Image-based Human Pose Estimation

Dongxu Gao

School of Computing

University of Portsmouth

A thesis submitted for the degree of

Doctor of Philosophy

November 2017

Whilst registered as a candidate for the above degree, I have not been registered for any other research award. The results and conclusions embodied in this thesis are the work of the named candidate and have not been submitted for any other academic award.

Dedicated to my father Chengfu Gao, my mother Shulan Guo, my sister Liang Gao and my wife Xiaohua Cui, with love.

Acknowledgements

This work would not have been possible without the guidance and support of many people. I would like to thank my first supervisor, Dr Zhaojie Ju, for his great guidance and inspiration and teaching me the importance of scientific research. I owe a huge debt of gratitude to him for many hours spent proofreading paper submissions at the last minute and this thesis. Fortunately, the whole process with him was an enjoyable experience. It is such a shame that I can not convey my gratitude with unlimited expressions of all thankful words. Further thanks to my second supervisor, Prof. Honghai Liu, for his thorough support during my whole PhD period. I am also grateful to my previous supervisor, Prof. Jiangtao Cao, for introducing me to pursue my PhD degree.

I really appreciate all the help I have received from the Intelligent Systems and Biomedical Robotics (ISR) Group members-particularly Dr. Yinfeng Fang, Mr. Haibin Cai, Mr. Dalin Zhou, Mr. Yiming Wang, Ms. Kairu Li, Ms. Bangli Liu, Mr. Charles Phiri, Mr. Uchenna Ogenyi, Mr. Peter Boyd, Mr Disi Chen, Mr. Tong Cui, all the visiting scholars and students. My study at Portsmouth would have been far less fruitful without all members in the ISR group. Research can sometimes be difficult, but my life here is far more creative and enjoyable with all of ISR members.

I also would like to express my sincere gratitude to the Chinese CSC funding, FP7 project funding and University of Portsmouth funding for the financial support, my study would not have been possible without funding.

I am also grateful to my parents, for always being there when I needed help and supporting my move to the UK for my PhD. Finally, special thanks to my wife for making my life better.

Abstract

Human pose estimation has become an active research topic in the field of computer vision. However, there are still some technical challenges because of the complexity of human motion. Although the depth sensors, such as Kinect and Xtion, open up new possibilities of handling with issues, they present some new challenges. In this thesis, we only address human pose estimation frameworks based on colour image and explore the possibility of the tradeoff between effective representing features and models.

Firstly, the task of human pose estimation can be treated as a regression model. So we propose a novel method based on the regression model, which is designed for estimating the upper joints and recognizing their special motions. We verified the proposed method on our recorded dataset and the experimental results show the proposed method is effective. This provides an important clue that the performance of human joints estimation contributes significantly for human motion estimation.

Secondly, the computation problems are always making it difficult for computer vision. For example, the pictorial structures normally use the interactions between connected joints such as elbow and shoulder, leading to a quadratic computation cost in the number of pixels for the inference process. Then a simple model for restricting themselves is proposed, which only measure the quality of limb-pair possibilities. Meanwhile, it allows the efficient inference in richer models, which exploit the data-dependent interactions.

Thirdly, to improve the effectiveness of the body pose estimation, we introduce a object tracking method to the body pose estimation process. In addition, we introduce structured prediction aggregate model, which only

need to focus on necessary computational effort. It can ensure the accurate output by filtering out many states cheaply. Meanwhile, our proposed decomposition method use cyclic dependencies on a tree model when imposing the model agreement. Thus it allows for efficient inference on a video or an image.

To sum up, we evaluate our proposed methods on public datasets and compare them with some popular methods to demonstrate both the efficiency and effectiveness. The model pairwise interaction potentials are afforded with data-dependent features and the aggregate model. The experimental results show that our model is worthwhile and features used are accurate for pose estimation on popular datasets.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

People are shocked by the recent novel technology such as big data, image processing and machine learning, achieved by related researchers spending most of their time in these areas. The general goal of computer vision enabless the computer to have the ability in the way of human to see and know the world around us. To research such a goal a machine must use a camera to sense the world and machine learning algorithm is employed to make sense of such world for different purposes. Recent human pose estimation algorithms have achieved promising performance with the help of the calibrated cameras in front of a clean, static background (Pons-Moll *et al.*, 2011; Rogez *et al.*, 2012). However, there are few studies on human pose estimation in unconstrained environments. Previous traditional pose estimation algorithms often rely on low-level appearance features, such as silhouette and optimal flow Bissacco *et al.* (2007); Ionescu *et al.* (2011); Navaratnam *et al.* (2006); Rogez *et al.* (2012). Experimental results have shown that these features are vulnerable to cluttered backgrounds, dynamic scenes and moving viewpoints in unconstrained videos. Therefore, the aim of this thesis is to explore robust features with balanced human body model and improve the performance of the human pose estimation on a wide range of scenario.

When we refer to the human pose estimation, it should be noted that related works such as human detection and tracking sometimes contribute the task of articulated hu-

man pose recognition. This thesis would mainly focus on the development of a human model and offer further insight into the task of estimating the pose of human with the prior knowledge of human detection and motion tracking. Humans can be considered a collection of related objects (body parts), or a single but highly deformable object. The parts themselves are the most difficult to detect in the literature. Typical objects that researchers work on to recognise faces, bicycles or even potted plants have distinguishing features, reliable patterns and limited intra-class variability. A body part such as a lower arm, on the other hand, is far more generic. It has a generic shape, at best it can be described as a projection of a cylinder or frustum is subject to much higher intra-class variability due to clothing, articulated pose, body type, and severe foreshortening. Features developed must be invariant to pose, lighting, texture and colour and still discriminate parts from clutter, or efficient searching procedures over these variations need to be developed.

Lots of challenges put the burden on the computer vision researchers such as lighting condition, viewpoint, occlusion, clothing, and background clutter. When it refers to a practical application, such as the human pose estimation and the hand pose recovery, more complex challenges make it extremely difficult to estimate the human body pose or the hand pose, because they share all the difficulties of computer vision issues as well as their own difficulties, for example, the appearance of pose is largely uncertain, and the scale or the pose in different camera angle makes it highly variable with uncountable appearance modes. The computation difficulty is also essential to be handled specially in practical application.

