10 the first two waveforms are the normalised distance between the outermost leg pixels (in blue) and the ratio of arm pixels to body pixels (in red). We then performed a Fast Fourier Transform (FFT) on these waveforms in order to measure the frequency components of the arm and leg motion. The magnitude spectrum of each waveform is displayed in the two bottom plots. These clearly show that in both cases the fundamental frequency of arm and leg motion is the same. Of the 20 sequences we tested, this correspondence of arm and leg motion was found in 17 of these sequences. This result shows that our model was able to segment arm from body pixels despite the poor quality of the data used.

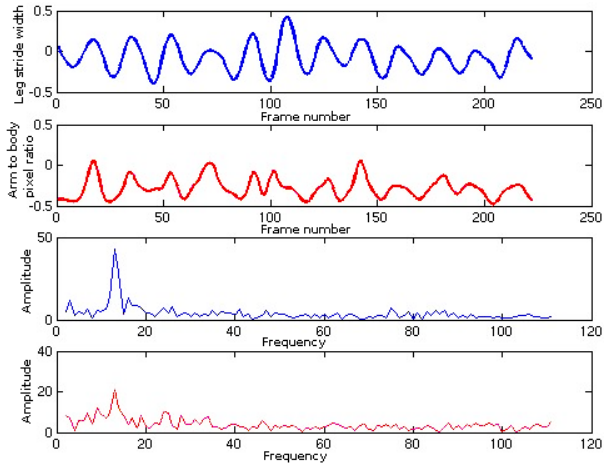

Fig. 9: The first two plots illustrate the leg stride width and the arm to body pixel ratio for the USF sequence 03774GOBR respectively. The next two plots illustrate the frequency spectrum for both waveforms. Both frequency spectrums show the same fundamental frequency.

Interestingly, Figure 10, as well as showing the same fundamental frequency, also displays a large component at half the fundamental frequency. This effect was noted in 7 of the 20 sequences. We are unsure as to the cause of this but it could be due to half the swing cycle

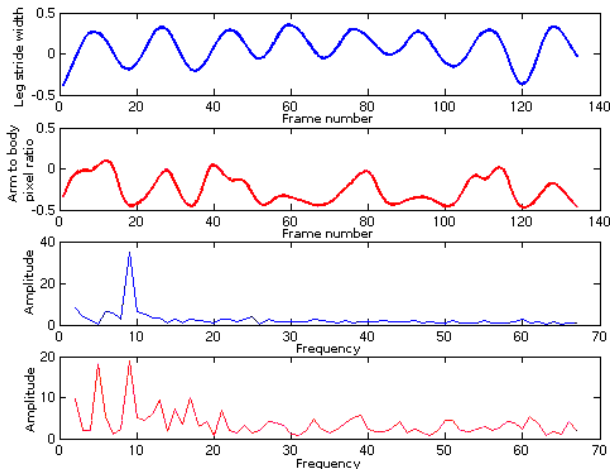

Fig. 10: The first two plots illustrate the leg stride width and the arm to body pixel ratio for the USF sequence 03788GOAR respectively. The next two plots illustrate the frequency spectrum for both waveforms. Both frequency spectrums show the same fundamental frequency.

of the arms not being seen, possibly because of body shape or the speed of the walker. This may even prove to be a useful feature for future experiments into person recognition using gait analysis.

6. Conclusions

We have presented here a system for shape classification that uses Hidden Markov Models and shape context features for statistical modeling of the human body shape. The system presented is used in the problems of body part segmentation, unusual activity detection in home environments and gait analysis. In the unusual activity detection application we used a global measure of confidence for our model based on a likelihood ratio. This was shown to be able to distinguish walking, sitting and standing from other unusual activities, such as falling or crouching. In the gait analysis experiment we showed that our segmentation model was able to recognise arms swinging in the USF gait recognition database.

The statistical models used require a lot of training data to avoid overfitting. We have presented our method for creating vast amounts of labeled synthetic training data to overcome this problem. This synthetic data can be produced from a variety of viewing angles with any desired labeling. We demonstrate that models produced using this synthetic data can be used to segment real body silhouettes and are robust to changes in viewing angle, noise and occlusions.

The weakness of the presented system is the requirement of a background subtraction method. The background of a typical room can be difficult to model due to moving objects (doors, chairs, etc.) and changing lighting conditions. To this date, no method has completely solved this problem.

In future we are planning to add occlusions to the training data. It is easy to add objects to a 3D scene and then render the database again. We believe that the performance increases if the training data includes occluded data also. The body part segmentation system proposed could be used to recognise activities. We are going to use the tracked body parts to train models to recognise the actions.

References

[1] A. Agarwal and B. Triggs. Recovering 3D Human Pose from Monocular Images. IEEE Transactions on Pattern Analysis and Machine Ingelligence, 28(7):1052-1062 (2006)

[2] M. Barnard and J. Heikkilä. On Bin Configuration of Shape Context Descriptors in Human Silhouette Classification. International Conference on Advanced Concepts for Intelligent Vision Systems (2008)

[3] S. Belognie, J. Malik and J. Puzicha. Shape Matching and Object Recognition Using Shape Contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(4):509-522 (2002)

[4] N. Boulgouris and Z. Chi. Human Gait Recognition Based on Matching of Body Components. Pattern Recognition, 40(6):1763-1770 (2007)

[5] C. Bregler and J. Malik. Tracking People With Twists and Exponential Maps. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, page 8 (1998)

[6] R. Collins, R. Gross, and J. Shi. Silhouette-based Human Identification from Body Shape and Gait. In proceedings of IEEE Conference on Face and Gesture Recognition (2002)

