Image based human body rendering via regression & MRF energy minimization

A thesis submitted for the degree of Doctoral of philosophy

by

Xinfeng Li

School of information system, computer science and mathematics

Brunel University

May 2011

Abstract

A machine learning method for synthesising human images is explored to create new images without relying on 3D modelling. Machine learning allows the creation of new images through prediction from existing data based on the use of training images. In the present study, image synthesis is performed at two levels: contour and pixel. A class of learning-based methods is formulated to create object contours from the training image for the synthetic image that allow pixel synthesis within the contours in the second level. The methods rely on applying robust object descriptions, dynamic learning models after appropriate motion segmentation, and machine learning-based frameworks.

Image-based human image synthesis using machine learning is a research focus that has recently gained considerable attention in the field of computer graphics. It makes use of techniques from image/motion analysis in computer vision. The problem lies in the estimation of methods for image-based object configuration (i.e. segmentation, contour outline). Using the results of these analysis methods as bases, the research adopts the machine learning approach, in which human images are synthesised by executing the synthesis of contour and pixels through the learning from training image.

Firstly, thesis shows how an accurate silhouette is distilled using developed background subtraction for accuracy and efficiency. The traditional vector machine approach is used to avoid ambiguities within the regression process. Images can be represented as a class of accurate and efficient vectors for single images as well as sequences. Secondly, the framework is explored using a unique view of machine learning methods, i.e., support vector regression (SVR), to obtain the convergence result of vectors for contour allocation. The changing relationship between the synthetic image and the training image is expressed as a vector and represented in functions. Finally, a pixel synthesis is performed based on belief propagation.

This thesis proposes a novel image-based rendering method for colour image synthesis using SVR and belief propagation for generalisation to enable the prediction of contour and colour information from input colour images. The methods rely on using appropriately defined and robust input colour images, optimising the input contour images within a sparse SVR framework. Firstly, the thesis shows how contour can effectively and efficiently be predicted from small numbers of input contour images. In addition, the thesis exploits the sparse properties of SVR efficiency, and makes use of SVR to estimate regression function. The image-based rendering method employed in this study enables contour synthesis for the prediction of small numbers of input source images. This procedure avoids the use of complex models and

geometry information. Secondly, the method used for human body contour colouring is extended to define eight differently connected pixels, and construct a link distance field via the belief propagation method. The link distance, which acts as the message in propagation, is transformed by improving the low-envelope method in fast distance transform. Finally, the methodology is tested by considering human facial and human body clothing information. The accuracy of the test results for the human body model confirms the efficiency of the proposed method.

Acknowledgement

First, I would like to express my gratitude to my supervisor Prof Feng Dong; this thesis will never exist without his constructive point of view of research and objective guidance. From his broad knowledge and deep understanding of computer graphics, I have received invaluable supervision in the past four years of my research. The influence applies far beyond research and learning experience for me is not only in research but also in life, his broad and deep professional perspective, rigorous style of life and work have given me the criteria and direction for my future work and life.

Second, I am grateful to my second supervisor Prof Marios Angelidas and the research staff: Mrs Ela Heaney for their enthusiastic help and generous advice.

Finally, I would like to send respect to my parents; they encouraged me to solve problems and taught me truth in life, also, I appreciate so much the time with my girlfriend, she gives my research life fulfilled exciting.

Contents

Abstract 2
Acknowledgement 5
Contents 6
1. Introduction 8
1.1 Brief background 9
1.2 Research contributions 13
1.3 Brief approach and thesis outline 14
2. Related work 18
2.1 Human pose analysis 20
2.1.1 Model based method of human pose analysis 21
2.1.2 Learning based method of human pose analysis 23
2.1.3 Regression methods for 3D human body pose from monocular Images 25
2.2 Definition of Markov Random Field (MRF) 27
2.2.1 Definition of neighbouring domain 28
2.2.2 MAP-MRF 29
2.3. Image based rendering 30
2.3.1 Method of IBR 31
2.3.2 IBR between stable rate adjacent frames 34
2.3.3 New IBR Approach Based on View Synthesis 35
2.4 Summary 37
3. Contour synthesis 38
3.1 Machine learning and SVM 39
3.1.1 Linear SVM 39
3.1.2 Non-linear SVM 43
3.2 ε-SVR 44
3.2.1 ε-band with hyperplane 45
3.2.2 Margin of SVR 46
3.2.3 Optimisation of the hyperplane 47
3.3 ε-SVR used in contour synthesis 49
3.3.1 SVR Parameters 53
3.3.2 SVR Parameters estimation and optimisation 55
3.4 Implementation 59
3.4.1 Data collection 60
3.4.1 SVR kernel functions 61
3.4.2 Parameters of SVR using different kernel function 67
3.4.3 Performance of parameters with optimal solution 68
3.5 Testing results 70
3.5.1 The testing results with different kernel functions 71
3.6 Summary 76
4. Pixel synthesis 78
4.1 Contour colouring via BP 81

4.2 Constructing the link distance field through an assigned BP program.....82
4.3 The initialisation of disparity..... 86
4.4 Message computed between two neighbours.....87
4.4.1 Message in BP.....87
4.4.2 Message computation.....92
4.5 Initialisation of eight distance..... 93
4.6 Results and performance analysis..... 94
4.6.1 Human poses without clothing information..... 98
4.6.2 Human poses with clothing information..... 102
4.6.3 Pixel synthesis for human body model from different viewing directions 105
4.7 Summary..... 107
5. Conclusion and future work.....109
5.1 Key contributions:..... 109
5.2 Possible future extensions:.....111
References..... 113

1. Introduction

Nowadays, computer graphics is a very popular field. One of the major sub-topics of computer graphics is human rendering. i.e., synthesis of virtual human images with appealing visual appearance.

A major feature of modern computer graphics is to create special visual effects by making use of existing data and images. One of the major characters of the current society is information big bang, which features exploitation of large data. For example, the amount of Internet images had risen by a factor of 56, from 5 exabytes in 2002 to 281 exabytes in 2009. While these images provide a rich resource and great potential for the current research of computer graphics rendering, the majority of them have not been utilized by the graphics community. The work on this thesis is a movement towards this direction.

This thesis brings the synthesis capability of image based rendering of computer graphics by using learning based method, and addresses the exploration of image-based rendering through machine learning. More specifically, it focuses on human image synthesis; the results are expressed in terms of synthesis of the object contours and pixel colours of synthetic human images. The advantage of the learning based method is to avoid dense sampling of rendered environment and the requirement of any geometry

information. The basic idea is to create new images through the learning from existing images which are largely available nowadays.

1.1 Brief background

The traditional approach to generating virtual views is direct rendering of 3D models, which can be produced using modellers, digitising tools, or stereo techniques. Kang (1997) reviewed three forms that 3D models can assume: polygonal and bicubic parametric patches, constructive solid geometry, and space subdivision representations. To enhance realism, assumed forms can be combined by applying texture or environmental maps, shading algorithms (e.g., Gouraud 1971 or Phong 1975), bump maps (Blinn 1978), or any combination of these techniques on 3D model surfaces.

Volume rendering or visualisation of voxel-based data (as opposed to surface-based data referred to in most earlier studies) is also another type of 3D model rendering. Depending on the application, the Marching Cubes algorithm (Lorensen and Cline, 1987) may be used for volume rendering. Volume may also be rendered with voxels, in accordance with the assigned volume data using different opacities or colour classifications.

The above mentioned approaches are 3D model based, which rely on

modelling transformation, view transformation, culling, and hidden surface removal. Therefore, object or scene complexity is a factor in the cost involved in rendering, specifically, in representation; this cost depends on the number of facets or voxels considered. For more complicated scenes, a 3D graphics accelerator, in addition to expensive software, is usually required in applying fast rendering.

The image-based rendering technique is an emerging and alternative means of producing virtual views. In contrast to 3D model-based rendering, image-based rendering relies on a set of trained or original images. One of the differences between the two methods is illustrated by object and scene representation, in which 3D model-based rendering uses a set of appropriately constructed 3D models, whereas image-based rendering applies a collection of original or real images—a convenient method for obtaining real objects and scenes.

By contrast to the 3D model based approach, image-based rendering techniques use original or real images, or pixel reprojection onto target images to produce virtual views from source images. As a result, the cost of rendering is independent of conventional restrictions, unlike model-based rendering, which relies on factors that may lead to a large number of computations. In representing wide scenes, however, considerable significant memory will be

required for the input images. Moreover, the realism in resultant images relies on the quality of source images.

To adopt this learning-based approach, numerous existing techniques in computer vision were also applied in the current work. Computer vision has become an increasingly important branch of artificial intelligence. The work revolving around human movements can be traced back to as early as 1973, at which time psychologists such as Johansson (1976) proposed the theory of object recognition based on research on perceptions of human movement. Human perception is more sensitive to dynamic objects than static ones (Barclay et al. 1978). The theory of computer vision was then represented by Marr (1982), who contended that computer vision involves a study of the processes necessary to obtain the corresponding dynamic scenes from single or multiple 2D images. In this theory, the vision system is divided into three phases, namely, low-level, middle-level, and high-level vision. Low-level vision involves image processing that uses image filtering, image enhancement, and edge detection to extract information about image contours, colours, appearances, and other basic features. The main task involved in middle-level vision is restoring 3D scenes, including features such as depth, surface characteristics, and other relevant information. Given these basic image features, high-level vision aims to complete the 3D information of the object, including object position and orientation. After two decades, computer vision

technology has rapidly developed in both theory and practice. Human motion analysis in advanced human-computer interactions, security monitoring, and medical diagnostics, among other fields, are some of the applications of this technology. Many important explorations have also been conducted (Bregler and Malik 1998; Sidenbladh et al. 2002; Klein et al. 2002; Grauman et al. 2003).

In general, despite many years of research, the current IBR require either densely and structurally collected images, or certain level of geometric information of the objects（Shum et al 2004; Kang et al. 2003）. Technically, this imposes a serious restriction on the applicability of these techniques to a general collection of unstructured images where the image setting is unknown and geometry information is not available.

Given the image resources available nowadays, this research will focus on IBR via learning based approach. It was designed to experiment a fundamentally new view reconstruction approach by leveraging learning and prediction strategy employed in machine learning and computer vision.. Prediction using models acquired through data learning is a common practice in machine learning. The rich image resource provides the possibility to build a reliable model for view prediction with desired accuracy. By doing so, the

learning based method can avoid the involvement of large image collections, and also, no geometry information is required.

1.2 Research contributions

The contributions of this work are summarised as follows:

1) Object contour synthesis method

Contour synthesis is the process of predicting new images from given sample images. The ε -Support Vector Regression (SVR) for contour synthesis is applied from the image samples. To the best of the knowledge, ε -SVR is proposed here for the first time in contour synthesis research. The sparse properties of the SVR method are exploited, and support vectors are created to estimate the regression function. Finally, the possibility of ε - and hard ε -bands are listed, and the optimisation is defined in terms of a linear regression. This synthesis method is implemented for the images which are created by human rendering package. The experimental result shows that the method proposed has relatively good effect on object outline.

2) Colour synthesis using Belief Propagation (BP)

Colour synthesis takes place based on the contours that are synthesised in the first step. In colour synthesis, the standard energy function is involved in

the proposed stages. First, Belief Propagation (BP) is applied to determine the solution through the minimisation of the standard energy function. The general definition of the energy function includes two items: P is used to represent the matching probability for a single pixel and V indicates smooth continuity between adjacent points. For the proposed stages, the assigned BP is firstly introduced to the assigned points in a message propagation process. This is done by either examining the set of candidate disparities $\Delta(p)$ in order to achieve energy minimisation, or denoting the pixel difference in eight directions. The probability of the difference and pixel value are defined, and the thesis subsequently creates colours for the pixels. The ultimate implementation of contour colouring indicates fair synthesis results.

1.3 Brief approach and thesis outline

In this thesis, 3D human images are applied as source images. A learning-based method is applied to image synthesis using regression and optimisation as basic tools to generate new images. The SVR method is used for contour synthesis to obtain poses from adjacent image frames. With optimisation and the proper initial parameter settings, a synthesised contour is obtained from a small number of training samples through the regression function and maximised propagation. Methods are data-driven, do not use any explicit human body models, and do not involve large collections of images. In

this thesis, contour synthesis results are used as inputs in colour synthesis. A standard energy function is introduced to the approach after considering synthesis problems in the MRF inference. The key technique here is to determine the values of the MRF nodes. The solution is obtained through minimisation of the energy function. To distill the corresponding points, assigned BP method is applied to define the link distance field for the construction of distances between the source image and the synthesised image. The colour information is transformed through an improved low-envelope method. The results have good adaptability, and are time-efficient.

The flow chart of the overall approach is shown below:

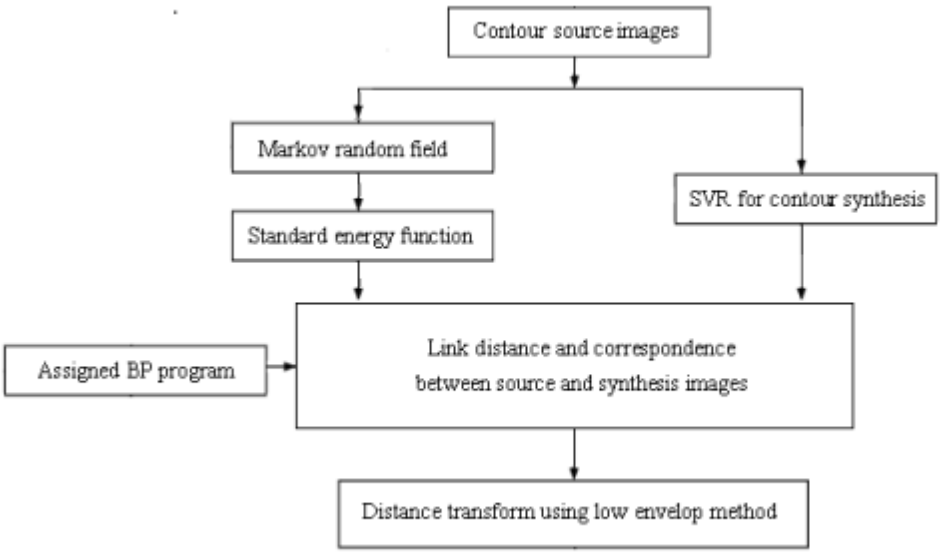


Figure 1.1 Contour and colour synthesis

The thesis is organised as follows:

Chapter 1, Introduction. Thesis presents the research background, contributions to the field, and an outline of the approach. The various problems involved are briefly discussed in the research.

Chapter 2, Related work. This chapter has brief reviews of the techniques employed in human pose analysis and image-based rendering, particularly in terms layers of image-based human images. The problems attributed to possible technical difficulties and future trends in the field are also summarized.

Chapter 3, Contour synthesis. The processing of contour prediction is extended to synthesis using a regression method based on optimisation. Regression is used to build a function that avoids the generative models found in most literature. The parameter-estimation method, and present results and conclusions are described in this chapter.

Chapter 4, Pixel synthesis. The link distance is constructed in contour colouring via BP. In this chapter, the link distance of two neighbouring pixels is constructed by using assigned BPs. The link distance is then incorporated into

the fast distance transform using a low-envelope method. The new and improved method is time-efficient, and features seamless colour merging when interferences occur.

Chapter 5, Conclusion. It sums up the thesis and presents prospects for future work.

2. Related work

Human motion analysis is an important concern in the pattern recognition and machine learning domains, as well as in virtual reality research. The analysis of human motion involves applying human tracking and human motion capture, and obtaining body motion parameters for human pose estimation. This research concerns the application of image based human motion analysis to the image-based rendering (IBR). The method proposed refers to the novel concept of prediction methodology using synthesis. The method is applied by using the parameters obtained from image-based human motion analysis. Therefore, human motion analysis, especially images from human motion, supply the source images for our research. Overall, considerable technical and expert support is needed to accomplish video-based rendering/image-based rendering (VBR/IBR). In the present study, IBR techniques are used to implement the parameterisation of the contour and colour synthesis of source images. In this part, the main task is managing integral factors, with emphasis on the various developmental aspects related to. This section discusses human motion analysis techniques and the future prospects of relative fields.

The analysis of human poses has the following fundamental elements:

(1) Contour of the human body. The most common approach is to use the background reduction algorithm to obtain foreground information (Gavrila, 1996; Deuscher, 2000; Sminchisescu, 2001). Another contour extraction method is traditional film production via Chroma-keying method, which uses a fixed colour background to separate the foreground colour from the background colour.

(2) Using colour pixel statistics as basis for rendering. For example, the Pfinder system is used to determine the attribution of pixels in accordance with the colour and location of the blob statistical information. The advantage of the statistical segmentation method is that the images obtained are more robust to noise and environmental changes. However, targeting the complex model is difficult.

(3) Segmentation using motion information. An example of such usage is the study of Krahnstoever (2003), in which optical flow analysis was used in accordance with the different values of each pixel. The movement into different regions was then divided. The difference in adjacent image frames can also be used in extraction, such as that performed by Leung (1995). This approach requires background stability.

(4) Using a source image template. The human body is expressed as one or

more templates, and matching is processed in the current frame through the template matching method. This method assumes that the surface of body colour (gray scale) remains unchanged, and human motion is slow and continuous (Ju, 1996).

(5) Feature-based tracking and matching point. In this method, the interest points or pixels in a frame are first identified, and then process matching in frames is conducted. Based on this method, the motion capture system requires the labelling of prescribed feature points (Vicon, 2003; Silaghi, 1998).

Undoubtedly, the analysis of human motions from such elements brought about innovative perspectives into the field. Further research in this field is also required, particularly on those related to upcoming technologies and the motion image representation industry (i.e., object representation, IBR, etc).

2.1 Human pose analysis

The analysis of human pose in coherent human motion from monocular images is a challenging issue for the following reasons:

1) The first has to do with variations in body types (stemming from different human facial features, clothing, etc.).
2) Second is the changeability and unevenness of body movement styles.

3) Finally, the 3D nature of the human body makes the analysis of real life images a challenging task (Guo and Qian, 2008).

Appropriate human pose analysis is important not only for the routine explanation of human poses in videos, but also in designing 3D computer and video games. It is also essential for articulated and emotional human body animation (Kakadiaris and Metaxas, 2000). The aim of this section is to provide a brief, yet comprehensive review of learning-based methods from the two common basic approaches to human pose analysis: the model- and the learning-based approaches.

2.2.1 Model based method of human pose analysis

The model-based method assumes a clearly defined body model prior to analysis, and estimates body poses through one of the following methods. a) Appearance-based methods for people detection, b) part-based methods for people detection (face or limbs), and c) 3D pose estimation in images using either matching-based methods, or inverse kinematics approach (that is, first determining body joints, then estimating body poses). Human pose tracking in the model-based method is by either gradient-based (track short successions of movements) or sampling-based techniques (Lee and Cohen, a2006).

As stated by Fan et al. (2002), the process of analysing motion begins with the performer who wears markers near each of his joints. The placement of markers helps identify motion and allows for a clear perception of the positions or angles between the predestined markers of an individual's joints.