1.2 Problems and challenges

The human pose estimation can be represented as: suppose x is the input image pixels, and y is a representation of the predicted pose which is the output. Then the above assumption can be formed as a scoring function f(x, y), which can evaluate the quality of any estimated pose y in the image x. Therefore, the solution of the assumption will provide us with the final result of the human pose estimation.

If the best pose is defined as the highest scoring, then y will be an infinite dimensional for continuous input. So the above problem could be dealt only with the condition that the determination of the maximiser can be confirmed in polynomial

time. Therefore, there are two sources of intrinsic computational complexity within the human pose estimation framework.

The complexity of the input: any given input x can map to the same pose y, which means the same body can look different for a different image. To handle the challenging issue, it seems there are only two options for us to use. The first one is to design features that are invariant to the model. The second one is to partition the space and model using separate model modes. A generic patch-based joint detector based on coarse edges could be one of the solutions for former approach. For the latter approach, the other difficult decisions are: it is difficult to get a definition of modes and a notion of joint position. In addition, it is extremely difficult to balance the richness of the model and the model fitting errors when the training data is finite.

The complexity of the output: The possible output poses can be enormous, which obviously increases computational complexity. In addition, the part interactions are computationally considerable because of the lack of discriminating features and the wide range of appearances. In another word, enumerating all possible joint pose configurations and estimating parts in isolation are extremely difficult. Then the pairwise model is a good option and the graph of part interactions can be formed to a tree model with regard to the part interactions at a time. In such a pairwise model, one of the challenging operation is to evaluate the quality of a pair of parts. So deciding the optimal global pose would combine all such pairwise scores together.

1.3 Overview of approaches and challenges

This thesis provides an explicit scheme of the pose estimation from image and video data. There are three main contributions for the whole thesis.

Firstly, one of the important contributions is that we provide an end-to-end process for the activity recognition with the advantage of the pose estimation. This pose estimation method uses only the colour image as the input, a deformable mixture-of-parts model is used to represent the body parts with the computational efficiency, and the upper body part is modelled as major joints set. The proposed clustering method can classify each body part with annotated ground truth and avoid the self-occlusion situation. This is because the maximum value of the formulation allows us to determine the appearances mode with the highest confidence with respect to the posterior probability,

which can overcome the ambiguous image data. The introduction of the random sampling strategy can efficiently decrease the complexity of the feature patches. Overall, with the benefit of the effective representation of joints information, even the classic SVM classifier can output a satisfied result.

Secondly, because of the computational issues discussed above, we introduce an improved spring model (pictorial structure model), which considers pairwise interactions between parts when resorting to a pose model in a restricted form. It is important to trade-off the individual score at any location for placement when deforming the default model positions. For example, the deformation penalty between an upper shoulder and wrist expresses the fact that they should about agree on the location of the elbow. An important property of this model is that the terms, pairwise and spring stretch are blind to the image content. So the individual part detector scores are extremely weak. Specifically, in all settings of environment, articulation, background and foreground, all these scores are isolated, so as the generalized limbs.

Thirdly, we introduce a tracking method to the pose estimation, to make the estimation process smoothly. The aggregate model we proposed in works for any tree-structured model. However, it is difficult to capture important interactions between frames when dealing with multiple parts tracking over time in a video. In addition, the part relationship, known to be exponential in the number and the union of edges, covers all relationships. We introduce an approximate approach for determining the best possible argmax answer over a graph of parts relationship by decomposing a cyclic model of pose to a tree sub-graph set. Thanks to the proposed aggregate model, all the interesting interaction terms in the model can be exploited in the original model with efficient inference. Moreover, all cues utilized in the model can be exploited to the benefit of colour symmetry across the body, location persistence information and temporal appearance.

1.4 Overview of thesis

The rest chapters are organised as follows:

Chapter 2 first reviews the related human pose estimation methods, then the modelling of the human body and features for representing human body are reviewed to give an overview of related research methodologies. These demonstrate the general

purpose of human pose estimation and provide a systemic understanding of the current development in image-based human pose features and modelling to the readers.

Chapter 3 provides an end-to-end human pose estimation framework. A novel recognition method is proposed, which is designed especially for estimating the upper body motion. The experimental results confirm the pose estimation method is effective and contributes the motion recognition method significantly.

Chapter 4 addresses one of the sub-problems of human pose estimation, this chapter mainly focuses on the features of the human pose. We introduce a novel pictorial structure feature for representing the human pose, as it can learn the pictorial structure even the pose resolution is increasing while the pose state space remains the same thus reducing the complexity of the model.

Chapter 5 addresses one of the sub-problems of human pose estimation. This chapter first introduces the object tracking method and then applies it to the human pose model to smooth the human pose results. In addition, we improve the performance by introducing an aggregate model. Moreover, we detail the inference and parameters of the proposed algorithm.

Chapter 6 provides the detail information of the used dataset and the implementation of the proposed methods (chapter 4 and 5). The experimental results are illustrated and evaluated for both single frames and video pose estimation. In addition, we provide a thorough discussion on the selection of human feature and model. Finally, the algorithm performance is presented in terms of comparison among some popular human pose estimation methods.

Chapter 7 provides a conclusion of the whole thesis and provides a direction for the future work to improve the current human pose estimation further. Then the current limitations, are still believed to be difficult to handle, are analysed. Finally, we provide a brief discussion of pose models on other problems such as the computational complexity issues and the weakness of the structured model.

Chapter 2

Literature Review

2.1 Human Pose Estimation

The human activities are complex and difficult to be researched on all aspects, according to the complexity of human activities, they could be categorized into four, which include gesture, actions, interactions, and group activities. Gestures are the atomic movement and with a meaningful expression of human motion. Actions are referred only one people activity which could be composed of multiple gestures such as waving and drinking.

The human pose estimation is a combination of a wide range of subjects and aspects. Many methods and frameworks of human pose estimation have been proposed. There are also some review papers tried to summarise the proposed method, but it is unrealistic that all aspects are described in only one paper. Aggarwal et al. (Aggarwal & Xia, 2014) has done a great work on grouping different methods and comparing them in a productive way. Basically, the previous methods can be seen as single-layered approaches and hierarchical approaches based on the complexity of the modelling. On one hand, the single-layered approaches consist of space-time approaches and sequential approaches. On the other hand, the hierarchical approaches are a set of methods including statistical, syntactic and description-based methods.