[7] S Dockstader, M. Berg and A Tekalp. Stochastic Kinematic Modeling and Feature Extraction for Gait Analysis. IEEE Transactions on Image Processing, 12(8):962-976 (2003)

[8] R. Hartley and A. Zisserman. Multiple View Geometry in Computer Vision. Cambridge University Press (2003)

[9] L. Ladický, P. Torr and A. Zisserman. Human Pose Estimation using a Joint Pixel-wise and Part-wise Formulation. In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2013)

[10] S. Zuffi, J. Romero, C. Schmid and M. J. Black. Estimating Human Pose with Flowing Puppets. In proceedings of the IEEE International Conference on Computer Vision (2013)

[11] B. Heisele and V. Blanz. Morphabple Models for Training a Component-based Face Recognition System. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2004)

[12] A. Kale, A Sundaresan, A. Rajagopalan, N. Cuntoor, A. Roy-Chowdhury, V. Kruger and R. Chellappa. Identification of Humans Using Gait. IEEE Transactions on Image Processing (2004)

[13] M. Körtgen, G. Park, M. Novotni and R. Klein. 3D Shape Matching With 3D Shape Contexts. In The 7th Central European Seminar on Computer Graphics (2003)

[14] L. Lee and W. Grimson. Gait Analysis for Recognition and Classification. In proceedings of the IEEE Conference on Face and Gesture Recognition (2002)

[15] J. Li, S. Zhou and R. Chellappa. Appearance Modeling Under Geometric Context. In Proceedings of the Tenth IEEE Conference on Computer Vision, pages 1252-1259 (2005)

[16] Z. Liu and S. Sarkar. Effect of Silhouette Quality on Hard Problems in Gait Recognition. IEEE transactions on Systems, Man and Cybernetics, Part B, 35(2):170-183 (2005)

[17] A. Mittal, L. Zhao and L.S. Davis. Human Body Pose Estimation Using Silhouette Shape Analysis. In proceedings of the IEEE Conference on Advanced Video and Signal Based Surveilance, pages 263-270 (2003)

[18] G. Mori and J. Malik. Recovering 3D Human Body Congfigurations Using Shape Contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(7):1052-1062 (2006)

[19] H. Nait-Charif and S.J. McKenna. Activity Summarisation and Fall Detection in a Supportive Home Environment. In the Proceedings of the 17th International Conference on Pattern Recognition (2004)

[20] P. Phillips, S. Sarkar, I. Robledo, P. Grother and K. Bowyer. The Gait Identification Challenge Problem: Data Sets and Baseline Algorithm. In the Proceedings of the IEEE International Conference on Pattern Recognition, pages 385-388 (2002)

[21] R. Poppe and M. Poel. Comparison of Silhouette Shape Descriptors for Example-based Human Pose Recovery. In Proceedings of the 7th International Conference on Automatic Face and Gestue Recognition (2006)

[22] L. Rabiner and B Juang. Fundamentals of Speech Recognition. PTR Prentice Hall (1993)

[23] S. Sarkar, P.J. Phillips, Z. Liu, I.R. Vega, P. Grother and K.W. Bowyer. The HumanID Gait Challenge Problem: Data Sets, Performance and Analysis. In IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(2):162-177 (2005)

[24] H. Sidenbladh, M.J. Black and L. Sigal. Implicit Probabilistic Models of Human Motion for Synthesis and Tracking. In Proceedings of the European Conference on Computer Vision (2002)
[25] A. Sundaresan, A. Roy-Chodhury and R. Chellappa. A Hidden Markov Model Based Framework for Recognition of Humans From Gait Sequences. In Proceedings of the International Coference on Image Processing (2003)
[26] T. Varga and H. Bunke. Generation of Synthetic Training Data for an HMM-based Handwriting Recognition System. In Proceedings of the International Conference on Document Analysis and Recognition (2003)
[27] A.J. Viterbi. Error Bounds for Convolutional Codes and an Asymptotically Optimal Decoding Algorithm. IEEE Transactions Information Theory (1967)
[28] D. Zhang, D. Gatica-Perez, S. Bengio and I McCowan. Semi-supervised adapted HMM:s for Unusual Event Detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Volume 1, pages 611-618 (2005)
[29] H. Zhong, J. Shi and M. Visontai. Detecting Unusual Activity in Video. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2004)
[30] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman and A. Blake. Real-Time Human Pose Recognition in Parts from Single Depth Images. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2011)
[31] L. Bourdev and J. Malik. Poselets: Body Part Detectors Trained Using 3D Human Pose Annotations. In Proceedings of the IEEE 12th International Conference on Computer Vision (2009)
[32] S. Knoop, S. Vacek and R. Dillmann. Sensor Fusion For 3D Human Body Tracking with an Articulated 3D Body Model. In Proceedings of the IEEE International Conference on Robotics and Automation (2006)
[33] D. Grest, J. Woetzel and R. Koch. Nonlinear Body Pose Estimation from Depth Images. In Proceedings of 27th Annual meeting of the German Association for Pattern Recognition (2005)
[34] M. Barnard, M. Matilainen and J. Heikkilä. Body Part Segmentation of Noisy Human Silhouette Images. Proceedings of the IEEE International Conference on Multimedia and Expo, pp. 1189 – 1192 (2008)
[35] M. Matilainen, M. Barnard and O. Silvén. Unusual Activity Recognition in Noisy Environments. Proceedings of the Advanced Concepts for Intelligent Vision Systems, pp. 389 – 399 (2009)