Many scholars [1] have devoted their efforts to this particular domain. Stephan and Sommer (2006) proposed a model for learning the articulated motion of the human arm, an aspect related to the act of the grabbing hand. The goal is to generate plausible trajectories of human body joints that mimic the acts of human movement through the specific depiction of deformation information. Trajectories are mapped to a constraint space and initiate the configuration of the human body and all task-specific constraints, such as those of avoiding an obstacle. This model is for principal component analysis and the dynamic cell structure network. In Jürgen's (2006) Bayesian framework, he made a vivid and elaborative study of the applicable poses that are usually captured in the current frame[2]. The prediction made by Jürgen (2006) was narrated through a dynamic model and updated through the next frame.

1 Dagstuhl Seminar Proceedings 06241 Human Motion - Understanding, Modeling, Capture and Animation. 13th Workshop "Theoretical Foundations of Computer Vision" http://drops.dagstuhl.de/opus/volltexte/2006/721 .

2 Pose distribution is learned from training samples using a Parzen-Rosenblatt estimator with a weighted Euclidean distance measure.

2.1.2 Learning based method of human pose analysis

This method avoids the assumed definite body model. It takes advantage of the hypothesis that the sets of representative human body poses are less than those of potentially kinetic poses. This can be achieved by learning (estimating) a model that diametrically retrieves guesses from evident image quantities. Consequently, the learning method occurs in one of two ways. The first is the storage and search for comparable training examples. Alternatively, a training data base can be filtered into a single compressed model using Bayesian regression analysis (Agarwal and Triggs, a2004).

R'omer (2001) presented a system for 3D hand pose retrieval from 2D colour images. It utilises a nonlinear learning context (supervised), called system specialised (being for a particular part of the body) mapping architecture (design). In other words, it is a system that charts image characteristics to possible 3D hand poses using a nonlinear supervised framework. This system has two basic components: a collection of specialised drawings (mappings) and a separate response (feedback) matching function. Forward mapping is approximated from training data, such as joint make-up and visual characteristics; joint angel data are obtained from CyberGlove®[3] in training, and a computer graphics model produces the visual characteristics that deliver a hand labelled as 22 joint angles.

3 CyberGlove® (http://www.mindflux.com.au/products/vti/cyberglove.html)

Cohen and Hongxia (2003) presented a method for assuming 3D body posture. They used a 3D illustrative body (hull) built from a collection of silhouettes to present a 3D shape based on appearance and independent view. Categorizing and recognizing body posture using the obtained 3D shape description are conducted with the help of a vector machine.

Another learning-based method for human pose estimation is the directional (top down and bottom up) generative identification model for 3D human pose estimation from monocular images. The basic concept of this method lies in the identification model, which is adjusted using samples from the producing (generative) model, which in turn, is enhanced to generate implications from the recognition model. The advantages of this context are the production of consistent 3D initialisation and retrieval of 3D human poses (Sminchisesco et al., 2006).

Ahmed (2008) used the entire systematic learning proceedings for the purpose of decomposable generative models from high dimensionality of the configuration space. These are also concerned with the subject and the explicitly decomposing acts that the internal body configuration supplies. The relationships are established in terms of time-invariant parameters.

Furthermore, interaction environments become further cluttered because of rapid lighting changes in non-static backgrounds. Image-based human pose representation and analysis in a specific conditioning environment needs to be supported by robust acts that involve real-time, accurate images and footage-captured motion. Grest and Koch (2005) discussed stereo algorithms exclusively[4]. Stereo algorithms can provide robust data with respect to variable lighting conditions and non-static backgrounds.

2.1.3 Regression methods for 3D human body pose from monocular Images

1) Elimination of cluttered images in high dimensional space

While working in high dimensional spaces, an example-based approach often raises problems in a densely covered space because creating enough example incorporation is difficult to accomplish. This is acceptable based on the estimation of human pose, which is recovered in many joint freedom degrees from a signal image using the regression approach (Shakhnarovich and Grauman, 2003). The merit of this approach is the advanced ability to recover observed images from pose parameters; however, the disadvantage is the requirement for discriminative and robust representation of image input (Leventon and Howe, 1999). Consequently, it uses the regression approach

[4] The presented approach estimates arm movement by using optimised search with 5–6 fps from up to 1000 correspondences

and extends its application to conform to cluttered background picture presence. Therefore, this method mainly focuses on the removal of relatives, which are nonsensitive. In this case, the approach needs to rely on the classical single valued regressor as the upper body gestures are relatively multimodal; there are fewer problems compared to the case of full body representation. However, if necessary, this regression method of multimodal multi-value can be employed (Blake and Isard, 1998), with attention mainly focused on the representation and outline of the images by eliminating clutter in the background of the target object

2) Regression method

This method summarises test set performance for various regression techniques such as the least squares regression and support vector machine. This is derived from the Kernelised and linear basis version for various possible subset poses and models of the full body.

Moreover, occasionally, pose regression from a silhouette in a single image leads to errors, which actually occur in multi-solutions, thereby causing the regressor to choose the wrong solution, or output compromised solutions that contradict one another. The most proficient way to reduce these errors is to incorporate stronger features, such as internal body edges, within the silhouettes (Agarwal and Triggs, c2008). However, the problem persists

because the important internal body edges are not visible and the irrelevant textures of clothing edges have to be considered (Thayananthan and Stenger, 2003). Through observed experimental poses, the single image method can reduce ambiguities that low-level vision problems are usually relying on significantly.

Furthermore, most studies focus on the retrieval of 3D human poses from silhouettes. The method that has been applied in relevance vector regression in the learning-based context was described by Agarwal and Triggs (b2006). The retrieval of 3D human poses is implemented by direct nonlinear regression against silhouettes vectors. Moreover, these vectors are extracted from shape descriptors in automatic computer behaviour from silhouettes. The advantage of this method is it makes having a precise body model unnecessary and requires no prior labelling of body parts in the image.

2.2 Definition of Markov Random Field (MRF)

The research on the Markov random field (MRF) began in the 1960s, and substantially progress after the equivalence between Markov random fields and Gibbs fields. The equivalence can be used to describe MRF using the Gibbs distribution. The problem, which is described by MRF, can be solved by the establishment of the effective link of distribution and energy function. In

recent years, researchers have achieved explicit advancements in solving early vision problems by applying MRF. MRF modelling is used in edge detection, moving object segmentation, image texture analysis, colour image segmentation, and other areas of computer vision. It provides a robust framework for early vision problems, such as optical flow, image representation, restoration, and stereo matching (Pedro et al., 2006). In practical research, the inference of MRF is also used in graph cuts, and belief propagation has gained accurate results.

2.2.1 Definition of neighbouring domain

The domain for MRF nodes is very important to the link between distribution and energy function. The definition of a neighbouring domain is first introduced (Berthod et al., 1996). For example, it assumes that η_s is the neighbouring domain of node s in set S. Therefore,

1) node s belongs to domain η_s;

2) domain η_s is in set S;

3) if s and r belong to set S, if $r \in \eta_s$, then $s \in \eta_r$.

Hence, for any node in set S, it can obtain

$\eta = \{\eta_s | \forall s \in S\}$,

and the neighbouring domain can be described in order as

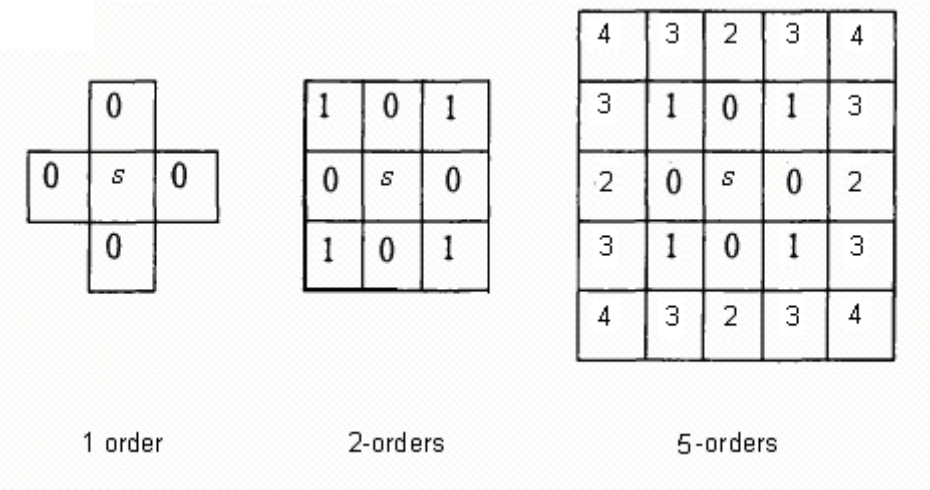


Figure 2.1. Neighbouring domain definition

In the thesis, it denotes the neighbouring pixels in eight different values using a 2-order domain, discussed in succeeding sections.

2.2.2 MAP-MRF

In accordance with Hammersley-Clifford and the links between distribution and energy function, Meier (1997) proposed a method for transforming the MRF problem into a maximum a posteriori (MAP) algorithm using

$$\hat{f} = \arg\max_{f \in \Omega} P(f|d)$$

$$P(f|d) = \frac{P(d|f)P(f)}{P(d)} \qquad (2.1)$$

Hence, the MAP problem can be transformed into minimum energy function as

$$\hat{f} = \arg\min_{f} U(f|d) = \arg\min_{f}(U(d|f) + U(f)) \qquad (2.2)$$

where $U(d|f)$ is the conditional energy function, and $U(f)$ is the priori energy function. For most computer vision problems, the process of MAP-MRF can be expressed as:

1) choosing the appropriate MRF to appoint specific matters for f;
2) energy function $U(f|d)$ is derived according to the known image data;
3) the solution for MAP is obtained through the minimisation of the energy function.

2.3 Image based rendering

A number of methods are currently available for IBR in the graphic field; image based rendering is a highly common term used in various situations that require understanding of the basis of graphics and other related terms (Fitzgibbon, 1998). Several algorithms are available for image- and video-based rendering; however, these require thorough understanding and implementation of IBR techniques (Genc, 1999). In particular, the understanding and implementation of image rendering techniques and methods in an appropriate manner (Faugeras, 1995) is an important requirement in the graphic field.

This thesis is related to the application of a method for providing unequal allocation of all IBR acts， and also thesis addresses all these terms to acquire a better perception of and more comprehensive follow up for the development of rendering. Functionality pertains to describing the process of recording all types of moments, and preferably obtaining the provisions required for translating information on the moment onto a digital model after vectorisation. Furthermore, there is a provision for implementation in the fields of entertainment and sports, as well as perspectives on medical applications that benefit mankind. Aggarwal and Cai (1999) stated that in this particular genre, a standard requirement for recording actions is necessary. Given that the movements and all matters related to actions are connected, problems can be converted to graphical representations.

All this information is then synchronised for animation in the form of 3D digital character models. The depiction or the graphical representation of the face, fingers, and expressions is transformed into digital coding and decoding. This is the act that has been termed IBR.

2.3.1 Method of IBR

IBR technology uses two-dimensional image information in expressing

rendering to achieve better realism and real-time images. It is irrelevant to 3D complexities (i.e., geometric information, human poses, etc.), and more effective as it does not require understanding of the 3D object structure.

Numerous techniques are available for rendering images, each of which has certain limitations and assumptions. Entire techniques and methods produce new and unique output based on different calculations and objectives (Mcmillan and Bishop, 1995; Levoy and Hanrahan, 1996). However, a few of them are briefly described as having different rendering requirements and approaches.

1) Panoramic view as basis

The basic idea is to obtain a panoramic view of one direction, and then this view is projected onto the inner surface or the inner surface of the sphere. Finally, the corresponding image is obtained according to the direction of the view, in which, the panoramic view is defined (from a fixed point of view) as the view of the image at 180° from the vertical direction and 360° from the horizontal direction.

2) Morphing as basis

Morphing technology considers feature information at different viewpoints, searching for correspondence instead of requiring geometric information. The correspondence transformation and projections are done in

accordance with the position of the point, as well as the feature of the sample. Moreover, the correspondence is summarised as grid, feature points, and field morphing.

3) Image depth information as basis

The basic idea is to use given point depth values and partial reconstruction of 3D scene geometry. Based on this 3D information, the visible point for direct projection is transformed, or the correspondence pixel between adjacent frames is created. The view interpolation algorithm (Eric, 1993) is the natural transition IBR algorithm between two images using the given depth of an image.

4) Light field information as basis

The light field reconstruction technique generates perspective by collecting large amounts of data rather than depth information or feature and geometric information (Gortler et al., 1996; Levoy, 1996).

Hundreds of researchers have been working on image rendering problems and majority of them use the previous background of image rendering as a basis for their research on new methodologies for different situations.

2.3.2 IBR between stable rate adjacent frames

This method features an efficient key frameless IBR technique. It is highly effective as it works on key frames that help create motion in images and videos. This approach is widely applied in intermediate imaging as an input to exploit the correspondence in neighbouring frames. The points in an input image are effectively rendered with the help of a ray-casting method (Chen, 2003), which is frequently used in rendering images to generate a nice and smooth to intermediate image for further processing. This method also uses an offset buffer to keep updates on the positions of obtained pixels from an input image, while every frame is generated in three steps: folding the input image into the frame, filling up the holes of an input image, and selectively rendering a group of old pixels. The number of these pixels in the last step allows for dynamic settings because the workload on every frame is balanced and proved effective in producing the desired output. The temporary key frames of an image need to be rendered frequently as input images always required updating in every frame at short intervals. These steps decrease the workload from each frame and provide balance to each frame, thereby providing more stable frames and a reasonable output image (Kaufman, 2000). The input image used in this approach is wrapped using the 3D rendering algorithm for obtaining expressive results. This technique is applied on the number of images for checking the output; in addition, it has produced substantial and acceptable results in specific applications in the combination of

the technique with a voxel-based terrain rendering system.

2.3.3 New IBR Approach Based on View Synthesis

Real world IBR has been one of the major research directions in computer graphics, in which image prediction has become advanced, with developments in terms of animation and motion analysis capabilities, and many other social requirements. In recent years, synthesising the number of human objects has become a famous prediction trend in the field of IBR.

Parke (1972) worked tirelessly on human face animation. When two main approaches were identified in facial animation (Parke, 1982), the technique based on IBR technologies for human face synthesis became a well-known approach in computer graphics. Currently, in terms of the motion and geometric information involved in the entire act, using the machine learning method for image-based human motion analysis has become a trend. However, major IBR techniques are restricted by the huge amount of requirements for geometric information and source image collection. In addition, in saving substantial efforts in creating 3D models, the image-based approach has the potential to achieve high realism in real images/videos (Buehler et al., 2001; Seitz et al., 1996; Shum et al., 2006; de Aguiar et al., 2008; Joshi et al., 2006).

A general definition of the IBR scheme is the synthesis new images, with results obtained from different viewpoints. Particularly, in this approach, an environment is displayed similar to the way it is seen in a sample camera controlled by user. Different experiments are applied to different types of images to generate effective results. This approach serves as a turning point in the image rendering field. It is applicable on numerous different images and specially designed for all types of objects (Ying 2006). It is reliable, conserves time, and always produces perfect and effective results, with environment conditions having tremendous effects on the approach. Therefore, the proposed technique in this thesis is a kind of image-based human rendering method, which targets human body synthesis. It uses the inference in the MRF model for rendering, and novel views are rendered even under a scarcity of source images. It is cost effective, time saving, and produces the best output depending on input type and smoothness.

2.4 Summary

Most recently, image-based rendering has evolved into an approach that merges computer vision and computer graphics. This is evident from the literature reviewed in this chapter. It makes use of the notion of segmentation and motion estimation, which is from the early vision problem, while the ideas related to rendering is heavily referred to in computer graphics

Currently, the challenges to most of the image based rendering techniques are that a certain level of geometry information needs to be involved otherwise a large image collection is required. And also, some of the current techniques become impractical due to high computational requirements. Therefore, by investigating the learning based IBR as proposed in this thesis, it is expected to render novel views without considering geometry information and a large collection of images and therefore overcome the technical challenge.

3. Contour synthesis

Support vector machine (SVM) is applied as a direct estimation function in the field of classification (pattern recognition). When extended to estimate a real function, application of support vector machine becomes a regression problem. Regression analysis is the most commonly used statistical method. It is used in dealing with various issues, including prediction, control, and optimisation. Through the introduction of a new loss function, application of support vector machine learning methods can be achieved with a strong robust regression. Regression estimation is incapable of retaining all the advantages of SVM. A learning and example-based method is proposed for the first time by using the SVR method to predict object contour in research.

This chapter describes a learning based method for synthesizing 3D human body pose from learning single images. The method developed here aims to use the existing methods in this domain (example based methods) to explicitly store a set of training examples whose 3D poses are known, and through learning the contour information from one single image by using sparse nonlinear regression to distill a large training database into a single compact model that has good generalization to unseen examples. The regression framework optionally makes use of kernel functions to measure similarity between image pairs and implicitly encode locality. This allows the

method to retain the advantage of example based methods. Despite the fact that full human pose recovery is very ill-conditioned and nonlinear, the method obtains enough information for computing reasonably accurate pose information via regression.

The method has chosen to base the system on taking image contour as input, which outline contour has been used as descriptor. To learn the contour of human body pose, the thesis takes advantage of the sparsification and generalization properties of Support Vector Machine (SVM) (Vapnik,1995) regression, allowing the contour outline to be obtained from a new image using only a fraction of the training database. This avoids the need to store extremely large databases and allows for very fast learning and synthesizing at run time.

3.1 Machine learning and SVM

Machine learning is an important aspect of modern data-based intelligence in the discipline of future data prediction from the observed data set. Linear and nonlinear SVM, classic methods in statistical algorithms for parameter estimation, are commonly used in statistical problems.

3.1.1 Linear SVM

Linear SVM is a linearly-separable case developed from the optimal

separating surface. The basic idea of SVM (Li and Huttenlocher,2008) is shown in figure 3.1.