There are many review papers on the human motion analysis. Ramanathan et al. (Ramanathan *et al.*, 2014) provided an overview of the existing human action recognition methods on challenges and robustness when handling these challenges. (Aggarwal & Xia, 2014) mainly review the human activity recognition from 3D data, Vrigka et al.

(Vrigkas *et al.*, 2015) analysed advantages and limitations of human activity methodologies and discussed multi-modal feature fusion method after comparing uni-modal and multi-modal methods. Ziaeefard et al. (Ziaeefard & Bergevin, 2015) provided an overview of semantic human activity recognition, which makes the recognition task more reliable. Therefore, it is required that more research on semantic action recognition is urgently needed for a better high-level human activity recognition. Many potential applications (such as video surveillance, video surveillance and video retrieval) cannot come true without a good human action recognition result. It is still far away from the perfect, despite many encouraging improvements have been achieved in the activity recognition.

Chen et al. (Chen *et al.*, 2013a) summarized the human motion analysis from depth data and believed the use of depth camera could simplify tasks such as background subtraction and illumination changes. However, it will be difficult to understand the image information without the colour information. In general, the depth cameras produce better quality 3D motion. To get an invariant feature for 3D joint positions, Wang (Wang *et al.*, 2012b) proposed a novel feature for representing human motion from depth information. Xiao (Xiao *et al.*, 2016), use deep learning to detect the human in a real-world scenario based on bounding boxes annotations.

Model is believed very important and it is proved that a good model can outperform the existing method. Zheng (Zheng *et al.*, 2016) addressed unconstrained video and multiple action instances in real applications, then action temporal localization usually considers temporal overlap and achieves high localization accuracy. (Wei *et al.*, 2013).(Hu *et al.*, 2016) used soft labels and go beyond accurate label restrictions so that the method could allow labels to be incomplete and uncertain. To represent the body, a cylinder-based model is utilised (Sigalas *et al.*, 2016) to extract body pose and tracking in RGB-D sequences.

Modelling 4D human-object interactions lie in three main tasks in vision simultaneously: segmentation, recognition and object localization. (Zhu & Lucey, 2015) presented practical 3D reconstruction results with trajectory basis Non-rigid Structure from skeleton information of just 2D projected trajectories. (Liu *et al.*, 2017) focused on joint human action grouping and recognition with using a hierarchical clustering multi-task learning method. (Yang & Tian, 2017) proposed a novel framework for recognizing human activities with depth information and achieved superior performance

on some public benchmark dataset including MSRAction3D, MSRDailyActivity3D, MSRGesture3D, and MSRActionPairs3D. Multi-layer Dynamic Bayesian Network is used to model the extracted discriminative features (Roudposhti *et al.*, 2016).

The time-related pose is also important. (Taylor *et al.*, 2007) proposed an undirected model with binary latent variables and real-valued visible variables for representing joint angles. The model can find a set of parameters for several different motions. Motion-based patterns (Ben-Arie *et al.*, 2002) used a multidimensional indexing method for recognising human activity. This view-based recognition method can identify the activity with just a few frames.(YU *et al.*, 2012) focused on the structure of interest points when using the spatio-temporal implicit shape model for predicting human activities. With the help of the multi-class balanced random forest, both the memory and computational cost could be saved simultaneously for multiple classes.(Wang *et al.*, 2012a) used random occupancy patterns to make the 3D action recognition robust, semi-local features are employed to deal with noise and occlusion. In addition, the random occupancy pattern features are robustly encoded with a sparse coding approach. (Taylor *et al.*, 2007) demonstrated that their model can effectively learn the transitions between different styles of motion.

Zhou et al.(Zhou *et al.*, 2016) estimated 3D full-body human pose estimation with only a monocular image sequence. The method can handle both cases of the locations of the human joints are provided or unknown. If the image locations are unknown, the image locations of the joints are treated as latent variables when integrating a sparsity-driven 3D geometric prior and temporal smoothness. Rafi et al. (Rafi *et al.*, 2016) proposed an efficient deep network architecture that is trained efficiently with a transparent procedure and exploits the best available ingredients from deep learning with a low computational budget. The network is trained only on the same dataset without pre-training and achieves impressive performance on popular benchmarks in human pose estimation.

As mentioned in the previous section, the integration of action and pose is beneficial for both action recognition and pose estimation tasks. While there exist some algorithms that recognise actions from pose estimation or structural constraints (Raja *et al.*, 2011; Yu *et al.*, 2010). The opposite direction, pose estimation from human action, is still a relatively new area.

Regarding the idea of pose estimation from action recognition, Yao (Yao *et al.*, 2012) used action recognition to assist achieving a multi-view 3D Human Pose Estimation (HPE) algorithm. An action class contains rich information of the spatio-temporal structure of the testing data. For example, when a "walking" action is detected from an input video, the subsequent pose estimator then constricts the possible output space to walking pose. Separate regression models were trained for each action class in the training data. Action classification was used to select the corresponding regression model that estimates the output of 3D poses. However, the above approach did not consider the temporal structure of an action. Action classification was performed on a frame-by-frame basis, class labels were only used as an indicator variable of pose estimator for each independent frame.

As a result, the proposed system seeks to investigate the feasibility of applying action detection to facilitate 3D HPE in monocular videos, particularly in an uncontrolled setting. Besides model selection via action classification, action detection forest also leverage spatial and temporal structure of actions, inferring a probabilistic pose estimate using a Hough voting scheme.

2.2 3D pose estimation

Holistic shapes and silhouettes, in particular, are common features for 3D pose estimation. Recent approaches achieve excellent performance by combining holistic shape features with new features or improved optimisation constraints. For instance, Agarwal and Triggs (Agarwal & Triggs, 2006) encoded foreground silhouettes using shape-context descriptor and estimated their corresponding poses with the sparse kernel-based regression methods. Bissacco *et al.* (2007) proposed a boosting classifier to compute human poses from both silhouettes and motion features. Andriluka used a deformable pedestrian detector to cover 3D walking poses in a cluttered street scene. Jiang (2011) presented consistent max-covering, which maximises the overlapping area of a projected 3D pose configuration and an input silhouette. A latent structured model is described by Ionescu (Ionescu *et al.*, 2011) to estimate 3D poses from silhouettes. Motion templates are also used in the 3D pose estimation, Rogez (Rogez

et al., 2012) used a global motion template to recognise poses using a tree-shape ensemble of rejectors. Pons-Moll (Pons-Moll *et al.*, 2011) presented a pose optimisation algorithm from silhouettes captured from multiple cameras.

On the other hand, thanks to the introduction of affordable depth sensors, Kinect, 2.5D depth images have emerged as a new direction for 3D HPE. Given the 2.5D information, foreground object segmentation is straightforward in a single depth image by adaptively thresholding pixel values. For instance, Zhu *et al.* (2008) presented an upper-body pose estimation method from sequences of depth images using visual tracking and inverse kinematics. Baak *et al.* (2011) proposed a data-driven algorithm for real-time 3D pose estimation. A variant of Dijkstra's algorithm was introduced to extract holistic pose features efficiently from depth images, and such features were combined with local estimation using the Hausdorff distance. Ye *et al.* (2011) matched a single depth image with a set of pre-computed motion exemplars to estimate a holistic body configuration. The initial result was subsequently refined by fitting the pose configuration back to the testing depth image. Sun *et al.* (2012b) estimated 3D pose configurations from depth image patches using regression forests. Using a similar regression forest algorithm, Taylor (Taylor *et al.*, 2012) performed the 3D HPE by computing dense correspondences from an input depth image to a deformable 3D articulated model.

However, the acquisition of depth data is one of the major limitations of the above 3D HPE approaches. They either require specialised hardware or some calibrated stereo cameras to capture depth images. Additionally, depth images still can not handle occlusions and are sensitive to noise.

2.3 Representation of human model

Similar to the object recognition methods, which are normally local feature based with the advantages of robustness to occlusion and translation, pose estimation methods, such as bag-of-words (Fei-Fei & Perona, 2005; Sivic *et al.*, 2005) and topic models (Fergus *et al.*, 2005), can also benefit from the categorical distribution of appearance features. On the contrary, part-based models, which were articulated by Fischler, Elschlager (Fischler & Elschlager, 1973) and Marr (Marr, 1982) as a collection of movable templates or shape primitives and were defined within a reference object frame. In

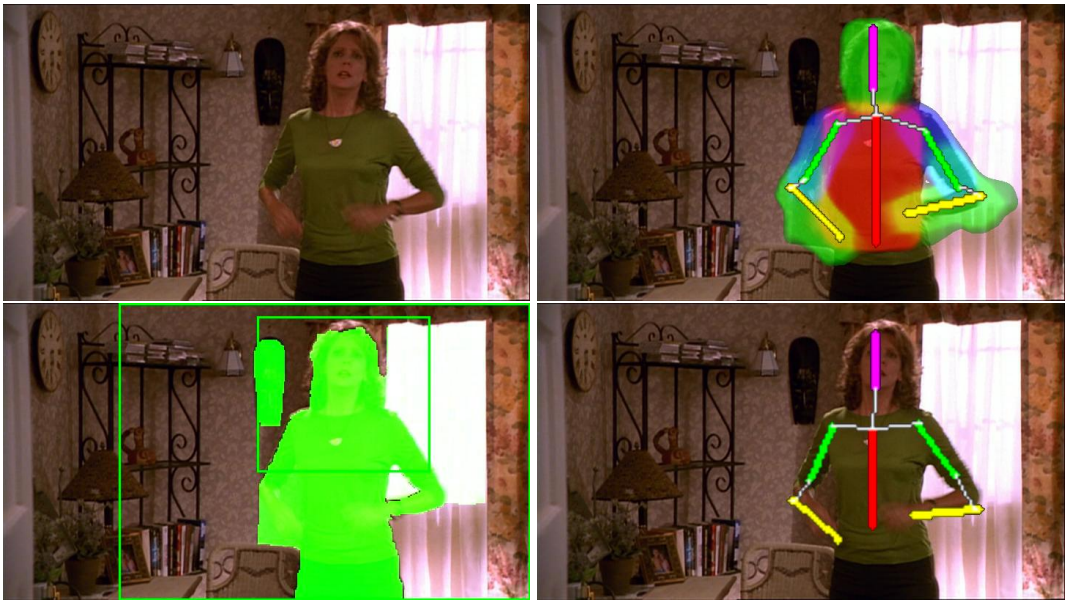


Figure 2.1: Human pose demo figure

addition, notable part-based models normally use constellations (Fergus *et al.*, 2007; Weber *et al.*, 2000) and pictorial structures (or deformable part models with geometric structures) (Felzenszwalb & Huttenlocher, 2005). Meanwhile, geometric structures (i.e. shape) of an object is either explained probabilistically by a model with fixed part positions (Fergus *et al.*, 2007), or discriminatively by movable parts (Felzenszwalb & Huttenlocher, 2005; Yuille *et al.*, 1989).

Early implementations of part-based models were class-specific, where parts arrangements and connections are defined before training. Meanwhile, some part models remain class-specific for deformable human pose estimation(Eichner *et al.*, 2012; Yang & Ramanan, 2011b). Recent constellation models have become more general. For instance, Burl (Burl *et al.*, 1998) trained a constellation model from hand-picked parts which were represented by Gaussian distributions. Weber (Weber *et al.*, 2000) improved the former approach by using an interesting point detector (Kadir & Brady, 2001) to learn feature words and approximate the feature-part assignment marginal. Fergus (Fergus *et al.*, 2007) introduced scale-invariance to the model and learned shape and appearance models jointly. Most recently, a constellation model was also applied in 3D shape recognition (Prasad *et al.*, 2011). Yuille (Yuille *et al.*, 1989) detected facial features using deformable templates.

Vote-based methods, Hough transforms (Barinova *et al.*, 2010; Woodford *et al.*, 2013), implicit shape models (Leibe *et al.*, 2008) and contour fragments (Shotton *et al.*, 2008a), which differ from constellation models in each feature vote, which consists of a fixed object pose and class. In vote-based methods, the internal object representation tends to be non-parametric, which bases on a codebook of appearance features and k-means clustering. There is no explicit model for background clutter, where false positives are discarded through a majority voting process. Such characteristics of vote-based methods allow much faster inference at the test stage. They have also been applied in image-based (Barinova *et al.*, 2010; Leibe *et al.*, 2008; Shotton *et al.*, 2008b) and 3D shape classification (Flitton *et al.*, 2010; Pham *et al.*, 2011; Woodford *et al.*, 2013).