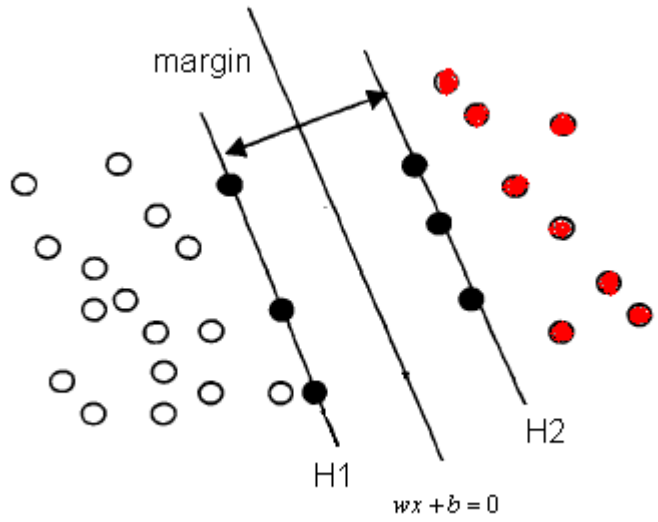


Figure 3.1. Optimisation hyperplane

where the hyperplane is described by $w\cdot x+b=0$; the points on H1 and H2 are the support vectors; and the distance is $2/\|w\|$, which is called the margin. The optimisation hyperplane allows for the correct identification and the computation of the largest margin. Linear SVM assumes that the training sample set $\{(x_i,y_i),i=1,2,...,l\}$ has two categories: $y_i=1$ and $y_i=-1$. The goal of training is to classify the two categories correctly. This means that, for any linearly separable sample set and hyperplane $w\cdot x+b=0$, linearly separable data have

$w\cdot x_i+b\geq 1, y_i=1$

and $w\cdot x_i+b\leq -1, y_i=-1$. (3.1)

These can be merged as follows:

$$y_i\left[(w\cdot x_i)+b\right]\geq 1, i=1,...,l \tag{3.2}$$

According to statistical learning theory, the hyperplane will be an optimised hyperplane if it separates the sample data with the largest margin. The optimised hyperplane is expressed as a quadratic problem:

$$\max \frac{2}{\|w\|}=\min \frac{1}{2}\|w\|^2 \tag{3.3}$$

s.t. $y_i\left[(w\cdot x_i)+b\right]\geq 1, i=1,...,l$

If the linear sample set cannot be separated, the optimisation of hyperplane can be expressed by introducing slack variable ζ as follows:

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\zeta_i \tag{3.4}$$

s.t. $y_i\left[(w\cdot x_i)+b\right]\geq 1, i=1,...,l$

where C is the penalty parameter. Introducing the Lagrange multipliers a, a^* for the linear quadratic problem, a dual optimisation problem is obtained as:

$$\max_{a,a^*}\min_{w,b,\xi} L=\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i - \sum_{i=1}^{l}a_i[y_i(w\cdot x+b)-1+\xi_i]-\sum_{i=1}^{l}a_i^*\xi_i \tag{3.5}$$

s.t. $a, a^*\geq 0$

Simplifying the equation by taking the derivatives provides the following:

$$\frac{\partial L}{\partial w} = 0 \rightarrow w - \sum_{i=1}^{1} a_i y_i x_i = 0$$
$$\frac{\partial L}{\partial b} = 0 \rightarrow \sum_{i=1}^{l} a_i y_i = 0 \qquad (3.6)$$
$$\frac{\partial L}{\partial \xi} = 0 \rightarrow C - a_i - a_i^* = 0$$

The dual optimisation problem can be simplified as follows:

$$\max \sum_{i=1}^{l} a_i - \frac{1}{2} \sum_{i,j=1}^{l} a_i a_j y_i y_j x_i x_j \qquad (3.7)$$

s.t. $0 \le a_i \le C$

Solving the optimisation problem, the optimal solution of a_i is

1) $a_i = 0$

2) $0 < a_i < C$

3) $a_i = C$

The optimal solutions of a_i in situations 2 and 3 correspond to x_i as the support vector.

3.1.2 Non-linear SVM

For the nonlinear SVM training set, a nonlinear function is used to transform the nonlinear training set from low-dimensional space into a linear function in a high-dimensional space. Then, the hyperplane for the classification function is conducted in a high-dimensional space. Hence, the hyperplane is described as $w \cdot \phi(x) + b = 0$

and the optimised hyperplane is described as

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l} \zeta_i \tag{3.8}$$

s.t. $y_i\left[(w \cdot \phi(x) + b\right] \geq 1, \quad \zeta_i \geq 0 \quad i = 1,...,l$

where $\phi(x) : R^n \to R^{n/h}$ transforms the input into a high-dimensional space.

After taking the derivation of w, b, ξ, the optimisation of dual problem becomes

$$\max \sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i,j=1}^{l} a_i a_j y_i y_j \phi(x_i)\phi(x_j) = \max \sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i,j=1}^{l} a_i a_j y_i y_j K(x_i, x_j) \tag{3.9}$$

s.t. $0 \leq a_i \leq C$

where $K(x_i, x_j) = \phi(x_i)\phi(x_j)$ is the kernel function that decides the structure of high-dimensional space. The kernel function is a symmetric function has two common main forms: the polynomial kernel function and the radial basis function.

3.2 ε-SVR

Support vector regression (SVR) is the extension of the SVM for estimation function. Its main feature is dimensional transformation. Through nonlinear transformation, the input can be transformed to a high-dimensional space in order to perform linear regression, thereby achieving nonlinear regression in original space. The nature of the algorithms guarantees that the regression model has good generalisation ability for solving the dimensional problem. For multi-dimensional spaces in SVM theory, ε-SVR is improved to achieve robust regression by introducing an insensitive loss function.

SVR has the following characteristics:

1) SVR has a solid theoretical foundation.

2) According to the quadratic problem, SVR has global optimal solutions.

3) Problems undergo nonlinear transformation to high-dimensional feature space. Using linear regression in a high-dimensional space, nonlinear regression is achieved in the original space. SVR has good generalisation ability and solves the dimensional problems in regression.

4) The SVR algorithm is entirely based on small training samples to construct the regression function.

The definition of ε is used to limit the support vectors of the training sample set by defining the other parameters needed to obtain a good

generalisation ability of the SVR method. The regressive model can be generalised for human motion. The function is established based on the relationship between the training data and the synthesised data.

The basic idea is to use the function performed to the training data to comply with the new input data to generate the new output information. In another words, the synthesis procedure is used on new data to generate new output by following the training data function.

3.2.1 ε-band with hyperplane

For the data set $(x_i, y_i)\Big|_{i=1}^{r}$, $x_i \in X^n$, $y_i \in Y$, the ε-band (Chen et al., 2004) is defined in figure 3.2 as follows:

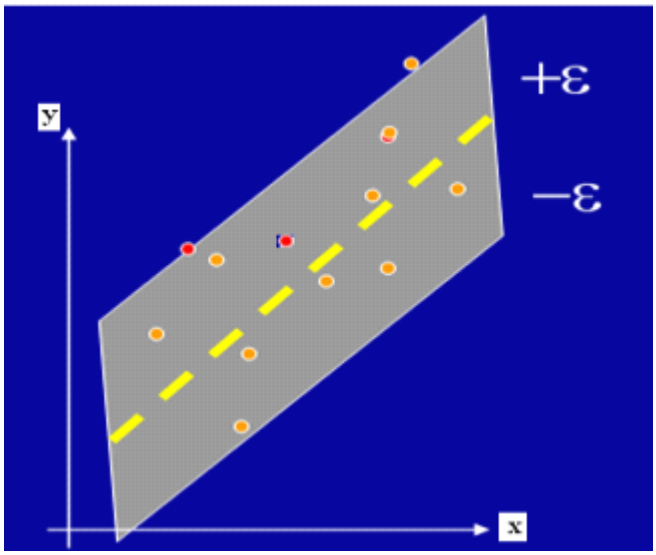


Figure 3.2. the definition of ε-band

In figure 3.2, if all the points are included in ε-band, the ε-band becomes a hard ε-band with a hyperplane, which can be expressed as follows:

$$y_i - ((w \cdot x_i) + b) \le \varepsilon, i = 1, 2, ..., l \tag{3.10}$$

This means that, for all the points in the set $\{(x_1, y_1), ..., (x_l, y_l)\} \in (X, Y)^l$ with the restraint in 3.10, the hard ε-band has hyperplane $y = (w \cdot x) + b$.

(Note that in this equation, x should be in terms of high-dimensional space.)

3.2.2 Margin of SVR

For the data set $(x_i, y_i)_{i=1}^r \; x_i \in X^n \; y_i \in Y$, first consider the hard ε-band. The distance d from any point (x_i, y_i) to the hyperplane $y = (w \cdot x) + b$ of ε-band is calculated as follows:

$$d_i = \frac{|(w \cdot x_i) + b - y_i|}{\sqrt{1 + \|w\|^2}} \tag{3.11}$$

In contrast, the hyperplane of hard ε-band is described as follows:

$$y_i - ((w \cdot x_i) + b) \le \varepsilon, i = 1, 2, ..., l \tag{3.12}$$

With the hard ε-band, the hyperplane includes all the training points.

Furthermore, with substitution of 3.12, the hyperplane has the following value:

$$d_i = \frac{|(w \cdot x_i) + b - y_i|}{\sqrt{1 + \|w\|^2}} \le \frac{\varepsilon}{\sqrt{1 + \|w\|^2}}, i = 1, 2, ..., l \tag{3.13}$$

where $\frac{\varepsilon}{\sqrt{1+\|w\|^2}}$ is the maximum margin for the hard ε-band.

Therefore, if the hard ε-band exists, the hyperplane with the maximum margin is the optimal regression hyperplane.

3.2.3 Optimisation of the hyperplane

The optimal hyperplane is conducted in accordance with the existence of the hard ε-band. The hyperplane of optimisation is described below:

1) With the hard ε-band, the optimisation of the hyperplane is expressed as follows:

$$\begin{aligned} &\min_{w,b} \frac{1}{2}\|w\|^2 \\ &(w\cdot xi)+b-yi\le\varepsilon, i=1,2,...,l \\ &yi-(w\cdot xi)-b\le\varepsilon, i=1,2,...,l \end{aligned} \tag{3.14}$$

The optimal solution of $\bar{w},\bar{b}$ will construct the optimal hyperplane $y=(\bar{w}\cdot x)+\bar{b}$

2) Without the hard ε-band, after introducing variable ξ,ξ^* as slack,

The optimisation equation becomes

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*)$$
$$(w \cdot x_i) + b - y_i \le \varepsilon + \xi_i, i = 1,2,...,l \qquad (3.15)$$
$$y - (w \cdot x_i) + b \le \varepsilon + \xi_i^*, i = 1,2,...,l$$
$$\xi_i, \xi_i^* \ge 0, i = 1,2,...,l$$

where slack variables ξ, ξ^* are used to measure the points staying outside the ε region, as shown in figure 3.3.

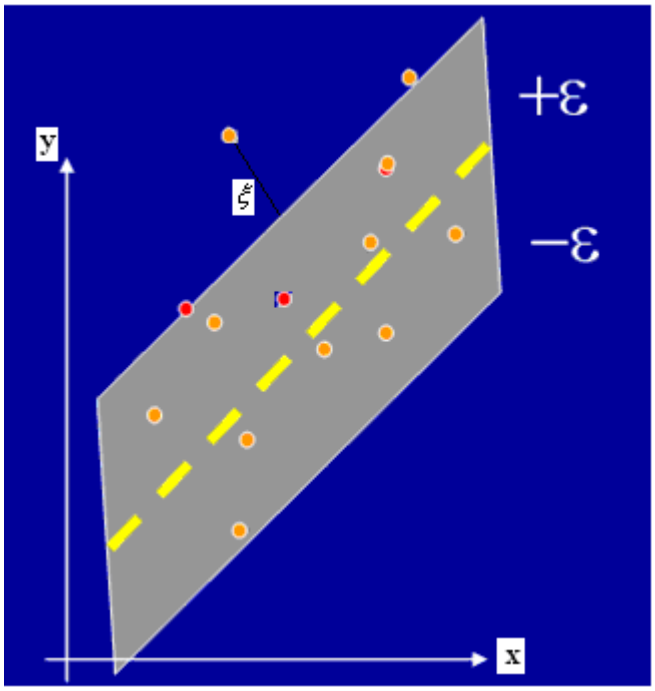


Figure 3.3. Tube with radius ε and slack variable ξ measures the points staying outside the tube

By introducing Lagrange multipliers, the solution of the equation can be obtained by optimisation, as shown below:

$$\min_{a(*)\in R^n} \varepsilon\sum_{i=1}^{l} a_i + a_i^* - \sum_{i=1}^{l}(a_i - a_i^*)y_i + \frac{1}{2}\sum_{i,j=1}^{l}(a_i - a_i^*)(a_j - a_j^*)(x_i \cdot x_j)$$

$$\sum_{i=1}^{l}(a_i - a_i^*) = 0 \qquad (3.16)$$
$$0 \le a_i, a_i^* \le C, i = 1,2,...l$$

The solution of equation 3.16 corresponds with the sample sets, which are the basis of the support vectors $w=\sum_{i=1}^{l}(\bar{a}_i^* - a_i)x_i$ for the regression function applied in this thesis.

3.3 ε-SVR used in contour synthesis

First, the SVR modelling is exploited in terms of mathematical functions by giving a sample of data $(x_i, y_i)_{i=1}^{l}$ with size l, where $x_i \in X^n$ is treated as the input of n-dimensional sample and $y_i \in Y$ is the sample output. Thus, the objective of nonlinear regression modelling is to find a function $y = f(x)$.

The generic SVR estimation function takes the form

$$f(x) = w \cdot \phi(x) + b \tag{3.17}$$

where ϕ denotes a nonlinear transformation into high-dimensional space.

The key for contour synthesis in the experiment is the regression function obtained from the training data. The basic idea is to apply the support vector regression algorithm as one of the contour synthesis methods. This determines the regression function through the training data. Based on the regression function, the input images with new variables will have a

corresponding output.

The ε -SVR method is used to search for the edge based on the ε -region method. For example, the training points are placed in the ε-region (refer to the cube with radius ε in figure 3.3) on the two-dimensional space. The selection of the parameter ε is very important because the contour points are allocated by using the size of the ε -region. They are then considered as the allocation on the contour by using the estimated regression function. The sample points from the edge based on manual segmentation of the sample images are selected. The key point for the regression estimation function is to search the support vectors in training samples forming the regression function for new output prediction. The traditional artificial neutral network (ANN) method based on the empirical risk minimisation (ERM) is used to determine the regression function:

$$\mathrm{R}_{\mathrm{e}mp} = \sum_{i=1}^{l} L(f(x_i) - y_i) \tag{3.18}$$

where $L(\cdot)$ is a cost function with $L(f(x_i) - y_i)$ denoting the deviation between the regression value and the concrete value.

For the regression function $y = f(x)$, the support vector regression method is the learning method based on the structure risk minimisation (SRM). The goal is to find the value of w, b. The value of x can be obtained by the

SRM as follows:

$$\min \ \frac{1}{2}\|w\|^2 + C \cdot \mathrm{R}_{emp}(f) \tag{3.19}$$

where C is a constant and $\mathrm{R}_{emp}(f)$ is the cost function representing empirical risk. The ε-insensitive loss function is the most widely used function to express $\mathrm{R}_{emp}(f)$ in 3.18:

$$\mathrm{R}_{emp}(f) = \begin{cases} 0, |y - f(x)| \le \varepsilon \\ |y - f(x)| - \varepsilon, |y - f(x)| > \varepsilon \end{cases} \tag{3.20}$$

Therefore, the SRM can be expressed by introducing the slack variables ξ_i and ξ_i^* as follows:

$$\min \ \xi_i \frac{1}{2}\|w\|^2 + C \cdot \sum_{i=1}^{l} (\xi_i + \xi_i^*)$$

$$\text{s.t} \begin{cases} y_i - wx_i - b \le \varepsilon + \xi \\ wx_i + b - y_i \le \varepsilon + \xi^* \\ \xi, \xi^* \ge 0 \end{cases} \qquad i = 1, 2...., l \tag{3.21}$$

where ξ_i and ξ_i^* are used to measure the errors outside the ε.

Therefore, based on the dual principle and the Lagrange multipliers $a_i^*, a_i (1, 2, ..., l)$, the SRM is equal to the maximisation of equation 3.19 with the constraints

$$\sum_{i=1}^{l}(a_i^* - a_i) = 0, a_i \in \left[0, \frac{C}{l}\right], a_i^* \in \left[0, \frac{C}{l}\right]$$

$$\max_{a(*)\in R^n} -\varepsilon \sum_{i=1}^{l} a_i + a_i^* + \sum_{i=1}^{l}(a_i - a_i^*)y_i - \frac{1}{2}\sum_{i,j=1}^{l}(a_i - a_i^*)(a_j - a_j^*)K\left(x_i \cdot x_j\right) \tag{3.22}$$

which is subject to

$$\sum_{i=1}^{l}\left(a_i - a_i^*\right) = 0, 0 \le a_i, a_i^* \le C, i = 1, 2, 3\cdots, l \tag{3.23}$$

The optimal solution of Lagrange multipliers $\bar{a}_i^*, a_i(1, 2, ..., l)$ in equation 3.22 is the basis for the construction of the support vector, which can be written in terms of data points as follows:

$$w = \sum_{i=1}^{l}(\bar{a}_i^* - a_i)\phi(x_i) \tag{3.24}$$

Therefore, the regression function can be written by expressing the support vector as

$$\begin{aligned} f\left(x\right) &= w \cdot x + b = \sum_{i=1}^{l}\beta_i \phi(x_i)\phi(x) + b = \sum_{i=1}^{l}\beta_i K(x_i, x) + b \\ \beta_i &= \bar{a}_i^* - \bar{a}_i, i = 1, 2, 3, \cdots, l \end{aligned} \tag{3.25}$$

where $K(xi, x)$ is the kernel function, which is used for high-dimensional transformation.

According to the optimal solution of Lagrange multipliers and the Karush-Kuhn-Tucker conditions, by substitution in equation 3.24, Bias b can be obtained as follows:

$$\bar{b} = y_j - \sum_{i=1}^{l} \beta_i K(xi, x_j) + \varepsilon \qquad \bar{a}_j \in (0, C) \tag{3.26}$$

or

$$\bar{b} = y_k - \sum_{i=1}^{l} \beta_i K(x_i, x_j) - \varepsilon \qquad \bar{a}_k^* \in (0, C) \tag{3.27}$$

In summary, to solve the problem of estimating the regression function using SVM, three parameters must be determined: non-sensitive value ε, regularisation parameter C and the kernel parameters (i.e., the width of RBF kernel parameters). The algorithm is used for regression estimation and controlling the generalisation ability of the SVM with the three parameters.

3.3.1 SVR Parameters

To construct the SVR model, choosing the width of the non-sensitive loss function is important. The parameter is the key for controlling the number of the support vectors. The point in the ε-region does not constitute a support vector; therefore, determining the data in the ε-region should be ignored in regression. When ε value increases, the number of support vectors decreases and the regression curve becomes flattened. However, increasing

ε may result in the prediction of performance degradation by the regression curve.

In contrast, parameter C is used as the penalty parameter to control the robustness of the regression model from the constraints of Lagrange multipliers:

$$\sum_{i=1}^{l}\left(a_i - a_i^*\right) = 0, 0 \le a_i, a_i^* \le C, i = 1, 2, 3\cdots, l \tag{3.28}$$

where a larger C assigns higher error penalties corresponding to a greater optimal solution of Lagrange multipliers. This will result in the lower generalisation of trained regression with minimised error.

Moreover, for the optimal answer to the question $a^{(*)} = (a_1, a_1^*, ..., a_l, a_l^*)^T$:

1) when a_i^* = a_i =0, the training point (x_i, y_i) should be on the edge or inside the ε-region area;

2) when $a_i \in (0, C), a_i^* = 0$ or $a_i = 0, a_i^* = C$, the training point (x_i, y_i) is definitely allocated on the edge of the ε-region; and

3) when $a_i = C, a_i^* = 0$ or $a_i = 0, a_i^* = C$ the training point (x_i, y_i) is allocated on the edge of the ε-region or outside the ε-region area.

According the description above, all the points inside the ε-region that correspond to a_i^* = a_i =0 are not support vectors, whereas, all the points (x_i, y_i) with $a_i^* \neq 0$ and $a_i \neq 0$ construct the support vector.

Therefore, for most of the properly selected values of ε, the penalty parameter C is a linear-based problem, whereas the SVM uses a nonlinear function to map the samples to a high-dimensional linear space to perform a regression function. The kernel function in the low-dimensional and nonlinear space can be used in the transformation from low-dimensional to high-dimensional.

In general, a kernel function is a symmetric function and has two common forms:

1) polynomial kernel function

$$K(x, x') = (xx'+1)^p \qquad p = 1, 2, ..., n \tag{3.29}$$

2) radial kernel function

$$K(x_i, x) = \exp(-\lambda \left\| x - x_i \right\|^2) \qquad (x_i \in X) \tag{3.30}$$

3.3.2 SVR Parameters estimation and optimisation

Training needs pre-determined learning parameters (C, ε, σ). Here, the cross-validation method (Wim and Cees, 2006) is used to determine these

parameters. The training set is divided into 10 equal portions. Each portion keeps a copy of the verification model performance, wherein 10% is used for testing and the remaining 90% is used for training models. The average performance of ten validation trials is used as the model performance parameter.

The parameters (C,ε,σ) use the steepest descent optimisation method to minimise SVR model:

1) Initial value of parameters (C,ε,σ) is set. The steps of parameter change after every iteration are denoted as $\Delta C,\Delta\varepsilon,\Delta\sigma$. Iteration denoted by the variable *iters* is equal to 1 and the number of unsuccessful iteration is set to 0.

2) The average performance under the current parameter values is calculated using the cross-validation method. The mean square error (MSE) under the current parameter value is $MSE(C,\varepsilon,\sigma)$.

3) For the parameter C, the values are changed to $(C-\Delta C,\varepsilon,\sigma)$ and $(C+\Delta C,\varepsilon,\sigma)$, corresponding to the change in parameter C. The expressions for the performance are $MSE(C-\Delta C,\varepsilon,\sigma)$ and $MSE(C+\Delta C,\varepsilon,\sigma)$. When the average $MSE(C,\varepsilon,\sigma)$ is less than $MSE(C-\Delta C,\varepsilon,\sigma)$ and

$MSE(C+\Delta C,\varepsilon,\sigma)$, the change in performance can be marked as $\Delta MSE_C = 0$. If $MSE(C-\Delta C,\varepsilon,\sigma)$ < $MSE(C+\Delta C,\varepsilon,\sigma)$, change in performance can be expressed as

$\Delta MSE_C = MSE(C-\Delta C,\varepsilon,\sigma) - MSE(C,\varepsilon,\sigma)$.

Otherwise, $\Delta MSE_C = MSE(C,\varepsilon,\sigma) - MSE(C+\Delta C,\varepsilon,\sigma)$.

4) For the parameter ε, the values are changed to $(C,\varepsilon-\Delta\varepsilon,\sigma)$ and $(C,\varepsilon+\Delta\varepsilon,\sigma)$, corresponding to the change in parameter ε. The performance values are $MSE(C,\varepsilon-\Delta\varepsilon,\sigma)$ and $MSE(C,\varepsilon+\Delta\varepsilon,\sigma)$. If the average $MSE(C,\varepsilon,\sigma)$ is less than $MSE(C,\varepsilon-\Delta\varepsilon,\sigma)$ and $MSE(C,\varepsilon+\Delta\varepsilon,\sigma)$, the change in performance can be marked as $\Delta MSE_\varepsilon = 0$. If $MSE(C,\varepsilon-\Delta\varepsilon,\sigma)$ < $MSE(C,\varepsilon+\Delta\varepsilon,\sigma)$, the change in performance can be expressed as $\Delta MSE_\varepsilon = MSE(C,\varepsilon-\Delta\varepsilon,\sigma) - MSE(C,\varepsilon,\sigma)$;

otherwise, $\Delta MSE_\varepsilon = MSE(C,\varepsilon,\sigma) - MSE(C,\varepsilon+\Delta\varepsilon,\sigma)$.

5) For the parameter σ, the parameter values are changed to $(C,\varepsilon,\sigma-\Delta\sigma)$ and $(C,\varepsilon,\sigma+\Delta\sigma)$, corresponding to the change in the parameter σ. The performance values are $MSE(C,\varepsilon,\sigma-\Delta\sigma)$ and $MSE(C,\varepsilon,\sigma+\Delta\sigma)$. If the average $MSE(C,\varepsilon,\sigma)$ is less than $MSE(C,\varepsilon,\sigma-\Delta\sigma)$ and $MSE(C,\varepsilon,\sigma+\Delta\sigma)$, the change in performance can be marked as $\Delta MSE_C = 0$. If $MSE(C,\varepsilon,\sigma-\Delta\sigma)$ < $MSE(C,\varepsilon,\sigma+\Delta\sigma)$, the

change in performance can be expressed as $\Delta MSE_{\sigma} = MSE(C, \varepsilon, \sigma - \Delta\sigma) - MSE(C, \varepsilon, \sigma)$;

otherwise, $\Delta MSE_{\sigma} = MSE(C, \varepsilon, \sigma) - MSE(C, \varepsilon, \sigma + \Delta\sigma)$.

6) The general performance under $\Delta MSE = \max(abs(\Delta MSE_C, \Delta MSE_{\varepsilon}, \Delta MSE_{\sigma}))$ is calculated. When $\Delta MSE = 0$, the non-change iteration number $iters = iters + 1$ and $\Delta C = \Delta C / 2, \Delta\mu = \Delta\mu / 2, \Delta\sigma = \Delta\sigma / 2$;

otherwise,

$$C = C + \Delta C \cdot \Delta MSE_C / \Delta MSE$$

$$\mu = \mu + \Delta\mu \cdot \Delta MSE_{\mu} / \Delta MSE$$

$$\sigma = \sigma + \Delta\sigma \cdot \Delta MSE_{\sigma} / \Delta MSE.$$

7) The equation $Iters = Iters + 1$ is set for next stage of iteration. If $Iters$ is greater than the maximisation of iteration, or if the non-change iteration number is greater than the maximisation, then there is an iteration escape. Otherwise, the iteration continues from Step 2.

In summary, the optimal values of parameters (C, ε, σ) can be obtained by following this iterative process. To post these parameters in all training data, the regression performance needs to be optimised for the synthesis.

3.4 Implementation

For the experiment, a data set composed of 40 human movements as human example data is rendered with several different human poses Then, the data is imported for pre-setting into the application Curious Labs Poser.

In order to minimise the effect of a cluttered background and limb ambiguities for human segmentation, the experiment highlighted the background and human limbs in the pre-processing. Based on the separated background and object, the sample data is outlined by a segmentation method or highlighted manually using Photoshop. Figure 3.5 summarises the sample data and its segmentation. All the segmentation and regression are performed on a 4200+ AMD Athlon machine. Forty data sets are obtained with sampling frequency of 10.36 KHz. Each data set includes 50 samples. To test the performance, the samples are further divided into 10 equal portions, in which 90% are used for training and 10% are used for testing the performance. The mean square error (MSE) and the linear correlation coefficients R are calculated to evaluate the quality of the regression.

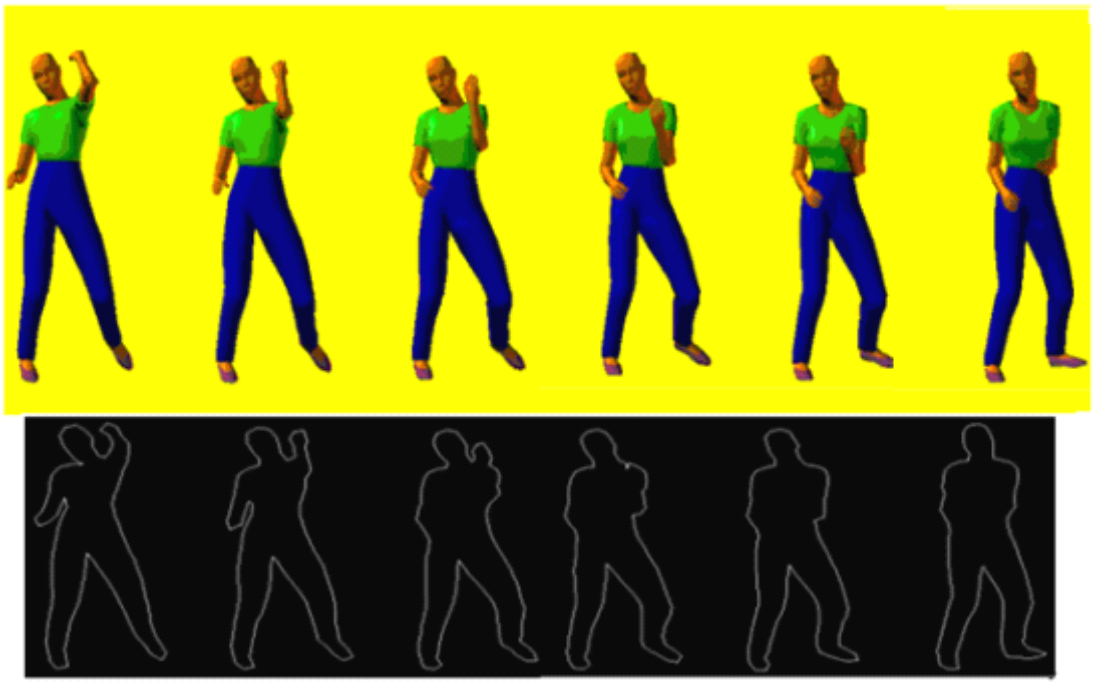


Figure 3.4. Sample data and segmentation have different background colour, which differentiates the limbs. Human data is created with a frame frequency of 12.36 KHz. The data is imported into Poser with a sampling frequency of 10.36 KHz.

For the algorithm for training and testing, the value of MSE and R are calculated in following experiments.

3.4.1 Data collection

Since this is a pioneer research, there are no publicly available human data for this type of research. Human motion is created with clear torso differentiation and a static unified background in BVH files. The images are rendered for each pose from Poser to capture inter-person variations and increase the amount of the training data to 300 pose-image frames (see figure 3.5) to study the effect of inter-person variations in appearance, which is a part

of the motion capture data used to create multiple training images in this manner.

Figure 3.5 A sample pose rendered using different torso angles

In order to test the extensive application of the methodology, the thesis obtained 300 human motion data frames in various kinds of movement as training data set. In addition, the data describes different upper body angles. For example, the main angles (with the range variation in parentheses) for some of the joints are as follows: body-head angle, 17°(360°); left shoulder angle, 7.5°(51°); and right hip angle, 4.2°(47°).

3.4.1 SVR kernel functions

The kernel function supports the transformation from nonlinear to linear in high-dimensional space. The kernel function has two common forms:

(1) polynomial kernel function,

$$K(x,x') = (.xx'+1)^p \qquad p = 1,2,...,n \tag{3.31}$$

and

(2) radial kernel function

$$K(x_i, x) = \exp(-\lambda \left\| x - x_i \right\|^2) \quad (x_i \in X) \tag{3.32}$$

The kernel function and the parameter λ determine the type of sample in the distribution of high-dimensional linear space. In the experiment, Gaussian radial basis function is used to represent kernel function as follows:

$$K(x_i, x) = \exp(-\frac{\left\| x - x_i \right\|^2}{2\sigma^2}) \quad (x_i \in X) \tag{3.33}$$

When the kernel function takes a greater value of σ, its performance is similar to the polynomial kernel function. In contrast, when the kernel function takes a smaller value of σ, it has a similar performance with a linear kernel function. (Wang et al., 2003; Walczack and Massart, 2000).

To compare the performance of these kernel functions in the SVR algorithm, the sample data is used for training with different parameters using MSE given by:

$$\frac{1}{l}\sum_{i=1}^{l}\left(\frac{y_i - f(x_i)}{\Delta y}\right)^2 \tag{3.34}$$

with the parameter R given by

$$R = \frac{\sum_{i=1}^{l}(y_i - \mu_y)(f(x_i - \mu_f)}{\sqrt{\sum_{i=1}^{l}(yi, \mu_y)^2 \cdot \sum_{i=1}^{l}(f(x_i) - \mu_f)^2}} \tag{3.35}$$

T evaluate the performance of the regression, in equation 3.34, $y_i - f(x_i)$ indicates the difference of value between the sample and the prediction with Δy given by $\Delta = \max(y) - \min(y)$ and μ_y is the calculation of the average of samples. This is given by

$$\mu_y = \frac{1}{l}\sum_{i=1}^{l} y_i \tag{3.36}$$

The variable μ_f is given by

$$\mu_f = \frac{1}{l}\sum_{i=1}^{l} f(x_i) \tag{3.37}$$

which is the calculation of the average of prediction. Moreover, this thesis uses a Gaussian radial basis function by using three parameters, (C, ε, σ) rather than (C, ε, p), in the polynomial kernel function.

Figure 3.6 summarises the test set performance of the various regression methods, which is a learning-based version of the SVM converted into a kernel. For the full human sample data and various subsets of it, an optimal regularising setting is computed using cross-validation and the steepest decent algorithm. The kernelised SVM with different parameters are normalised and compared by training and testing variances.

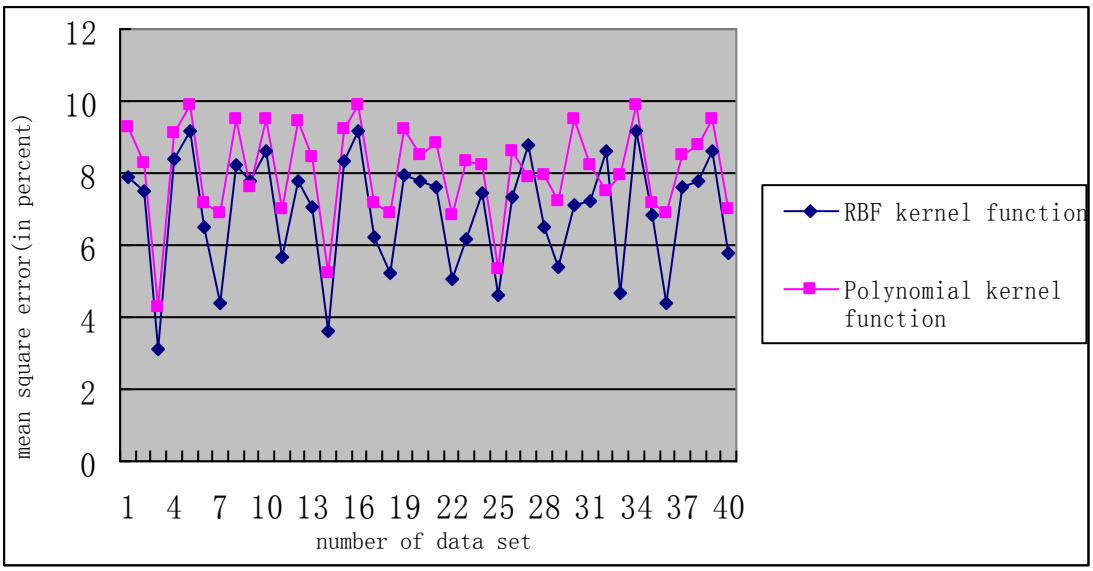


Figure 3.6.. Performance of the regression subject to the different kernelised SVM for the sample data training

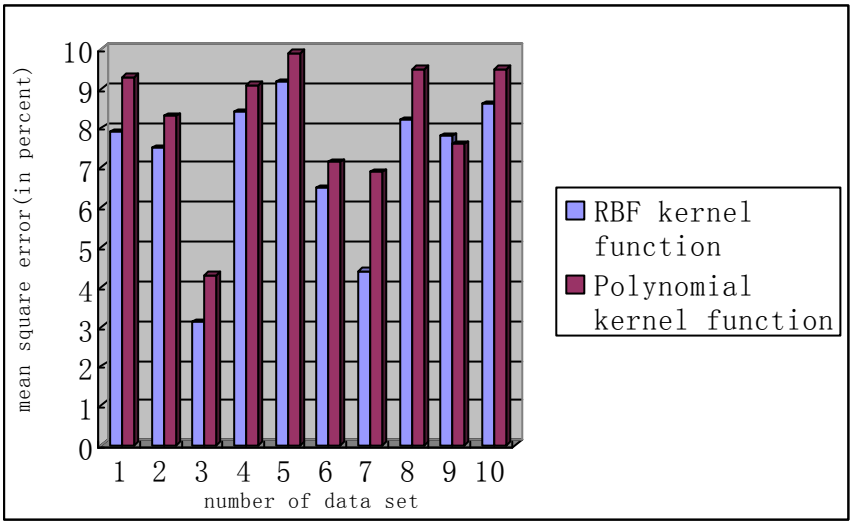


Figure 3.7. Testing results of MSE for the regression with different kernel functions and parameters.

Figure 3.6 shows the average performance based on 40 sample data sets using different kernelised SVM. The RBF kernel function has an average of 6.91% MSE in the 40 training sample data sets. The experiment uses 10%

of training samples for testing, thereby utilising the average performance in the ten testing experiments. Figure 3.7 summarises the average performance of ten experiments. Compared with the average performance (MSE=8.2%) of the polynomial kernel function in the 10 tests, the radial basis kernel function has a performance of 7.1% in MSE.

Moreover, experiment uses MSE and the retained support vectors to illustrate the influence of different training and testing sets with the different kernel functions to the performance.

Training	8	36	45
Testing	42	14	5
Training MSE	0.17	0.13	0.11
Testing MSE	0.23	0.18	0.20
% of support vectors retained	13	69	86

Table 3.8. A summary of SVR with the polynomial kernel function. The first two rows show the numbers of training and testing. The bottom rows show the error measures for SVR using the polynomial kernel function bases with the number of support vectors retained.

Training	8	36	45
Testing	42	14	5
Training MSE	0.11	0.08	0.07
Testing MSE	0.18	0.15	0.11
% of support vectors retained	15	71	89

Table. 3.9. Summary of SVR with the radial basis function. The first two rows show the number of training and testing. The bottom rows shows the error measures for SVR using the polynomial kernel functions bases with the number of support vectors retained.

Tables 3.8 and table 3.9 show part of the training and the testing results using the different kernel function in sample data sets. The number of sample data sets is 40, and each data set has 50 samples. Figures. 3.8 and 3.9 also summarise the performance of the different training and testing with the kernel-based SVR for one data set. With an increase in the number of training examples, the training and the testing MSE decrease. The MSE performance is improved from the polynomial kernel function to the radial basis function. For the experiment, the Gaussian radial basis function is used with σ estimating a scatter matrix of training samples. Furthermore, the regression parameters are given as $(C = 15, \varepsilon = 0.16, \sigma = 3.5)$. The average performance curves of the polynomial kernel function and the radial basis function in 40 data sets are shown in figures 3.6 and 3.7.

3.4.2 Parameters of SVR using different kernel function

The experiment applies a cross-validation method in the 40 human data sets. It uses 90% of sample data for training with the other 10% for testing. The results are based on the average performance of ten experiments. For the experiment, the regression performance used the different kernel functions, as shown in table 3.10.

ε	C	p	Training MSE	Maximum training error	Testing MSE	Maximum testing error
0.01	10	2	0.15	0.38	0.31	0.61
0.01	10	1	0.08	0.21	0.10	0.26
0.01	20	1	0.03	0.10	0.09	0.24
0.05	10	1	0.16	0.40	0.20	0.45
0.005	40	3	0.006	0.03	0.49	0.80

Table. 3.10. SVR parameters on the effects of regression results (polynomial kernel functions)

Table. 3.10 shows the regression performance of the different parameters with polynomial kernel function. The parameter p has an ideal value of 1. When p>1, such as that of the last row, the value of p=3 is displayed. The training MSE and the maximum training error have an increase in training accuracy; however, the generalisation is decreased, and the testing MSE and maximum testing error decrease by 30% and 40%, respectively. This shows the influence of the parameters ε and C on the regression performance. Hence, based on table 3.10, the value of the SVR parameters is not only associated with the training accuracy, but is also related to the generalisation. In this experiment, the MSE in the training and the testing are used to evaluate the regression.