In addition, each codeword in an implicit shape model contains a vote vector that points to the reference point of a target object (Leibe *et al.*, 2008). Target objects are detected by finding the local maxima in the Hough space. Barinova (Barinova *et al.*, 2010) redeveloped the traditional Hough transform within a probabilistic framework, which detects multiple objects without performing the non-maxima suppression step. Woordford (Woodford *et al.*, 2013) proposed two variants of Hough transform. The intrinsic Hough transform minimises the memory requirement by utilising the sparsity in the Hough space. The minimum-entropy Hough transform improves the recognition and registration accuracy by explaining the incorrect votes.

Both part-based and vote-based approaches achieve better performance in classifying textureless objects or objects with large appearance variations, cars and pedestrians. They leverage the spatial structure of an object class at a cost of lower flexibility to pose changes. Feature detection is an essential step for many 3D object recognition algorithms. Various 3D interest point detectors have been proposed for different applications. There are some comprehensive reviews for 2D interest point detectors (Mikolajczyk & Schmid, 2004), but 3D interest points have not been studied extensively.

Local feature-based methods are widely used in object recognition tasks, thanks to their robustness to occlusion and translation. While some approaches, such as bag-of-words (Fei-Fei & Perona, 2005; Sivic *et al.*, 2005) and topic models (Fergus *et al.*, 2005), use only the categorical distribution of appearance features.

Although part-based and vote-based methods can be used to infer pose at test time, they all require registered training instances for learning. Registration can be estimated from matching local features in a pre-processing step, using ICP (Pham *et al.*, 2011), RANSAC (Moreels & Perona, 2007), bunch graph matching (Wiskott *et al.*, 1997) and matrix factorisation (Arie-Nachimson & Basri, 2009), but these require either good initialisations or manual annotations for bootstrapping. Alternatively, Learned-Miller (Learned-Miller, 2006) proposed a data-driven method for registering an image collection. However, we know of no method which learns a shape and appearance model and infer pose of training instances *jointly*.

2.3.1 Early methods

Human pose estimation has been studied for decades, most of the early approaches focused on estimating poses in 2D images, including pictorial structure (Fischler & Elschlager, 1973) and template matching (Ioffe & Forsyth, 1999). These methods, however, lack an automatic part detector, which signifies that manual labelling is required for both training and testing data. Hence, their potential applications are greatly limited. On the other side, 3D human pose estimation is a more sophisticated task than its 2D counterpart because of occlusions and high dimensionality of the pose space. Nevertheless, various techniques have been investigated to estimate 3D poses from video data. For instance, Hogg (Hogg, 1983) used image edges to infer the 3D pose of a walking person captured from a carefully controlled scene. Micilotta (Micilotta *et al.*, 2006) used several appearance-based part detectors and boosting detectors for face and hand to compose a simple 3D upper-body pose. Navaratnam (Navaratnam *et al.*, 2006) designed a semi-supervised regression algorithm, where unlabelled training data were used to learn a Gaussian mixture based pose regressor and shape-context features extracted from silhouettes. Two extensive literature reviews on traditional 3D human pose estimation were presented by Aggarwal (Aggarwal & Cai, 1999) and Poppe (Poppe, 2007). More recent human pose estimation techniques are discussed below according to representations used for human pose.

2.3.2 Multiple-cameras approaches

Pose ambiguity is the central problem of 3D HPE. During data acquisition, 3D body poses are projected to 2D video frames or depth images. Occluded parts are difficult to recover from the data. Hence, each observation can be explained by more than one 3D pose configuration.

A standard approach to resolve the pose ambiguity issue is to maximise the area of view by simultaneously capturing multiple images with calibrated cameras (Pons-Moll *et al.*, 2011; Yao *et al.*, 2012). Although these methods guarantee the excellent accuracy, their potential applications are restricted to a calibrated and fixed multi-camera system. Meanwhile, resolving the pose ambiguity from multiple views is, still, a sophisticated optimisation problem. As a result, existing solutions are often computationally expensive, which further limit their potential applications.

2.3.3 Deformable part model

Most deformable part models (DPM) for 2D HPE are built upon the original seminal work of pictorial structures by Fischler and Elschlager (Fischler & Bolles, 1981). In a pictorial structure model, an object is recognised by evaluating the spatial arrangement of its constituent parts in the image. Early work by Felzenszwalb (Felzenszwalb & Huttenlocher, 2000, 2005) designed a probabilistic approach for the training and testing of pictorial structure models in an image. Recent proposed DPMs are often extensions of traditional pictorial structure models with new features or improved machine learning algorithms. For instance, Sapp (Sapp *et al.*, 2011) described the pictorial structure in a more complicated graph by decomposing it into smaller and stretchable components. Yang and Ramanan (Yang & Ramanan, 2011b) captured location-dependent appearances and spatial relations of parts with a structured SVM model. Similarly, a branch-and-bound algorithm was proposed by Sun (Sun *et al.*, 2012a) to extend the traditional pictorial structure beyond star-shaped or tree graphs. Hua (Hua & Wu, 2007) combined visual tracking with part detection in order to estimate articulated human pose. Ardriluka (Andriluka *et al.*, 2009a) revised the traditional pictorial structure using dense shape context and boosting algorithm. Eichner (Eichner *et al.*, 2012) presented a multi-phase algorithm that detects 2D body parts from unconstrained images using various clues such as face detection and graph cut.

As DPMs have shown encouraging performances in 2D HPE, especially in uncontrolled environments, it is suggested that similar techniques can be applied to improve 3D HPE. For example, by applying suitable inverse kinematic constraints, 3D poses can be estimated accurately from images with manually labelled parts (Ramakrishna *et al.*, 2012; Wei & Chai, 2009). In addition, the poselet algorithm (Bourdev & Malik, 2009) estimated a rough 3D pose by learning 2D part templates. Andriluka (Andriluka *et al.*, 2010) used a pedestrian detector with an automatic deformable DPM algorithm to estimate rough 3D poses in street scenes. Simo-Serra (Simo-Serra *et al.*, 2012) applied inverse kinematics to an optimise algorithm for the most probable 3D pose configuration from multiple noisy 2D DPM hypotheses. To summarise, the above-mentioned approaches demonstrate a greater flexibility than the traditional holistic-based 3D HPE systems. Hence, the use of part-based and mid-level features in multi-action 3D HPE is a topic with great research potential.