3.4.3 Performance of parameters with optimal solution

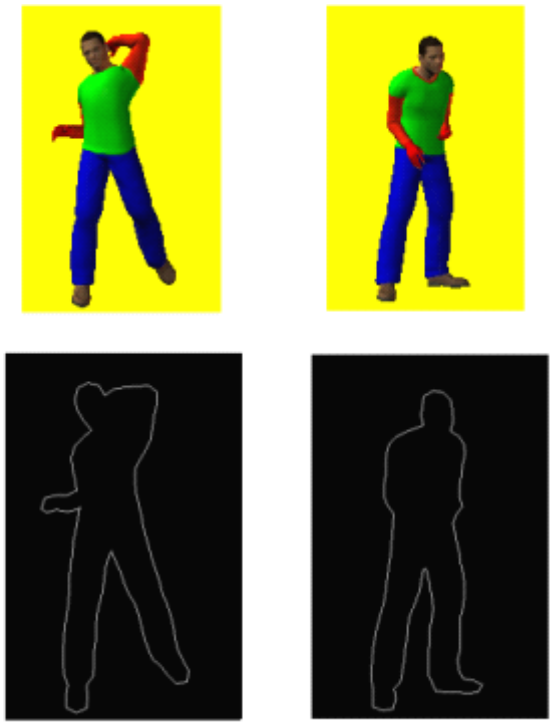
The SVR method is based on the idea of SRM. Using-cross validation and the steepest descent algorithm, the three parameters of (C,ε,σ) optimise the solutions after iteration.

Training set	SVR with optimised parameters	Artificial Neutral Network
Training set		
MSE	0.091	0.104
R	0.825	0.712
Testing set		
MSE	0.098	0.104
R	0.810	0.698

Table. 3.11 Average MSE and correlation sufficient measures for the data sets using the Gaussian kernel function.

Table 3.11 shows the comparison of SVR with optimised parameters and the ANN method. The SVR method obtains 22% improvement in MSE performance, where the optimised parameters are $(C = 15, \varepsilon = 0.16, \sigma = 3.5)$.

3.5 Testing results



(a) (b)

Figure 3.12.. Outline of two source images. To simplify the testing process, the backgrounds of source images are highlighted prior to processing.

Figure 3.12 shows the data imported by using Curious Labs Poser. To distinguish the object from background, the background colour is highlighted differently from the human body. The segmentation method is applied for image contour extraction. The depth information is not expressed in the outline images because it is irrelevant for the synthesis in this section. In order to simplify the calculation, 100 contour points are manually selected.

3.5.1 The testing results with different kernel functions

For our experiments, 40 data sets of motion capture data are divided into training sets of several motion sequences (1800 poses in all), including various kinds of limbs motion viewed from same direction, and a test set of 200 frames sequences of a person's movement. The contours corresponding to these poses are extracted using Open Source Computer Vision (OpenCV).

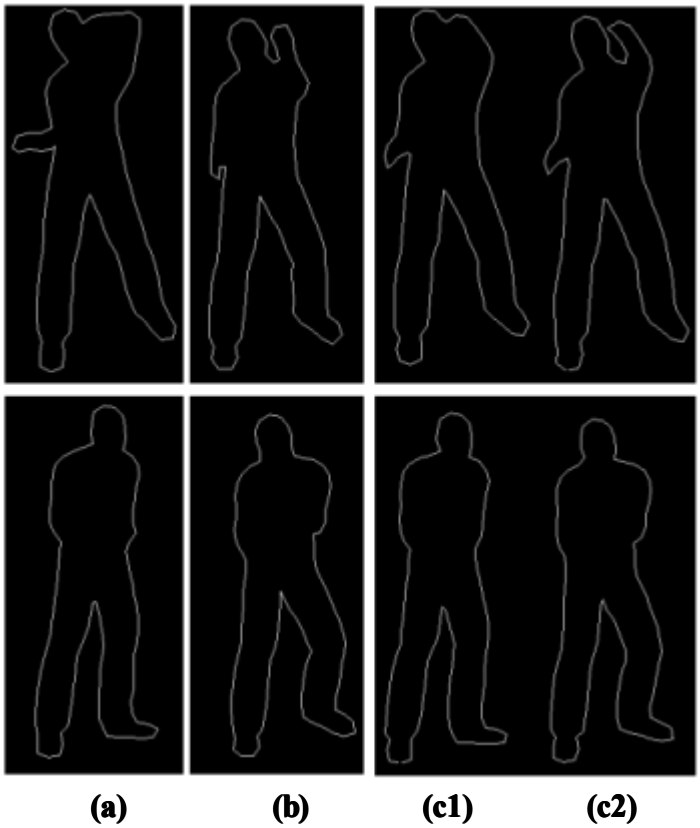


Figure 3.13. Image synthesis based on one-sample regression. The results are compared using different kernel functions. The synthesis is computed with the polynomial kernel function and Gaussian radial basis kernel function using 200 images from 40 data sets as input images: (a) input images, which are source images; (b) contour images of the training examples; (c1) synthesis result of the

regression with Polynomial kernel function; and (c2) synthesis result of the regression with Gaussian radial basis kernel function. Using the Gaussian radial basis, the kernelised regression will approximately have 9% incorrect result synthesis, compared to 13% incorrect results if the polynomial kernel function is used.

The results of the experiment are shown in figure 3.14 below.

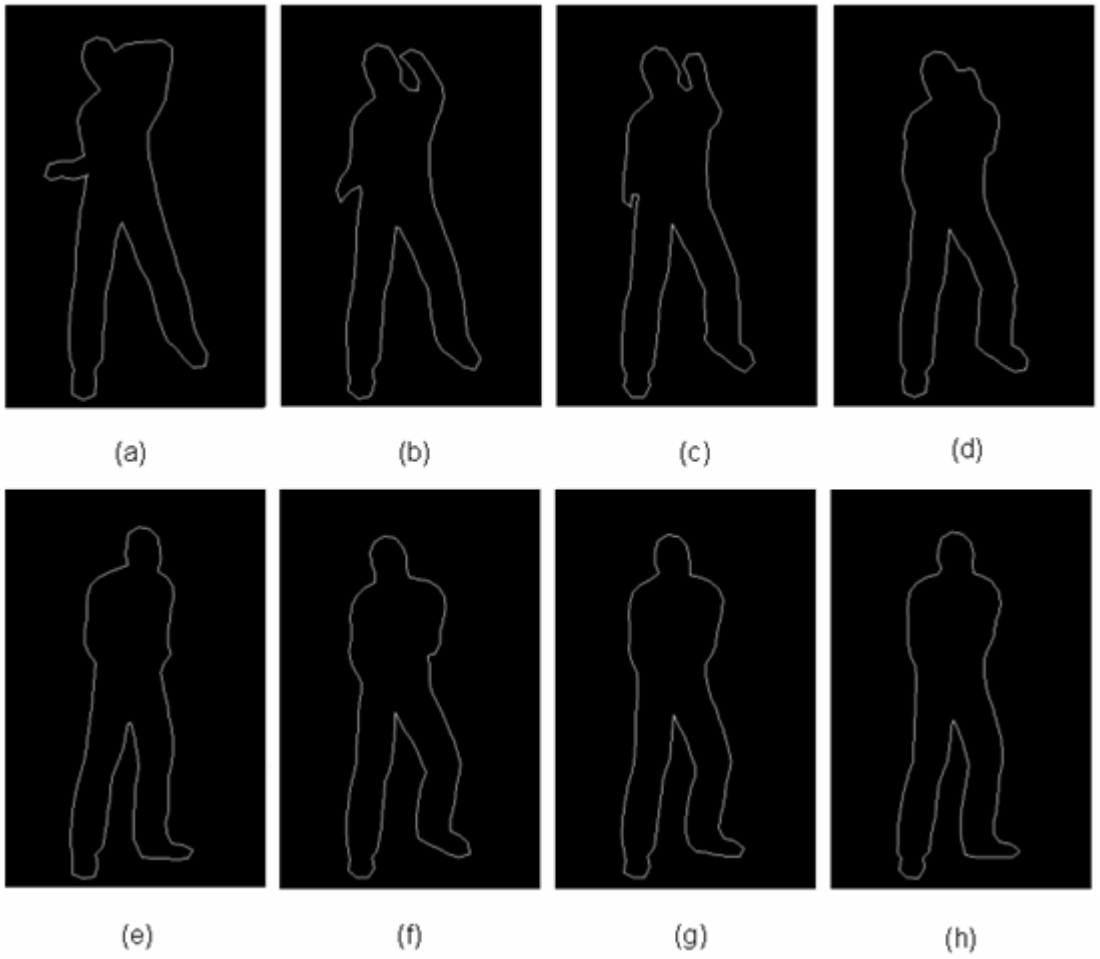


Figure 3.14. Source images and synthesis images: the source image is (a) whose synthesis results are generated as (b), (c) and (d) by using three different regression functions. For (e)–(h): image depth information is not considered with the human pose information. Source image (e) has three synthesised images, as shown in (f)–(h). On average, compared with the training examples, the result has an acceptable MSE of 7.2%.

Figure 3.14 summarises a part of result frames from two source images. For

this experiment, the contour of the source image is extracted by the segmentation method using a static background. The open source codes from OpenCV are applied for contour extraction in the 40 sample data sets. Moreover, to simplify the regression calculation, 100 independent contour points are selected manually for the regression functions generated by training with the data sets.

The human body is a non-rigid object in image processing. Human body movement has many degrees of freedom and a high degree of nonlinear characteristics. In addition, the variability of the human body appearance leads to a large number of variations. In figure 3.14, the contour synthesis shown at the bottom row illustrates the ambiguities of human movement, including the internal information such as human limb information, facial information, and others.

Different persons have different ranges of postures and wear different clothing. Therefore, the human body structure involves great differences in appearance and is difficult to express in a unified model. Without considering the human facial information, our experiment involves human body model with varying clothing information to test our methodology.

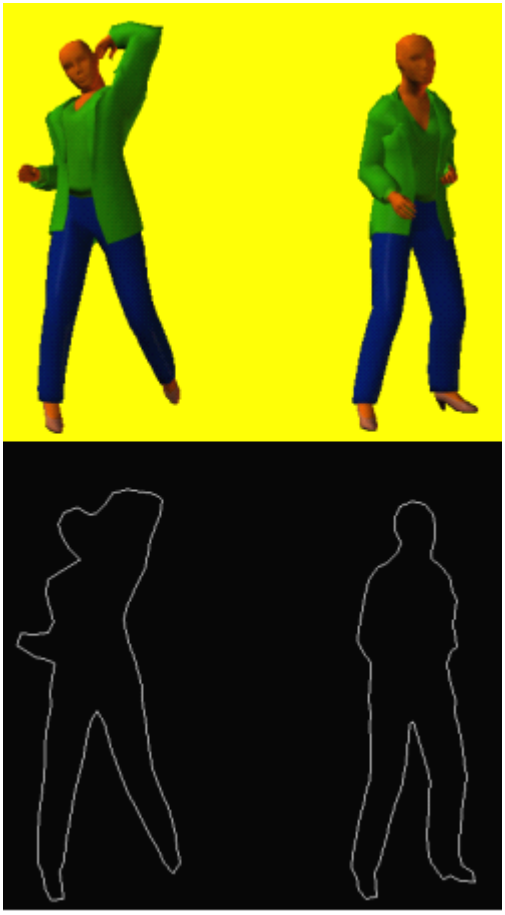


Figure 3.15. Contour for source images with clothing information

In figure 3.15, human clothing information is one kind of human body outline considered in our contour method to make the outer lines continuous and smooth. To conveniently synthesise the outline, the clothes colour is set to be clearly different from the background and the human body colour in Curious Labs Poser.

Figure 3.16. Synthesis results for the human body with clothing information

In figure 3.16, the leftmost and the rightmost images are input images, which are treated as source images, whereas the other six images are synthesised images. The results when human clothing information is considered are shown in figure 3.16. Due to the ambiguities of human clothing, two input images may have considerable pose differences. However, as the edge and corner of human clothing are bounded by the outline, the technique mentioned above produces good synthesis results. For the source images, the contours of the input images are extracted by the segmentation method with 100 contour points manually selected for regression functions.

3.6 Summary

This chapter has introduced the approach to incorporate the SVR. The regressive function is conducted by training the contour points from example data sets. The whole body regressor retains 100 contour points for training as follows: the torso, 50; the arm, 25; and the leg, 25. Selection of the right solution by using regression of the contour points and considering the two current images to produce reasonable results is a challenging problem. Glitches happen occasionally due to the presence of multiple solutions. The regressor either needs to select a wrong solution or the regressor outputs a compromised solution. Notably, the contour points are manually selected in the source images to fit the regression function and to deduct the influence of multiple solutions from regression.

Moreover, the steepest descent algorithm and cross-validation methods are applied to obtain the optimised parameters and minimise the MSE. Due to the reasonable results and quite robust process, these methods match the training outline with lower MSE. The Gaussian radial basis kernel function is applied to show the robustness in regression.

Finally, regression with the different kernel functions and regression parameters are conducted to determine their performance. The effectiveness and efficiency of the synthesis method have been proven with this pioneering

research. The novelty of this research is the implementation of regression in an image-based method. However, the limitation of the method proposed is that it is only concerned with the outline of human body. For the internal details, especially when the human model involves facial information and clothing information, the synthesis method can be achieved through contour colouring, which is presented in the next chapter.

4. Pixel synthesis

The key proposal of the technique is to apply the inference in the Markov random field (MRF) to the synthesis of pixel colours. Therefore, the problem has become a process of determining the values of the MRF nodes (i.e. pixels).This purely image-based rendering uses contours synthesised from Chapter 3 and carries out colouring within the contours through belief propagation (BP). Photographs of the human body have been used as the input images.

Over the past few years, there have been exciting advances in the development of algorithms for solving early vision problems such as stereo, optical flow and image restoration using the MRF model. The MRF formulation of these problems yields to an energy minimisation that is an NP hard problem. Good approximation algorithms based on graph cuts (Boykov and Veksler, 2001) and on the BP (Weiss and Freeman, 2001; Sun and Zheng, 2003) have been developed and demonstrated for the problems of stereo and image restoration. These methods are good both in the sense that the local minima they find are over ’large neighbourhoods’, and in the sense that they produce highly accurate results in practice. A comparison between the two different approaches for the case of a stereo is described in the work of Tappen and Freeman (2003).

Despite these substantial advances, both the graph cuts and BP approaches require several minutes of processing time, even on today's fastest desktop computers, for solving stereo problems. Furthermore, the approaches are too slow for practical use when solving optical flow and image restoration problems. Thus, one is faced with choosing between methods that produce good results but are slow, or local methods that produce substantially poorer results but are fast. This chapter presents a new algorithmic technique applied in a synthetic research that substantially improves the running time of BP for solving the colour synthesis problem

Based on the results from the previous chapter of contour synthesis, this chapter describes a learning-based method for synthesising the colour from a source image sequences to a synthesised contour image. The most relevant synthesis work is primarily focused on 3D geometry (Koch, 1995; Scharstein, 1999; Scharstein and Szeliski, 2002). In contrast, this thesis primarily transforms the problem to consider the image colour rather than depth. The key proposal of the technique is to apply the inference of MRF to the pixel colour synthesis problem. Hence, the problem has become a process of determining the values of the MRF nodes. BP is selected as the MRF inference. It makes use of photographs of the human body as input images.

BP, as a primary and effective method in computer vision and graphics has been used in MRF (Rabiner, 1989) research, max-product belief propagation (Pearl. 1988), optimal control methods (Bertsekas, 2001), and resource allocation (Bellman and Karush, 1961). Moreover, by ignoring the influence of the lighting environment, the difference between pixels is generally invariant. Therefore, the colours are computed from the input source images to the synthesised images by constructing the link distance field using a controlled BP program for matched points between source and synthesised contours. The synthesis is applied as a rendering process in two distinct steps:

1) Link distance construction by using a controlled BP scheme; and

2) Complete the colours within the synthesised contours via BP with the link distance.

Based on the proposed techniques, the colouring process is as shown in figure 4.1.

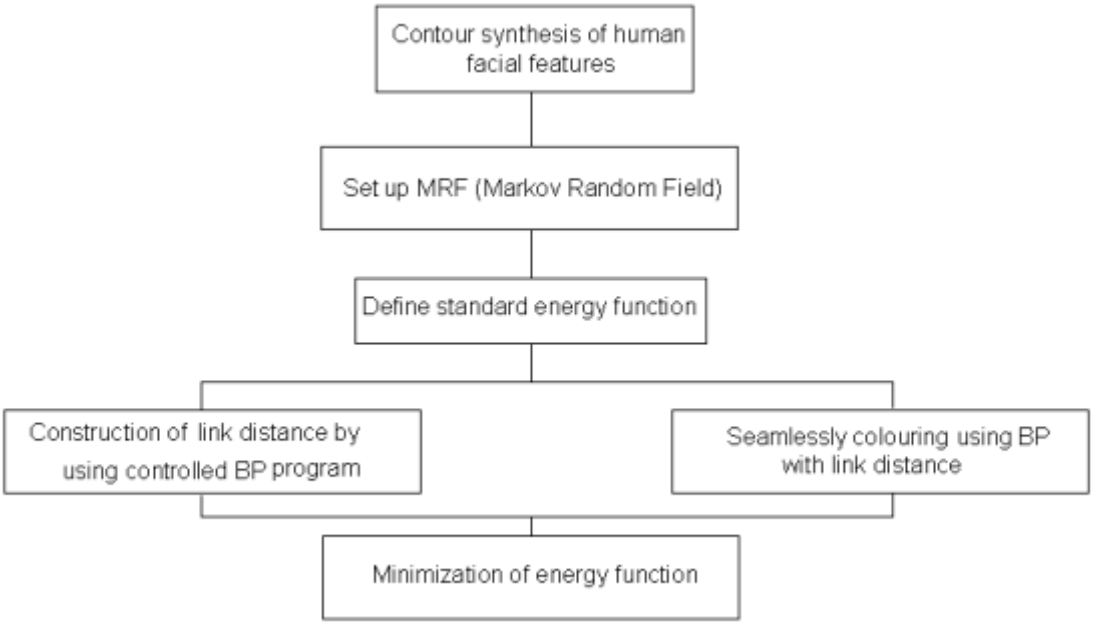


Figure 4.1. Flowchart of the pixel synthesis

4.1 Contour colouring via BP

The general definition of an energy function includes two items: 1) D is used to represent the data cost for one single pixel and 2) V is used to indicate the smooth adjacent continuity between two neighbouring pixels that gives the link cost. The solution is obtained by minimising the energy function. Hence, the standard energy function is

$$E(f) = \sum_{p \in N} D\left(f_p\right) + \sum_{p,q \in N} V\left(f_p, f_q\right) \tag{4.1}$$

where N is the edge of the pixels of the graph with eight connected neighbours. The first item denotes the data cost for the pixel p when the pixel is assigned a single value f_p. At the same time, the second item is the

adjacent smooth continuity that gives the link cost of assigning the values f_p, f_q to two neighbouring pixels p and q. The minimum energy corresponds to the MAP estimation problem with the defined MRF.

4.2 Constructing the link distance field through an assigned BP program

The controlled BP program can identify the correspondence for each pixel in a synthesised image and a colour source image by examining the set of corresponding pixel disparity. In the message propagation process, assigned points are involved. The explicit setting value of assigned points that come from the contour data points determines the desired results.

To identify the details of the process of the BP program, the input images are denoted as L_n where $n = 1, ..., N$. For each data point d on the contour Z in the synthesised images R, there is a corresponding d' in S_n, to identify each pixel p and its corresponding input and synthesised images. The assigned BP program is applied to examine the set of disparity of the correspondence pixels to achieve energy minimisation.

Therefore, for the candidate set, the disparity needs to be initialised by placing a window for p. The size of the window should be sufficient for the

disparity of p in the synthesised image and the correspondence p' in the source images:

$$\Delta Gpp' = p' - p, \| p' - p \| < w_p \tag{4.2}$$

All the assigned points come from the data points d on the contours Z in R. The form of the set of data points d on the contours Z in R is as follows:

$$v = \{ d | d \in Z, \forall Z \in R \} \tag{4.3}$$

Thus, the collection of disparity for the contour points is

$$M = \{ \Delta d' | \Delta d' = d - d', \forall d \in v \} \tag{4.4}$$

where d' stands for the correspondence point of d in S_n.

Hence, the data cost is defined as follows:

$$D(f_p) = \begin{cases} 0, \ if \ p \in v, \Delta p \in M \\ \infty, if \ p \in v, \Delta p \notin M \\ a, \ if \ p \notin v \end{cases} \tag{4.5}$$

Here, the data cost will contain the lower cost when pixel p is assigned a point and the assigned value of the disparity Δp comes from the value of

the disparity for the data points.

As the result of the assigned BP, each pixel p will have disparity Δp in the synthesised image. To identify the correspondence between the synthesised image and the input image, the correspondence is computed as

$$T(R, L_n) = \left\{ p' \middle| p' = p + \Delta p, \ \forall p \in R \right\} \qquad n = 1, ..., N \tag{4.6}$$

where N is the number of colour source images, and p' is the correspondence of p in S_n.

From this, the link distance construction is represented using a probability method. Two neighbouring pixels, p and q, are assumed to have the link distance $L(pq)$. The density distribution function of the pixel value at p is also assumed to obey the Gaussian distribution. Therefore, for each distance p in the synthesised image, the probability of having the pixel value at p is defined by its correspondence in S_n using a mixture Gaussian distribution given as

$$P(p) = \sum_{n=1}^{N} W_n(p) \mathrm{N}(p \middle| \mu_n, \sigma^2) \tag{4.7}$$

where N is the mixture Gaussian distribution with centre $\mu_n = m_n(p')$. This is

also the pixel value of p' from S_n with a standard deviation σ. In practice, the standard deviation σ is set to around 10. $W(p)$ is the weight of one colour source image at p. For all the colour source images S_n, $W(p)$ needs to be normalised as $W_n(p)$ over the whole S_n. Intuitively, giving the fixed value to pixels will lead to a low data cost. The data cost has an anti-proportional relation to the pixel value probability that can be obtained from equation 4.7.

According to the low data cost, fixed value pixels are often observed as the one with the high pixel value probability and $W_n(p)$.

Moreover, the density distribution function of $L(pq)$ obeys the Gaussian distribution. Therefore, for each distance pq in the synthesised image, the probability of having the distance of $L(pq)$ is given by mixture Gaussian distribution as

$$P(L(pq)) = \sum_{n=1}^{N} W_n(pq)\mathrm{N}(L(pq)\left|g(p'), \sigma^2\right.) \tag{4.8}$$

where $W(pq)$ is the weight of one colour source image at pq and is given by the ratio of window size T_{pq} and the correspondence T'_{pq} for all the colour source images S_n. Here, $W(pq)$ needs to be normalised as $W_n(pq)$ over all the S_n.

4.3 The initialisation of disparity

To calculate the colour information of the labelled pixel node corresponding to Yoon's (2005) work, it needs to be initialised at the disparity image. This means the matching correlation of pixels should be obtained with their adjacent ones.

Hence, p,q are assumed from the same pixel data in one frame of the contour. The link distance is described as

$$\Delta G_{pq} = \|p - q\|_2 \tag{4.9}$$

Therefore, the weight of distance of p, q is

$$W_{pq} = \exp(-\gamma \Delta G_{pq}) \tag{4.10}$$

According to equation 4.1, if a window is placed with the size T_{pq} around pixel p and pixel q to decide the link distance pq, the window size T_{pq} can find the correspondence as

$$T'_{pq}(R, L_n) = \left\{ T'_{pq} \middle| T'_{pq} = S_{pq} + \|p - q\| \right\}, \|p - q\| < w_{pq} \tag{4.11}$$

4.4 Message computed between two neighbours

As the standard energy function indicated in 4.1, messages in the propagation can be used in an optical flow, restoration, and image representation. The general framework can be defined through the function between the pixel set and the labelling set.

4.4.1 Message in BP

The research begins by briefly reviewing the BP approach for performing inference on Markov random fields (Weiss and Freeman, 2001). In particular, the max-product algorithm can be used to find an approximate minimum cost labelling of the energy functions in the form of equation 4.1. Normally, this algorithm is defined in terms of probability distributions. However, an equivalent computation can be performed with negative log probabilities, where the max-product becomes a min-sum.

The BP method is an approximate calculation method. It is based on performing inference on the MRF method to solve the problem of a maximum a posterior estimate. In particular, the algorithm can be used to find an approximate minimum-labelled energy function. Furthermore, the BP algorithm works by transforming the message between the two neighbour nodes in the graph defined by the four-connected image grid. In this case, the thesis has

represented a premier practice in disparity defined by eight-connected compared results, as shown in the energy curves. This formulation is used because it is less sensitive to numerical artefacts, and it uses the energy function definition more directly. The max-product BP algorithm works by passing messages around the graph defined by the eight-connected image grid in this thesis. Each message is a vector of the dimension given by the number of possible labels. Let $m^{t}_{p\to q}(f_q)$ be the message that the node p sends to a neighbouring node q at time t. When using negative log probabilities, all entries in $m^{0}_{p\to q}(f_q)$ are initialised to zero. At each iteration, new messages are computed in the following way.

Thus, for any $D(f_p)\propto -\log\Phi(f_p,f_q)$ and $V(f_p)\propto -\log\Psi(f_p,f_q)$, the iteration of information transmission can be expressed as

$$m^{0}_{p\to q}(f_q)=\min_{p}\left(D(f_p)+V(f_p,f_q)+\sum_{k\in N(p)\setminus q}m^{t-1}_{k\to p}(f_p)\right) \tag{4.12}$$

Figure 4.2 is the indication of (4.12), wherein $N(p)\setminus q$ refers to the neighbour of p other than q. In figure 4.2, p and q are adjacent nodes. The message from p to q is $m_{p\to q}$, and $(d_1,d_2,d_3,\cdots,d_M)$ is the set of labels of pixels at p and q.

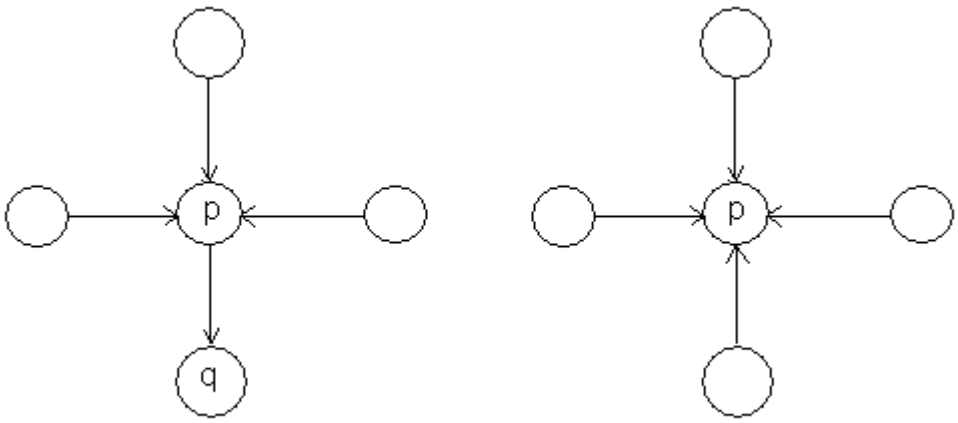
BP can be written as

$$b_q(f_q)=D(f_q)+\sum_{k\in N(q)}m^{T}_{k\to q}(f_q) \tag{4.13}$$

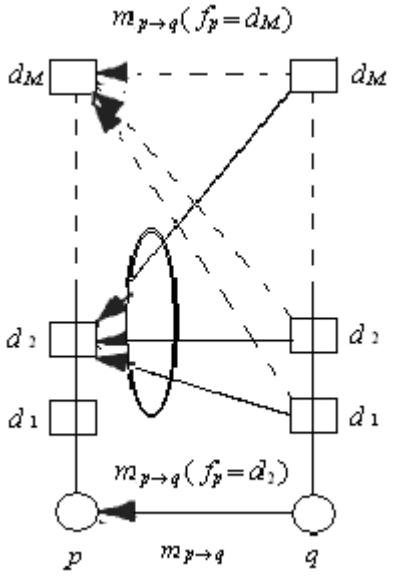
where t is the number of iterations. This can be expressed as figure 4.2. The value of the label for pixel q can be estimated as

$$f_q^{MAP} = \arg\min_q b_q(f_q) \tag{4.14}$$

Thus, finding the label with the minimum energy corresponds to the MAP problem in a correct MRF wherein the estimation of the label at each pixel can refer to the quantities of disparity, vectors, etc. (Pedro et al., 2006).



a **b**



c

Figure 4.2. Message propagation between adjacent pixels

Figure 4.2 illustrates the message passing in the adjacency graph of the given pixel. If node i is left of node j then node i sends a message to node j at each iteration. Node i contains the messages already received from its neighbours. In parallel, each node of the adjacency graph computes its message. Then, those messages are sent to adjacent nodes in parallel. Based on these received messages, the next iteration of messages is computed. For each iteration, each node uses the previous iteration's messages from adjacent nodes to compute the messages sent to its neighbours. The larger the Dp(fp), the more difficult the task of passing a message to an adjacent node. This means that the influence of an adjacent node decreases when the cost at this node increases.

According to Pedro et al. (2006), an efficient algorithm divides the

image pixel nodes field into two main interval maps. In figure 4.3, image pixel nodes have been divided into two parts, A and B, where the circle represents A and the rectangle represents B. Therefore, if the message from A to B is known during the iteration t-1 (Figure 4.3 b), then the message from B to A at iteration t can be obtained as follows (Figure 4.3 c):

$$\overline{m}^{t}_{p\to q} = \begin{cases} m^{t}_{p\to q} & if\ p \in A\left(p \in B\right) \\ m^{t-1}_{p\to q} & otherwise \end{cases} \tag{4.15}$$

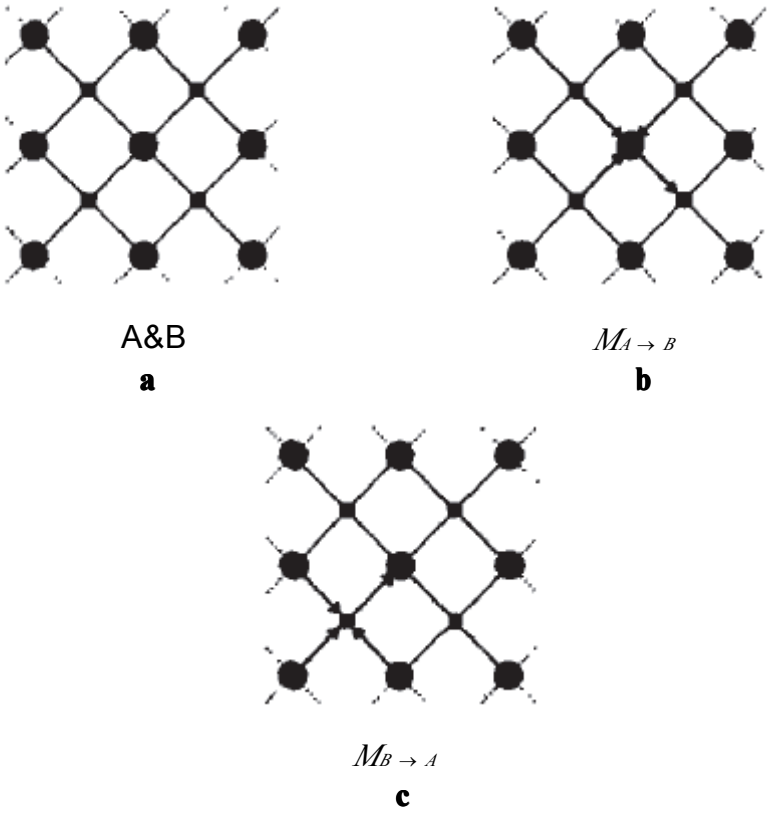


Figure 4.3. Efficient algorithm for propagation (Pedro et al., 2006)

Therefore, an assigned set f is the main part of the message that can

refer to vectors, intensity, etc. In this thesis, the labels correspond to the disparity for two neighbouring pixels p and q, and the same disparity value between neighbouring pixels may be implied as

$$m^t_{pq}(f_q) = \min_{fp}(d(p,q) + h(f(p))) \tag{4.16}$$

where $d(p,q)$ is the measure of the distance between p and q. Intuitively, for each point p, the nearest point q will have smallest value of f(q). Note that if f has a small value at some point and a nearby point, the equation will have a small value at that location and at the point nearby.

4.4.2 Message computation

The cost function computed in quadratic is useful in one-dimensional distance. For the squared Euclidean one-dimensional distance (Felzabszwalb, 2004) applied in this thesis, the description of the distance transform is given by

$$m^t_{pq}(f_q) = \min_{fp}((p-q)^2 + h(f(q))). \tag{4.17}$$

Note that the distance transform of f will be bounded by a parabola rooted at (q, f(q)).

4.5 Initialisation of eight distance

To justify the $g(p')$, the set of the distance at pixel p is denoted as

$$g(p) = m(p) - m(N(P)) \tag{4.18}$$

where $N(P)$ is the set of neighbouring pixels of p. Eight distance values are utilised as top, down, left, right, top left, top right, lower left, and lower right. The relation is illustrated in the figure below.

$g_{tl(p)}$	$g_{t(p)}$	$g_{tr(p)}$
$g_{l(p)}$	p	$g_{r(p)}$
$g_{ll(p)}$	$g_{d(p)}$	$g_{lr(p)}$

Figure 4.4. Distance value of p at eight distance directions

Therefore, according to equation 4.1, with the correspondence p' in an input image, the eight-distance value illustration can be constructed from S_n as

$$g(p') = m(p') - m(N(p')) \tag{4.19}$$

which is shown in the figure below.

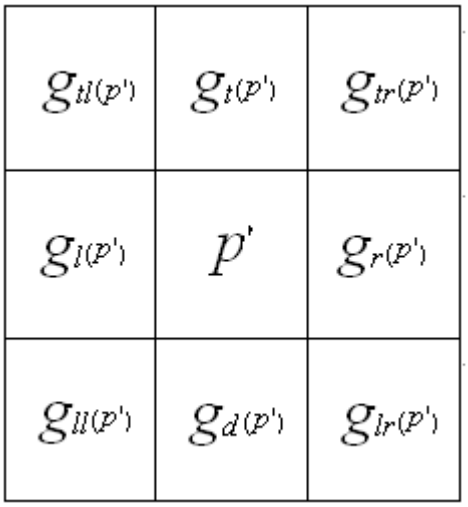


Figure 4.5.The distance value of p' at eight different directions

4.6 Results and performance analysis

The human body is a non-rigid object. In addition, the variability of the human body appearance in complicated conditions will lead to large variations. Different persons will have a range of postures and have different clothing information. Due to the colour synthesis needed to consider human pose information, facial, and clothing information, different situation combinations are used under the superior merge result in the contour synthesis method for the human body outline. This is applied practically to test the proposed methodology with different human body situations, which are divided by considering the pose information, clothing, and facial information. Experiments are applied in three steps:

1) Human poses without clothing information;

2) Human poses without facial information;

3) Combination of the three situations from different view directions.

To test the efficiency of the method, this section analyses the accuracy of the 4-connected and 8-connected distance in identifying max-deviation and synthesising running time. The thesis uses motion-capture data along with the corresponding synthesised contour image sequences for training the pixel synthesis method. For this part, the thesis renders each pose with several different human models (from Poser), and tests the 300 pixels synthesis from clear torso movement using different pixel distance initialisation.

The figure shows some comparison results using the 8-connected distance and 4-connected initialisation methods on 300 pixels employed from portions of the torso (left and right arms, hips and shoulders) in a spiral walking motion capture sequence. The mean synthesis error over all pixels for 8-connected bases in this test is 6.0. The test uses different angles of left and right arms to select the correlated pixels because the synthesis errors for individual body angles depends on their observability. The ranges of the variation of these angles, can vary significantly from angle to angle.

Figure 4.6 (top) plots the synthesis and the actual values of the overall body heading angle during the test sequence using 8-connected distance initialisation method. Much of the error is seen in the form of occasional large errors. They are referred to as ‘glitches’ and are associated with poses which have ambiguous contours. These events are strongly correlated with the synthesis errors in the degrees of torso angles, as illustrated in figure 4.6 (middle and bottom).

The comparison results for are applied in figure 4.6.