2.4 Detectors of human feature

A performance evaluation of volumetric 3D interest points will be presented later in this chapter. It will, first of all, provide an overview of current volumetric interest point detectors found in the literature. Then a selection of interest point detectors is evaluated quantitatively using the repeatability area score, which is a new unified performance metric that describes the repeatability and accuracy of an interest point detector. Finally, The qualitative characteristics of the interest points are compared.

Generally, proposed new volumetric interest point can be applied in many areas including medical imaging (Criminisi *et al.*, 2011; Donner *et al.*, 2011), shape retrieval,classification (Knopp *et al.*, 2010; Prasad *et al.*, 2011; Riemenschneider *et al.*, 2009) and video classification (Willems *et al.*, 2009; Yu *et al.*, 2010). Although much efforts on interest point detectors have been extensively devoted for images, (Tuytelaars & Mikolajczyk, 2008), how to evaluate 3D interest points is still a challenging work need to pay attention.

To demonstrate the basic principles of the formulations of 3D interest points and their evaluation, we describe some classic and well-known methods first. Below are the most common used methods such as DoH and Harris-based interest points (Laptev,

2005),DoG (Flitton *et al.*, 2010), VFAST (Yu *et al.*, 2010), SURF (Willems *et al.*, 2008) and MSER (Donoser & Bischof, 2006).

2.4.1 Scale-space and sub-voxel refinement

Creating a scale-space of the volumetric data as the input can provide the scale covariance of interest points, so the introduction of a Gaussian smoothing kernel to the input volume can create an octave of linear scale-space. Then fine-scale structures can be suppressed by applying such a smoothing kernel on the volume recursively. After that, the down-sampling is used for input volumes from the previous octave to create a new octave. Within these steps, a lot of volumes can be created with multiple levels of detail. For more information, on the subject of the interest point detection, the detail implementation of scale-space can be found in (Lindeberg, 1998).

However, the representation of a scale-space is not designed for computing the MSER because the salient regions are detected in different scales. In another word, the interest points are located by MSER through fitting an ellipsoid to the detected salient region (Matas *et al.*, 2004). The computing of saliency responses is in every volume within the scale-space for other interest point detectors. Moreover, all these detectors need the subpixel refinement process in SIFT (Lowe, 2004). The introduction of 4D quadratic functions, which is used for fitting around the local scale-space maxima, can help to locate the interest points at the sub-voxel level and select the maxima of these functions instead.

2.4.2 Determinant-of-Hessian

With respect to the formulation in (Lindeberg, 1998) and similar to the Harris detector, the DoH (Determinant-of-Hessian) interest point is also one of the common used detection method. The difference is that the second-moment matrix, a Hessian matrix $\mathbf{H}$ is computed from $\mathbf{v}$:

$$\mathbf{H} = \begin{bmatrix} \mathbf{v}_{xx} & \mathbf{v}_{xy} & \mathbf{v}_{xz} \\ \mathbf{v}_{yx} & \mathbf{v}_{yy} & \mathbf{v}_{yz} \\ \mathbf{v}_{zx} & \mathbf{v}_{zy} & \mathbf{v}_{zz} \end{bmatrix}, \tag{2.1}$$

where $\mathbf{v}_{xy}$ denotes the second derivative of the volume at scale σ_s, along x and y axes, such that

$$\mathbf{v}_{xy} = \frac{\partial \mathbf{v}(\mathbf{x}; \sigma_s)}{\partial x \partial y}. \tag{2.2}$$

The saliency response is the scale-normalised determinant of the Hessian matrix $\mathbf{H}$:

$$S_{\text{Hessian}} = \sigma_s^3 \det(\mathbf{H}). \tag{2.3}$$

Subsequently, interest points are located at the 4D scale-space local maxima of S_{Hessian}.

2.4.3 Difference-of-Gaussians

One of the blob detection technique is the DoG (Difference-of-Gaussians) operator, which is used for feature localisation popularised by the SIFT algorithm (Lowe, 2004). The DoG detects features of a particular size by approximating the Laplacian-of-Gaussian filter. By subtracting two Gaussian smoothed volumes, the saliency response of DoG detector S_{DoG} can be computed. To make it more clearly, the volumes are taking the absolute values of the difference, which is usually the adjacent scale-space representations of the same input data.

Interest points are detected at the 4D local maxima and 3D space plus scale, with respect to the saliency response S_{DoG} within each octave of $\mathbf{v}(\mathbf{x}, \sigma_s)$:

$$S_{\text{DoG}}(\mathbf{x}; \sigma_s) = \left| \mathbf{v}(\mathbf{x}; \sigma_s) - \mathbf{v}(\mathbf{x}; \sigma_{s-1}) \right|. \tag{2.4}$$

Volume $V(x, y, z; \sigma_s)$ indicates the scale-space representation of the input volumetric data at scale σ_s.

2.4.4 Harris corner detector

The local window sliding is used to examine the image gradients for a Harris corner detector, at the same time, the interest points can be detected at positions with observed large changes in all directions (Harris & Stephens, 1988). Laptev (Laptev, 2005) uses separate scale parameters to detect the first 3D extension of the original Harris corner for the heterogeneous space and time axes. In the work, the scale σ_s is the only one, which is shared among three homogeneous spatial axes for simplicity.