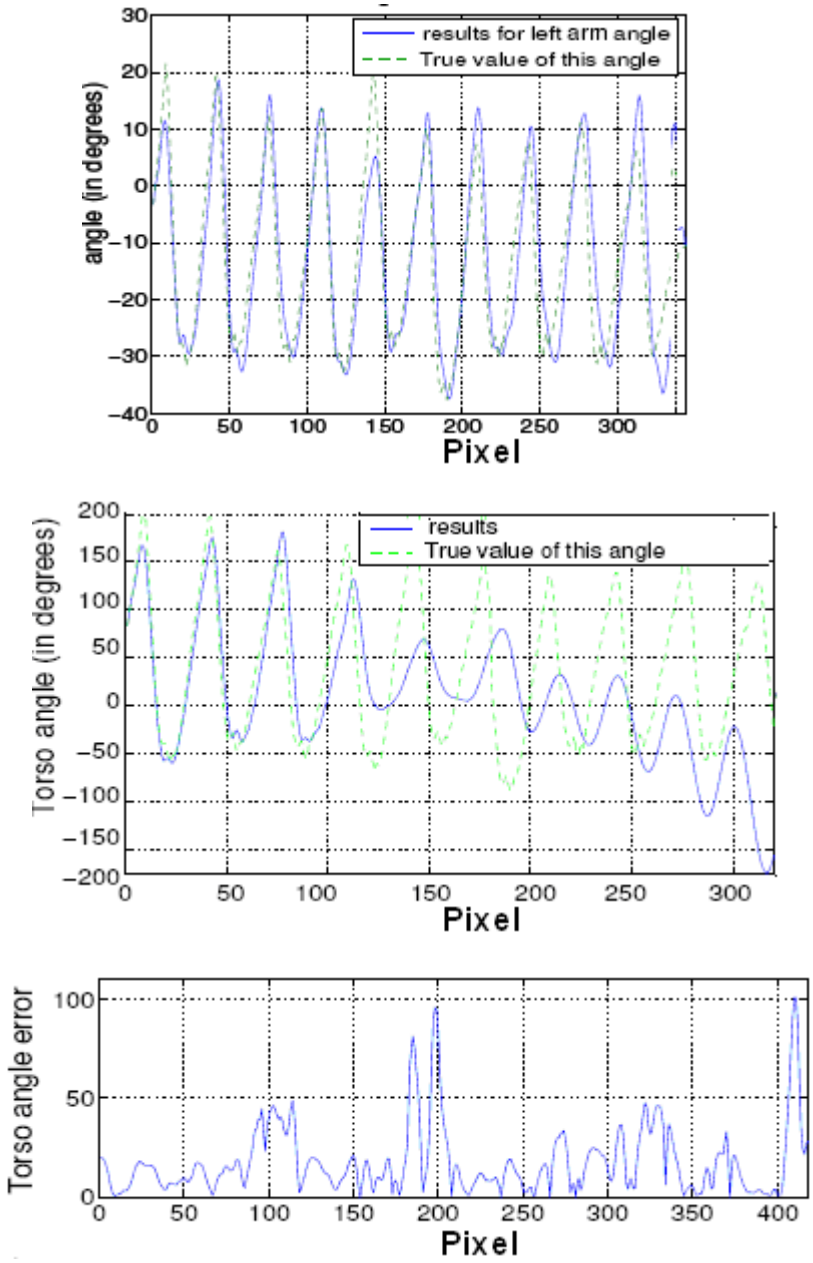


Figure 4.6. (Top): The estimated torso and left arm angle over 300 pixels of the spiral moving test sequence, compared with its actual value from the motion capture. (Middle): The estimated torso and torso angle over 300 pixels of the spiral moving test sequence, compared with its actual value from the motion capture. (Bottom): Episodes of high estimation error are strongly correlated with the torso pixels synthesis applied.

Figure	Max-deviation 4-connected	Running time 4-connected	Max-deviation 8-connected	Running time 4-connected	Iteration
4.7	70 pixel	6 m	45 pixel	8 m	9
4.9	85 pixel	9 m	60 pixel	12 m	10
4.13	96 pixel	11 m	85 pixel	12 m	13
4.14	109 pixel	12 m	90 pixel	13 m	13

Table 4.7. Error measures for the full body using 4connected and 8-connected bases distance initialisation

4.6.1 Human poses without clothing information

Figure 4.8 Source images and outline from contour synthesis

Without considering the human facial information and clothing information, figure 4.8 represents the colour source images and the results of the contour synthesis. To synthesise the colour information using the method

of contour colouring, six synthesised frames are generated to test the method proposed above.

Figure 4.9. Results of contour synthesis

Figure 4.10. In human body contour colouring, the link distance is constructed from two source images at the far left to the far right. The other four images are the synthesis results.

Figures 4.9 and 4.10 present the results of contour synthesis and colour synthesis. Four frames for synthesised images are generated. Notably, due to the occlusion condition, two input images may have occlusion problems (e.g., limb occlusion). However, through conducting the link distance, the above-mentioned technique allows the good synthesis of colours from different images.

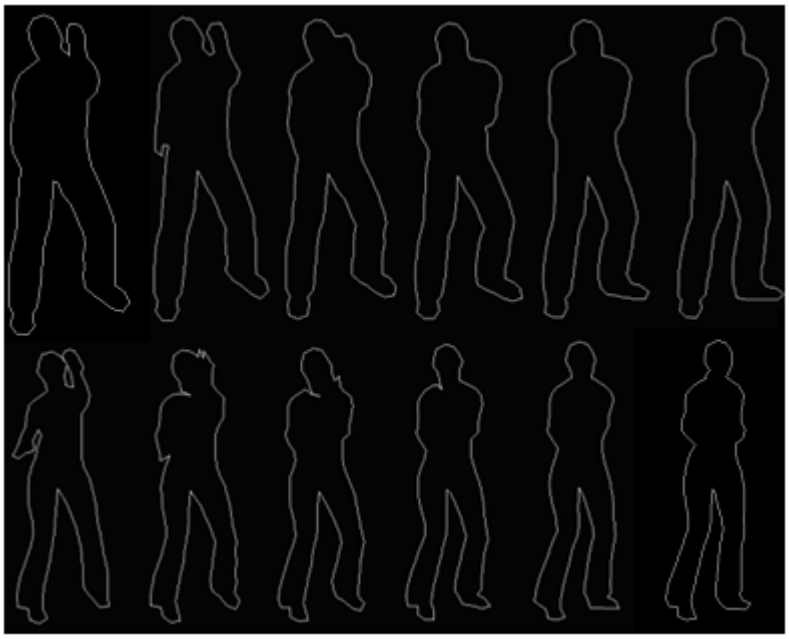


Figure 4.11. Human body contour synthesis with facial information

Figure 4.12. Human body contour colouring with facial information (male). The rightmost and the leftmost images are in-out images. The others are pixel synthesis results using the proposed method.

Figure 4.13. The human body contour colouring with facial information (female). The rightmost and the leftmost images are in-out images, the others are pixel synthesis results using the proposed method

The synthesis results of the human pose with the facial information are shown in figures 4.11 and 4.12. Source images are input images at the left most and right most. The results of synthesis contour as reference in figure 4.13 is also presented. The computation time to process the body and the facial synthesis together is challenging. Based on the proposed synthesis method, colours merge well from the source images. In contrast, the computation time is calculated as the link distance construction time takes 6–9 minutes for each input images. Combining the BP with the link distance takes another 5–8 minutes.

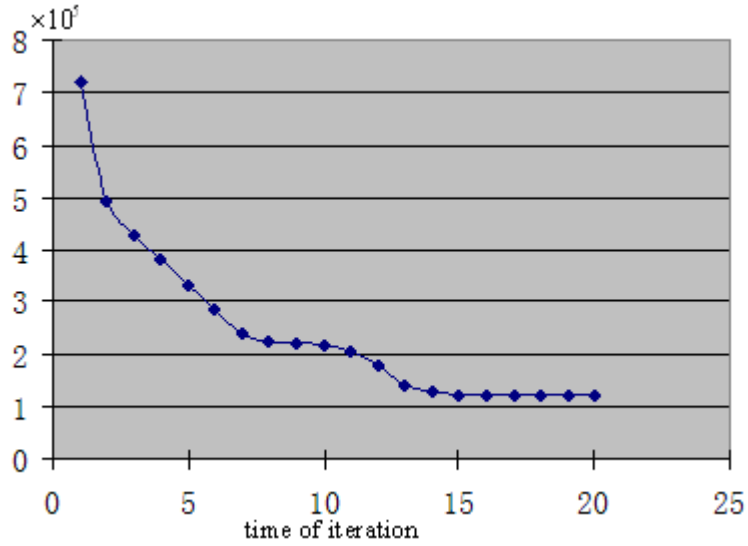


Figure 4.14. Energy minimisation based on 20 iterations

Figure 4.14 shows the energy minimisation process in pixel synthesis involving human facial features and posture. The average values are conducted based on 20 iterations. Based on the synthesis results shown in figure 4.14, the curve of the average value for energy minimisation has the minimal value of 13 iterations.

4.6.2 Human poses with clothing information

Figure 4.15. The results of the contour synthesis for human body model with clothing information. The rightmost and the leftmost images are in-out images. The others are pixel synthesis results using the proposed method.

Figure 4.16. The rightmost and the leftmost images are in-out images. The others are pixel synthesis results using the proposed method

To include the clothing information with the human poses in the contour colouring, the colour is set in a clearly different to background and the body colour. To make the process time-efficient, considering human facial information, the synthesis result in figure 4.16 is not only well-synthesised for clothing information, but also has a 5-7 minutes construction time for the link distance and the 4–7minutes combination processing time. The energy minimisation is shown in the figure below.

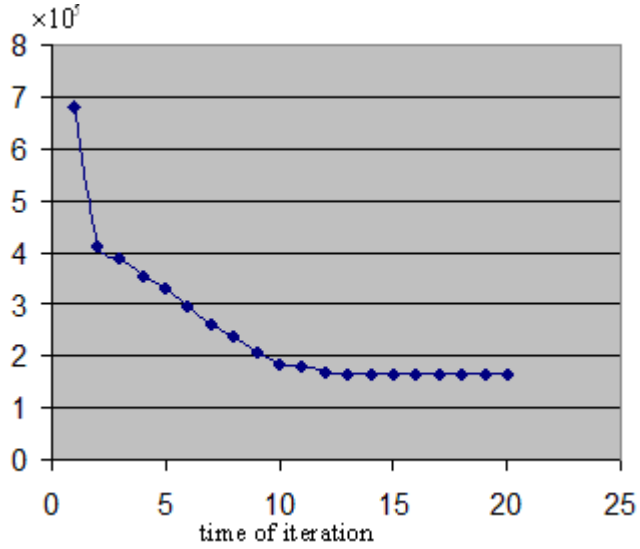


Figure 4.17. the energy curve for 20 iterations

Figure 4.17 shows the energy minimisation process in pixel synthesis which involving clothing information of the human body model. The average values are conducted based on 20 iterations. Based on the synthesis results in Figure 4.17, the curve of the average value for energy minimisation has the minimal value at 13 iterations.

4.6.3 Pixel synthesis for human body model from different viewing directions

Figure 4.18. The synthesis of the static human model with clothing information from different view directions. The rightmost and leftmost images are in-out images. The others are pixel synthesis results using the proposed method

In practice, the clothing information is employed and tested for rendering using our contour colouring method. Figure 4.18 shows the pixel synthesis results based on the contour synthesis. Human facial information and posture are not considered in the source (leftmost) and the synthesis (the rest of the images) images. For each of the synthesis experiments, 20 iterations are conducted. Figure 4.19 shows the curve of the energy minimisation process for the synthesis of images generated in figure 4.20. For each synthesis, the average value of the energy minimisation is based on 20 iterations. In Figure 4.21, the method has the energy approximately minimised at six iterations. Furthermore, The acceptable computation time of is controlled in 8 minutes. Typically, this construction that takes 4–5 minutes. Then, the BP with the link distance processing takes 3–4 minutes.

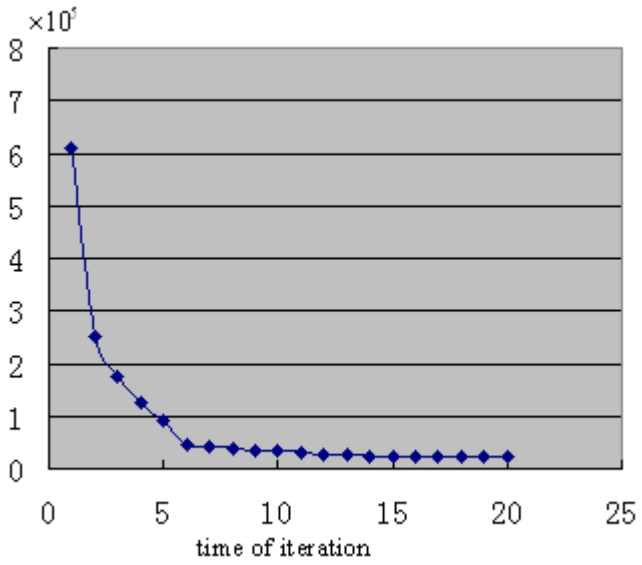


Figure 4.19. Energy minimisation based on 20 iterations

Alternatively, the three situations are combined together. The results are shown as:

Figure 4.20. Synthesis of human body with facial and clothing information from different view directions. The leftmost and rightmost images are recorded as source input images with minor changes in knees. For the computation efficiency, clothing information is set to same colour in black.

Human facial information is integrated with the slightly bent knees and the same colour in clothing. The combination has a processing time of 8–10 minutes in constructing the link distance. Another 7-9 minutes is taken by the processing of BP with the link distance.

4.7 Summary

This chapter presents an approach for complete pixel synthesis of contour-synthesised images. Given the coloured images as the source images, the method uses a robust belief propagation method to obtain contour colouring by involving a link distance. A novel improved application of the distance transform is demonstrated for this purpose. Its ability to transform the pixel information (from source images with the introduced link distance) can be used to allocate the corresponding pixels. Furthermore, it can be transformed into the synthesised images. This method is likely to prove useful in other applications, including Viterbi decoding and optimal control.

The approach developed here directly links image-based human motion understanding, and motion capture with the broader problem of image representations. This is achieved by constructing a link distance to allocate the corresponding pixels in the source images, and synthesised images to understand which pixels are suitable for transforming. Given that this is a largely unexplored area, the research implementation is focused on the human face. However, there certainly remains plenty of scope of the human body for further work. The method, as it is, can be extended to include cluttered human images and dynamic background to code more pixel information [e.g., multi-scale edge histograms are used in AKiranyaz's work (2008)] and environmental impact. The more basic question, however, is regarding the use

of disparity descriptors tied to the construction of the link distance. The four-connected disparity representation has not been approved as the effective method. Nevertheless, given the comparison, a more detailed classification of disparity (eight-connected) representation would be robust to the location of the corresponding pixels for the human body. Moreover, another possibility is to introduce the link distance forms for the distance transform method (Felzabszwalb and Huttenlcher, 2004) to loosely encode the geometry.

5. Conclusion and future work

This thesis has exploited several aspects of image-based rendering, particularly human body image synthesis, achieved by combining methodologies related to computer graphics and machine learning. The thesis has shown that effective human body rendering can be achieved by applying prediction to process contour and colour synthesis that can be realised from small numbers of input images. Beyond the methodology proposed, synthesis has been applied to the human body, providing a framework for future research in the field.

5.1 Key contributions

1) Learning and image-based contour syntheses

For image-based human images with complex human body postures, synthesising contour for outlines proves to be an effective method. Despite loss of information on human body postures, the regression method enables the synthesis of the human body from small numbers of images for object outlines. A total of 300 human motion data frames are obtained, with various kinds of movement as training data sets. Images are created with clear torso differentiation and a static unified background in BVH files using Poser (Curious Labs). Different parameters are used in testing the experiment, which yields reasonable results. These results are considered accurate for colour

synthesis.

2) Pixel synthesis using assigned belief propagation

Because of the variability and invisibility of human posture, human body analysis is extremely complex, especially when synthesising pixels hidden by the outline of contour synthesis. To overcome the corresponding features hidden by the outline of contour synthesis from source images to synthesis images, an assigned belief propagation method is employed to construct a link distance for corresponding pixels. The method shows the correspondence of neighbouring pixels between the input images and synthesised images. To the best of my knowledge, this study is the first to attempt human body synthesis by using a learning and image based method.

3) Fast distance transform for the human body

Fast distance transform is not the first method proposed for synthesis. Despite its lack of efficiency in computation time, the fast distance transform incorporated with the link distance for human body is widely employed. The method can obtain reasonably accurate results compared with those proposed in human facial research.

5.2 Possible future extensions

On the basis of the training data selected, the following fields as directions for future work are proposed:

1) The algorithms developed in this thesis are machine learning and image-based types. Human contour is encoded as a simple vector, and synthesised using regression methods. In accordance with the input on human images and other contour information (i.e. human facial information, clothing information), the regression method for contour synthesis can be extended to include the synthesis of the internal information within the outline. However, this method necessarily disregards a considerable part of the contour representation in the internal space. Therefore, the current methods proposed do not attain the same precision for internals as do outline synthesis. This discrepancy is due in large part to the huge computation time spent in addressing inherent difficulties associated with the variability of internal contour synthesis. Furthermore, the image-based approach is often limited by the range of its training examples. The effective introduction of an explicit and wide range of training example images would allow the generation of accurate information using the learning-based method.

2) The current method employed in synthesising contour is essentially based on SVR. Several other possibilities such as the least square regression

can likewise be explored. Meanwhile, considering statistical methods such as linear or kernel canonical correlation analysis (Lai and Fyfe, 2000) can prove helpful in understanding the vectors. Moreover, providing more reasonable uncertainty measures such as the relevance vector machine is important, particularly for the training database.

3) For complicated human data with numerous ambiguities, the effective construction of pixel link distance is suggested to reduce the computation time. This can be the primary focus of future research. A relatively straightforward extension to the algorithms presented in Chapters 3 and 4 would be the combination of an explicit body model and top-down optimisation (Agarwal, 2004). Such optimisation is computationally expensive, but presents improved accuracy and provides an appropriate likelihood model for understanding the internals.

Reference

Agarwal A and Triggs B (a).（2004）, earning to Track 3D Human motion from Silhouettes. 21st. International Conference of Machine Learning. Canada, Banff, pp. 9-16.

Agarwal A and Triggs B (b). (2006), Recovering 3D Human Pose from Monocular Images." IEEE Transactions on Pattern Analysis and Machine Intelligence vol. 28(1), pp.1-15.

Agarwal A and Triggs B (c). (2004). Learning Methods for Recovering 3D Human Pose from Monocular Images. English version. Institut National Recherche en Informatique et en Automatique. Project LEAR - Learning and Recognition in Vision. Available at http://www.inria.fr/rrrt/rr-5333.html. [Accessed 22 August 2008]

Agarwal A and Triggs B (d). (2006), A Local Basis Representation for Estimating Human Pose from Cluttered Images. 7th Asian Conference on Computer Vision. International Institute of Information Technology. India, Hyderabad, pp. 50–59

Aggarwal, JK and Q Cai (1999), Human Motion Analysis: A Review. Computer Vision and Image Understanding 73, pp. 428-440

Anderson. C, Bert. P and Walvander. G. (1985), Change detection and tracking using Pyramids transformation techniques. Proc SPIE Conference on Intelligent Robots and Computer Vision, Cambridge, MA, 579.pp. 72-78.

Arseneau. S and CooPerstock. J, (1999), Real-time image segmentation for action recognition. Proc IEEE Pacific Rim Conference on Communications，Computers and Signal Processing, Victoria, Canada, pp. 86-89

Ayache. N, and Faverjon. B, (1987), Efficient registration of stereo images by matching graph descriptions of edge segments, international journal of computer vision, pp.101-137

Bartlett. P, and Traskin. M, (2007), AdaBoost is Consistent. The Journal of Machine Learning Research vol.8, pp.2347-2368

Barclay. CD , Cutting. JE, and Kozlowski. LT (1978), Temporal and spatial factors in gait perception that influence gender recognition. Perception Psychophysics, vol. 23(2), pp. 145-152

Barron. J, Fleet. D, and Beauechemin.S. (1994), Performance of optical flow techniques. International Journal of Computer Vision, 12(1).pp. 42-77.

Bellman. R. and Karush .W (1961), On a new functional transform in analysis: the maximum transform, Bull. Amer. Math. Soc. 67, pp. 501-503

Berthod. M, Kato. Z, Yu. S and Zerubia.J (1996), Bayesian Image Classification Using Markov Random Fields. Image and VisionComputing, 14, pp. 285-295

Bertsekas. D (2001), Dynamic Programming and Optimal Control, vol.1 and

vol. 2, pp. 704

Blake. R, and Shiffrar. M (2007), Perception of Human motion. ANNU. Rev. Psychol, vol. 58, pp. 47-73.

Blake, A and Isard, M (1998), Propagation for Visual Tracking. Computer Vision, vol. 29(1), pp. 5-28

Blinn. J. F (1978), Simulation of wrinkled surfaces. Computer Graphics (SIGGRAPH'78), vol.12 (3), pp. 286–292.

Boulfani-Cuisinaud, Y. and Antonini, M (2007), Motion-Based Geometry Compensation for DWT Compression of 3D mesh Sequences. ICIP (1), pp. 217-220

Boykov. Y, Veksler. O, and Zabih. R (1998), Markov Random Fields with Efficient Approximations," IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'98), pp. 648

Boykov. Y., and Kolmogorov. V (2004), An Experimental Comparison of Min-cut/Max-flow Algorithms for Energy Minimization in Vision, IEEE Trans on Pattern Analysis and Machine Intelligence (SO162-8828), 26(9), pp. 1124-1137

Boykov, Y., Veksler, O., and Zabih, R (2001) Fast Approximate Energy Minimization via Graph Cuts", IEEE Trans on Pattern Analysis and Machine Intelligence, pp. 1222-1239

Brain.L and Walton J N, et al（1969） Brain's diseases of the nervous system. Oxford: oxford university press

Bregler. C, and Malik. J (1998) Video motion capture In Proc of SIGGRAPH98, Newyork: ACM press.

Buehler, C., Bosse, M., McMillan, L., Gortler, S., and Cohen, M（2001） Unstructured lumigraph rendering, Proc. Siggraph, pp. 425-432

Chen S. E, and Williams. L (1993) View interpolation for image synthesis, Proc SIGGRAPH' 93, pp. 279-288.

Chen. W (2003), Real-time Ray Casting Rendering of Volume Clipping in Medical Visualization. Journal of Computer Science and Technology, vol.18 (6), pp 804-814

Chen. Y, Zhu. W, Sun.Y, Yin. B, and Jiang. D (2004), SVR-Based Facial Texture Driving for Realistic Expression Synthesis Third International Conference on Image and Graphics (ICIG'04), pp. 456-459.

Collins R (2000), A system for video surveillance and monitoring: VSAM final report. Carnegie Mellon University, Technical report, CMU-RI-TR-00-12

Cohen. I and Hongxia. L (2003), Inference of Human Postures by Classification of 3D Human Body Shape. IEEE International Workshop on Analysis and Modelling of Faces and Gestures. IEEE computer Society. France, Nice.

Cutler.R and Davis.L. (2000) Robust real-time Periodic motion detection, analysis, and applications. IEEE Trans Patten Analysis and Machine Intelligence, 22(8), pp. 781-796.

Deaguiar, E., Stoll, C., Theobalt, C., Ahmed, N., Seidel, H.P., and Thrun, S.,

(2008) Performance capture from sparse multi-view video, ACM Trans Graphics, 27(3), Article 98

Demirdjian. D (2003) Enforcing the constraints for human body tracking. MIT CSAIL Vision Research. MIT Computer Science and Artificial Intelligence Laboratory. Available at http://group.csail.mit.edu/vision/vip/papers/demirdjian_iccv2003.pdf [accessed 03 August 2008]

Deutscher, J., Blake, A., and Reid, I (2000) Articulated body motion capture by annealed particle filtering [A]. Proceedings of Conference on Computer Vision and Pattern Recognition [C]. vol.2, pp.1144～1149.

Dhome, M and Jurie, F (2002) Hyperplane Approximation for Template Matching." Machine Intelligence and Pattern Analysis. vol. 24(7). pp. 996-1000

Elgammal, A. and Lee, C.-S (2008) The role of Manifold learning in Human Motion Analysis" computational imaging and vision, vol. 36, pp. 25-56

Fan, J, EA El-Kwae, M-S Hacid, and F Liang (2002) Novel tracking-based moving object extraction algorithm. J Electron Imaging 11, pp. 393-400

Faugeras. O (1995) Stratification of 3D vision: Projective, affine and metric representations". Journal of the Optical Society of America A, 12(3), pp. 465-484.

Felzenszwalb, P. F., and HuttenIcher, D. P (2004) Distance transforms of sampled functions. IEEE CVPR, pp. 261-268

Felzenszwalb, P. F., and HuttenIcher, D. P (2006) Efficient Belief Propagation for Early Vision. International Journal of Computer Vision, vol. 70, pp. 41-54

Fitzgibbon. A, and Zisserman. A (1998) Automatic 3D model acquisition and generation of new images from video sequences. European Signal Processing Conference, Rhodes, Greece, pp. 311- 326.

Friedman. N and Russell. S (1997) Image segmentation in video sequences: a Probabilistic approach, Proc the Thirteenth Conference on Uncertainty in Artificial Intelligence, Rhode Island, USA.

Gavrila. DM., and Davis. LS (1996) 3-D Model-Based Tracking of Humans in Action: A Multi-View Approach [A]. Proceedings of Conference on Computer Vision and Pattern Recognition. pp. 73-80

Genc. Y, Ponce. J, Leedan. Y, and Meer. P. (1999) Parameterized image varieties and estimation with bilinear constraints". Proc. IEEE Conf. Computer. Vision and Patten. Recognition, Fort Collins, CO, pp. 67- 72.

Gortler. SJ., Grzeszczuk. R, Szeliski. R and Cohen. M (1996) the lumigraph Computer Graphics, Proc. SIGGRAPH 96, pp. 43.

Gouraud. H. (1971) Continuous shading of curved surfaces. IEEE Transactions on Computers, vol.20 (6), pp. 23-628.

Grauman. K, Shakhnarovich. G, and Darrell. T. (2003) Inferring 3D structure with a statistical image-based Shape model. Proceedings of the IEEE international conference on computer vision, pp. 641-647

Greig D.M., Porteous B.T. and Seheult A.H (1989) Exact maximum a posteriori estimation for binary images. Journal of the Royal Statistical Society Series B, 51, pp. 271–279.

Grest. D, and Koch. R. (2005) Multi-camera person tracking in a cluttered interaction environment. Computer analysis of images and patterns. International conference N°11, Versailles, France.

Guo. F and Qian G. (2008) Monocular 3D Tracking of Articulated Human Motion in Silhouette and Pose Manifolds. Journal on Image and Video Processing vol. 08(04), pp. 1687-5176

Haritoaglu, Harwood. D, and Davis. L.S. (2000) W4:Real-time surveillance of people and their activities, IEEE Trans, on Patten Analysis and Machine Intelligence, vol. 22,no.7,pp. 809-830

Hartley. R, Gupta. R, and Chang. T. (1992) Stereo from uncalibrated cameras. Proc. IEEE Conf. Computer Vision Patten Recognition, pp. 761-764.

Hill DR, Pearce A and Wyvill B. (1987) Animating Speech: an Automated Approach Using Speech Synthesised by Rules. The Visual Computer, vol.3, pp. 23

Hu. Z, He. Y, and Ou. Z. (2006) A New IBR Approach Based on View Synthesis for Virtual Environment Rendering, 16th International Conference on Artificial Reality and Telexistence--Workshops (ICAT'06), pp. 31-35,

Joachims, T. (1999) Learning Practical in Making large-Scale SVM. Support Vector Learning. Boston, MIT Press

Jordan, M and Jacobs, R. (1991) Local Experts & Adaptive Mixtures. Neural Computation, vol. 3, No. 1, pp. 79–87

Joshi. N., Matusik, W., and Avidan, S. (2006) Natural video matting using camera arrays, ACM Trans Graphics, 25(3), pp. 779-786

Ju, Shanon X. Black, Michael J. and Yacoob, Yaser. (1996) Cardboard people: A parameterized model of articulated image motion. In Proc. Gesture Recognition, pp. 38-44

Jürgen.Gall. (2006) Learning a Dynamic Independent Pose Distribution within a Bayesian Framework. Human Motion - Understanding, Modeling, Capture and Animation. 13th Workshop "Theoretical Foundations of Computer Vision", IBFI Schloss Dagstuhl, Germany

Kakadiaris I and Metaxas D. (2000) Model-Based Estimation for 3D Human Motion. IEEE Trans on Pattern Analysis and Machine Intelligence vol. 22(12), pp. 1453-1459

Kang, S.B., Wu, M., Li, Y., and Shum, H. Y. (2003) Large environment rendering using plenoptic primitives. IEEE Trans on Circuits and Systems for Video Technology, 13(11). pp. 1064-1073.

Kilger M. (1992) A shadow handler in a video-based real-time traffic monitoring system. Proc IEEE Work shop on Applications of Computer Vision， Palm Springs, CA, pp.1060-1066.

Klein. K，Malerczyk. C，Wiebesiek. T，and Wingbermuhle. J. (2002) Creating a Personalized Immersive sports TV experience via 3D reconstruction

of moving athletes. W. Abramowicz. (ed.), Business Information Systems, Proceedings of BIS, Poznan

Kolmogorov, V., and Zabin, R. (2004) What energy functions can be minimized via graph cuts? IEEE Trans on In Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 26, No. 2, pp. 147-159

Krahnstoever. N, Yeasin. M, and Sharma. R. (2003) Automatic acquisition and initialization of articulated models, Machine Vision and Applications, vol.14 (4), pp. 218-228

Kuno.Y， Watanabe.T， Shimosakoda.Y and Nakagawa.S. (1996) Automated detection of human for visual surveillance system. Proc IEEE International Conference on Patten Recognition, Vienna , pp. 865-869.

Lee, M W and Cohen, I (a). (2006) A model-Based Approach for Estimating Human 3D poses in Static Images. IEEE Transactions on Pattern Analysis and Machine Intelligence vol. 28(6), pp. 905-916.

Lee, M W and Cohen, I (b). (2004) Human Upper Body Pose Estimation in Static Images. Computer Vision - ECCV 2004, 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14. Proceedings, Part II

Leung. MK and Yang. YH. (1995) First Sight: A Human Body Outline Labelling System.IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 17(4), pp. 72

Leventon, M and Howe, N. (1999) Single-Camera Video Reconstruction of 3D Human Motion. Neural Information Processing Systems.

Levoy. M, and Hanrahan. P. (1996) Light field rendering, SIGGRAPH 96 Conf. Proc. Annual Conf. Series, pp. 31-42

Lewis JP, and Parke FI. (1987) Automated Lip-synch and Speech Synthesis for Character Animation, Proc. Human Factors in Computing Systems and Graphics Interface '87, Toronto, pp. 143-147.

Li. Y, Sun. J, and Shum. H.Y. (2005) Video object cut and paste. ACM Trans. Graph. vol. 24(3), pp. 595-600

Lipton. AJ. (1999) Local application of optical flow to analyse rigid versus non-rigid motion· In ICCV99 Workshop on Frame-Rate Applications,

Lipton. AJ, Fujiyoshi. H and Patil. R. (1998) Moving target classification and tracking from real-time video. Proc IEEE Work shop on Applications of Computer Vision. pp. 8-14.

Lorensen. W. and Cline. H. (1987) Marching cubes: A high resolution 3D surface construction algorithm. Computer Graphics (SIGGRAPH'92), vol. 21(4), pp. 163–169.

Lowe, D. (1999) Local Scale-invariant Features Object Recognition. Computer Vision, pp. 1150-57,

Malik, J and Mori, A. (2002) Using Shape Context Matching for Estimating Human Body Configurations. Computer Vision, vol. 3, NO. 56, pp. 666-680

Masi, CG. (2006) Vision improves bat performance. Vision Systems Design.

June, available at http://www.vision-systems.com, [Accessed 28 August, 2008]

McLachlan.G and Krishnan.T. (1997) The EM algorithm and Extensions .Wiley Interscience, Hoboken.

McMillan. L, and Bishop. G. (1995) Plenoptic Modeling: An Image-Based Rendering System. SIGGRAPH, pp. 39-46

Meyer. D, Denzler. J and Niemann.H. (1997) Model based extraction of articulated objects in image sequences for gait analysis. Proc IEEE International Conference Image Processing, pp. 78-81.

Müller, M., Röder, T., and Clausen, M. (2005) Efficient Content-Based Retrieval of Motion Capture Data". ACM Trans. Graph., vol. 24(3), pp. 677-685

Parke F.I. (1972) Animation of Faces, Proc. ACM Annual Conf., vol.1

Parke F.I. (1982) Parameterized Models for Facial Animation, IEEE Computer Graphics and Applications, vol.2, No9, pp. 61-68.

Pearl, J. (1988) Probabilistic reasoning in intelligent Systems. Networks of plausible inference. San Francisco: Morgan Kaufmann, pp. 286–292

Pedro F. Felzenszwalb and Daniel P. Huttenlocher (2006) Efficient Belief Propagation for Early Vision. International Journal of Computer Vision archive. Vol. 70 (1), pp. 41-54.

Phong.B. (1975) Illumination for computer-generated pictures. Communications of the ACM, vol.18 (6), pp. 311–317.

Qu. H, Wan. M, Qin. J, and Kaufman. A. (2000) Image Based Rendering with Stable Frame Rates. Proceedings of the conference on Visualization, pp. 251-258

Rabiner .L. R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc. of the IEEE, vol.77 (2), pp. 257-286..

Rosales. R, Siddiqui. M, Alon. J, and Sclaroff. S. (2001) Estimating 3D Body Pose using Uncalibrated Cameras". CVPR (1), pp. 821-827

R'omer. R, Athitose. V, Sigal. L and Sclaroff. S. (2001) 3D Hand Pose Reconstruction Using Specialised Mappings. IEEE international Conf. on Computer Vision (ICCV). IEEE Computer Society. Canada, Vancouver, pp. 378-387

Scharstein.D. (1999) View Synthesis Using Stereo Vision. Conference on Computer Vision and Pattern Recognition, pp. 852

Seitz, S.M., and Dyer, C.M.（1996） Viewing morphing, Proc. Siggraph, pp. 21-30

Selinger. A and Wixson. L. (1998) Classifying moving objects as rigid or non-rigid without Correspondences, Proc DAPRA Image Understanding Workshop, Monterey, CA1, pp. 341-358.

Shakhnarovich, G and Grauman, K. (2003) Statistical Image Based Shaped Model. Computer Vision, pp. 641-48

Shum, H.Y., Chan, S.C., and Kang, S.B. (2006) Feature based light field morphing in Image Based Rendering. Springer, Berlin, Chapter 17

Shum, H. Y., Sun, J., Yamazaki, S., Li. Y., and Tang, C.K. (2004) Pop-up light field: An interactive image based modelling and rendering system, ACM Trans on Graphics, 23(2), pp. 143-162

Sidenbladh. H, Black. M， and Sigal. L. (2002) Implicit Probabilistic models of human motion of synthesis and tracking. European Conference on Computer Vision, ECCV, pp. 784-800

Silaghi. M.C, Plänkers. R, Boulic. R, Fua. P and Thalmann. D. (1998) Local and Global Skeleton Fitting Techniques for Optical Motion Capture. Lecture Notes In Computer Science; vol. 1537, pp. 26-40

Sminchisescu. C, Kanaujia. A and Metaxas. D. (2006) Learning Joint Top-Down and Bottom-up Processes for 3D Visual Inference. IEEE Computer Conf. on Computer Vision and Pattern Recognition. IEEE Computer Society. USA, New York.

Sminchisescu. C, Triggs. B.Covariance. (2001) Scaled Sampling for Monocular 3D Body Tracking[A].Proceedings of the Conference on Computer Vision and Pattern Recognition[C], vol. 1, pp. 447-454.

Sparks CE, Hinrichsen RL, Friedmann D. (2005) Comparison and Validation of Smooth Particle Hydrodynamics (SPH) and Coupled Euler Lagrange (CEL) Techniques for Modeling Hydrodynamic Ram. 46th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference, Austin, Texas.

Stauffer. C. (1999) Automatic hierarchical classification using time-based co-occurrences. Proc of IEEE CS Conference on Computer Vision and Patten Recognition, pp. 333-339.

Stauffer. C and Grimson. W.E.L. (1999) Adaptive background mixture models for real- time tracking. IEEE Conf. On Computer Vision and Patten Recognition, pp. 246-252.

Stringa.E. (2000) Morphological change detection algorithm for surveillance applications. British Machine Vision Conference, Bristol, UK, pp. 402-411

Sun. J, Zheng. NN and Shum. HY. (2003) Stereo Matching Using Belief Propagation. Pattern Analysis and Machine Intelligence, vol. 25, pp. 7

Szeliski.R and Scharstein.D. (2002) Symmetric subpixel stereo matching. Seventh European Conference on Computer Vision, Copenhagen, Denmark, vol. 2, pp. 525-540.

Tappen. M and Freeman.W. (2003) Comparison of Graph Cuts with Belief Propagation for Stereo, using Identical MRF Parameters. IEEE International Conference on Computer Vision (ICCV), pp. 900 - 907

Thayananthan, A and Stenger, B. (2003) Tree Based Estimators. Computer Vision, Cambridge, UK, pp. 343–356

Triggs, B and Sminchisescu, C. (2001) Monocular 3D Body Tracking. IEEE International Conference on Computer Vision and Pattern Recognition, vol.1, pp. 447-454

Troje. N F. (2002) Decomposing biological motion: A framework for analysis

and synthesis of human gait patterns. Journal of vision, vol. 2, pp. 371-387

Vapnik .V (1995) The Nature of Statistical Learning Theory. Springer.-Verlag, New York, pp. 332

Verri.A, Uras.S and DeMieheli.E. (1989) Motion Segmentation for optical flow, Proc the 5th Alvey Vision Conference, Brighton, UK, pp. 209-214.

Walczack. B, and Massart. D L. (2000) Local modelling with radial basis function networks. Chemometrics and Intelligent Laboratory Systems 50, pp. 179–198

Wang. W, Xu. Z, Lu. W, and Zhang. X. (2003) Determination of the spread parameter in the Gaussian kernel for classification and regression. Neuro computing, vol.55, pp. 3-4

Wei, L.Y., and Levoy, M. (2000) Fast Texture Synthesis using Tree-structured Vector Quantization, Proc. Siggraph, pp. 479-488

Weiss .Y and Freeman .W. T. (2001) On the optimality of solutions of the max-product belief propagation algorithm in arbitrary graphs. IEEE Trans. Information Theory, vol.47, pp. 723–735.

Welch.G and Foxlin. E. (2002) Motion tracking: no silver bullet, but respectable arsenal.” IEEE Computer Graphics and Applications vol.23 (1), pp. 24-38

Wren. CT, Azarbayejani. A， Darrell. T and Pentland. A. (1997) Pfinder: Real-time Tracking of Human Body, IEEE Trans. On Patten Analysis and Machine Intelligence，vol.19(7), pp. 780-785.

Yoon. K.J and Kweon. I.S. (2005) Locally Adaptive Support-Weight Approach for Visual Correspondence Search," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. 924-931.

Yunpeng Li, and Huttenlocher, D.P. (2008) Learning for stereo vision using the structured support vector machine. IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008. , Anchorage, AK, pp. 1-8

Zhang. Z and Troje. N F. (2007) 3D Periodic Human Motion Reconstruction from 2D Motion Sequences. Neural Computation, vol.19, pp. 1400-1421