Smoothing the first derivatives of the volume in scale-space $\mathbf{v}(\mathbf{x};\sigma_s)$ can provide the second-moment matrix $\mathbf{M}$ with a spherical Gaussian kernel $g(\cdot;\sigma_{\text{Harris}})$ is given, which can be derived as follows:

$$\begin{aligned}
\mathbf{v}_x(\mathbf{x};\sigma_s^2) &= \frac{\partial \mathbf{v}(\mathbf{x};\sigma_s^2)}{\partial x}, \\
\mathbf{v}_y(\mathbf{x};\sigma_s^2) &= \frac{\partial \mathbf{v}(\mathbf{x};\sigma_s^2)}{\partial y}, \\
\mathbf{v}_z(\mathbf{x};\sigma_s^2) &= \frac{\partial \mathbf{v}(\mathbf{x};\sigma_s^2)}{\partial z}, \\
\mathbf{M}(\cdot,\sigma_{\text{Harris}},\sigma_s) &= g(\cdot;\sigma_{\text{Harris}}) * \begin{bmatrix} \mathbf{v}_x^2 & \mathbf{v}_x\mathbf{v}_y & \mathbf{v}_x\mathbf{v}_z \\ \mathbf{v}_x\mathbf{v}_y & \mathbf{v}_y^2 & \mathbf{v}_y\mathbf{v}_z \\ \mathbf{v}_x\mathbf{v}_z & \mathbf{v}_y\mathbf{v}_z & \mathbf{v}_z^2 \end{bmatrix},
\end{aligned} \tag{2.5}$$

The second moment matrix $\mathbf{M}$ denotes the auto-correlation along different directions in a local neighbourhood of size σ_s. While $\mathbf{v}_x$,$\mathbf{v}_y$,$\mathbf{v}_z$ represents the partial derivatives of the volume in scale-space $\mathbf{v}(\mathbf{x};\sigma_s)$ along x, y and z axes respectively.

The coordinates (x, y, z) are the located candidate interest points, and the second moment matrix matrix $\mathbf{M}(x, y, z;\sigma_{\text{Harris}},\sigma_s)$ has large eigenvalues. With the idea used in (Mikolajczyk & Schmid, 2004), the window size σ_{Harris} is proportional to expected feature scales σ_s by a factor of 0.7. With the sub-voxel refinement method, locations of interest points can be refined. In the scale-space of S_{Harris}, the interest points are selected at the 4D maxima.

The saliency response of Harris corner, S_{Harris}, is computed from the determinant and trace of the second-moment matrix $\mathbf{M}$:

$$S_{\text{Harris}} = \sigma_s^3 \det(\mathbf{M}) - \kappa \operatorname{Tr}(\mathbf{M})^3, \tag{2.6}$$

where κ is a tunable sensitivity parameter that controls the rejection of edge points. The saliency response S_{Harris} is normalised by its scale σ_s.

2.4.5 VFAST

The introduction of FAST corner detector is very successful in many image-based area (Rosten *et al.*, 2010). The improvement of the FAST algorithm has been introduced for video-based object classification. The top two famous algorithms are VFAST (Yu

et al., 2010) and FAST-3D (Koelstra & Patras, 2009). As we only consider 2D cased, so the VFAST is introduced briefly here. In a video sequence, the VFAST interest points are detected by directly comparing intensities. The VFAST detector can be further accelerated based on the theory of the FAST by learning a decision tree-based corner detector from training videos (Rosten *et al.*, 2010).

2.4.6 Speeded up robust features

Speeded up robust features (SURF) is a well-known and efficient feature extraction algorithm (Bay *et al.*, 2008). Willems first introduced the 3D volumetric version of SURF for video classification (Willems *et al.*, 2008). It can also be used in many other vision-based tasks such as 3D shape object recognition (Knopp *et al.*, 2010).

SURF can be seen as a special and efficient approximation of the DoH detector. In the DoH detector, six Haar wavelets and box filters, the second-order of Gaussians can be derived approximately. The integral videos or volumes can greatly accelerate the computing process of the convolutions of the Haar wavelets. In addition, there is a similarity between the saliency response of 3D SURF and the aforementioned DoH detector.

2.4.7 Interest point detectors

The first stage of many image-based tasks, such as object classification and pose estimation, is the interest point detection process. The basic task of an interest point detector is to locate the salient features or interest area from input image data/ video data for the next step processing. Then the following task uses the detected interest points to match corresponding points across two or more similar sets of data. Most researchers mainly focus on the feature detection on 2D images, which makes the algorithm of 2D image feature detection dominate the technique on 2D images. Consequently, some reviews have been performed extensively on locally-invariant feature detectors such as Tuytelaars and Mikolajczyk (Tuytelaars & Mikolajczyk, 2008).

The recent novel image sensors, such as Kinect, advance the data acquisition techniques greatly and make it possible to use 3D data as the input of an algorithm. More researchers have been attracted to the shape-based pose estimation systems. For example, Google Warehouse (Lai & Fox, 2010) and the B3DO dataset (Janoch *et al.*,

2011) work hard on the large-scale synthetic and realistic 3D repositories. As a result, more and more interest point detection techniques based on 3D data have been proposed in the computer vision research field. According to the representation of input data, the existing techniques for 3D interest point detection can be grouped as two, one is geometry based detectors and one is volume based. Geometry-based interest point detectors use the geometric information to locate features. The information consists of contours, surface normals and surface patches, while the general data representations of this technique cover synthetic meshes and point clouds (Aanæs *et al.*, 2012; Glomb, 2009; Sipiran & Bustos, 2011; Unnikrishnan & Hebert, 2008; Zaharescu *et al.*, 2009)

Volume-based detectors use the pixel or voxel values directly from the volumetric scalar data. This type of data range from time-varying video data (Koelstra & Patras, 2009; Willems *et al.*, 2008; Yu *et al.*, 2010), CT scan (Dalvi *et al.*, 2010) volumes and binary volumes generated from depth images (Hadfield & Bowden, 2013) to 3D meshes (Knopp *et al.*, 2010). Geometry-based interest point detectors were introduced in some literature review papers such as Salti (Salti *et al.*, 2011) and Dutagaci (Dutagaci *et al.*, 2011). Nevertheless, unlike 2D interest points, performance evaluation of 3D interest points remains an unexplored topic.

2.4.7.1 Blob detection

Similar to the determinant of the Hessian (DoH), which is introduced before, scale-covariant interest points used the Laplacian-of-Gaussian kernel. It is equivalent to the trace of the Hessian from the mathematical point of view (Lindeberg, 1998). The introduction of a difference of Gaussians operator for approximating the above kernel was proved to have the advanced computational performance. The DoG algorithm was also used in some other computer vision areas such as volumetric scans (Flitton *et al.*, 2010),3D object detection, recognition of synthetic meshes (Wessel *et al.*, 2006) and multi-view stereo data (Pham *et al.*, 2011). Recently, Hadfield (Hadfield & Bowden, 2013) proposed 4D and 3.5D extensions of Harris corner, this novel method uses a corner detection method with the complimentary 3D spatiotemporal volumes, appearance and depth sequences.

The proposed speeded up robust feature has been proved to accelerate the computation time of the determinant of Hessian operator if the integral images and box

filters are used (Bay *et al.*, 2008). Since then SURF has been applied to spatiotemporal data (videos) (Willems *et al.*, 2008) and volumetric data generated from synthetic 3D mesh models (Knopp *et al.*, 2010). In addition to Harris corners, Hadfield (Hadfield & Bowden, 2013) also extended the Hessian-Laplace interest point for recognising 3D spatiotemporal volumes.

Similarly, the hessian-Laplace detector detects the interest points by evaluating the Hessian matrix of an input image (Mikolajczyk & Schmid, 2004). In contrast, the region-based detector tries to find the salient regions. The maximally stable extremal region is one example of this method (Matas *et al.*, 2004). Specifically, both SURF and DoG are grounded on the approximation of the Laplacian-of-Gaussian kernel, while maximally stable extremal region method applies the threshold changes to the threshold regions in the maximally stable areas. Three dimensional MSER has already been applied to volumetric data for both the context of segmentation of MRIs and the video data (Riemenschneider *et al.*, 2009). So this is inherently multi-scale, in another word, it is invariant to affine intensity variations and covariant with affine transformations.

2.4.7.2 Corner detection

Its 3D adaptations have been applied to registration of volumetric CT scans (Dalvi *et al.*, 2010; Ruiz-Alzola *et al.*, 2001). With respect to the success of Harris corner detector, Mikolajczyk (Mikolajczyk & Schmid, 2004) found the Harris corners in the spatial domain thus improved the scale-covariant Harris-Laplace detector. They are maxima of the Laplacian in the scale domain. Laptev extend this method to space-time interest points for video classification(Laptev, 2005).

The corner detection is also a type of interest point detector, it detects only the corner area of the input data. Corner detectors operate directly on image pixels instead of detecting corners by image gradients. One of the classic image-based corner detection algorithms is the Harris interest point detector, which searches points of large gradient changes in orthogonal directions (Harris & Stephens, 1988) by analysing the eigenvalues of the first order derivative (second-moment matrix). Smith and Brady (Smith & Brady, 1997) presented the SUSAN corner detector. In contrast with the central pixel value, it leveraged the proportion of pixels in a neighbourhood. The fast detector

measures the largest number of contiguous pixels in a circle, these pixels are significantly brighter or darker than the centre pixel. Rosten (Rosten *et al.*, 2010) proposed the fast corner detector with an accelerated segmented test which is a relaxed version of SUSAN for locating stable corners in an image. The computation speed could be further improved because there is no need to compute the derivative at each pixel and the feature detecting process is to learn a decision tree classifier. Benefit from the efficient performance, this feature detector can be applied to many areas including the video classification (Koelstra & Patras, 2009; Yu *et al.*, 2010).

2.4.7.3 Covariant characteristics

It is noted that image-based detector has been made affine-covariant, that is to say, the perspective distortion caused by the projection of 3D world onto the 2D image plane need to be approximating(Mikolajczyk & Schmid, 2002). When it undergoes the same transformation as the data, a feature characteristic is considered as covariant. As most shape acquisition techniques do not have the 3D-2D projection process, so it is not always necessary for the covariance with 3D shapes or in a variant to viewpoint changes. However, when the data is acquired, the object very likely has many poses or rigid shapes, such as translation, rotation and scaling. Therefore, the rotation and scale covariance become extremely essential when processing the 3D shape data. Moreover, the data is robust to illumination and lighting conditions except for texture-mapped meshes (Zaharescu *et al.*, 2009). To sum up, the quality of shape data is determined by many factors such as sampling artefacts, noise, occlusions and holes from the reconstruction process.

2.5 Summary

This chapter reviewed related human pose estimation on both human model representation and features used in computer vision. To sum up, researchers have put much effort on the challenges, but how to balance the features and models is still one of the main issues that this thesis will focus on. So the understanding of both models and features could promote the research of the human pose estimation.

Chapter 3

Deformable Parts ConvNet Based Pose Estimation

Human motion recognition is a trending topic and could be applied in many areas. The motion estimation is extremely challenging because of the high uncertainty of human activities. We thus introduced a novel method which is designed for estimating the upper joints and recognising their special motions. In addition, we verified the proposed method on a dataset and the experimental results show the proposed method is effective on the dataset.

3.1 Introduction

The activity recognition usually means to learn about the activities from video sequences and identify similar actions with machine learning method. Human activity recognition is very important in computer vision research area today, as it can be applied in many fields including the surveillance system, human-machine interfaces, video indexing, virtual coaching, VR games, patient monitor system(Aggarwal & Ryoo, 2011)and some motion related application (Kyriazis *et al.*, 2016).

In general, the activity recognition system needs to have the ability to track the human motion (Liu *et al.*, 2017) and recognise complex human motions from a continuous video sequence or from only a static image. Such a system usually can be classified into two approaches according to the input data in a contactless method,

which is known as the computer vision-based activity recognition (Shahroudy *et al.*, 2016), instead of a wearable-based method (Kumari *et al.*, 2017). As the wearable-based method could limit the human pose and affect the possible motion, we only focus on the vision-based method in this paper.

Different features have been used in activity recognition methods. Michel (Michel *et al.*, 2014)adopted a tracking method to capture the articulated motion including the 3D position and orientation with two RGB-D cameras. Spatio-temporal and bag-of-words features are used to represent human motion in many works of literature. Semantic features are used to explain the meaning of a motion. For example, it is understandable that a car appears on a road while it is not acceptable for some people that a giraffe appears in a kitchen.

Although lots of researchers work hard on the activity recognition using different methods, many factors, including the diversity of appearances, the variation of the camera angles, background clutter, illumination changes in a scene, and occlusion by other objects, pose a challenge on the performance of the activity recognition. Some research methods were proposed to handle some of these issues or one aspect of them. For example, to handle the illumination changes, depth information based method (Vemulapalli *et al.*, 2014) were used for more accurately estimating the human pose. Multi-view based method (Gall *et al.*, 2010) was used in the activity recognition system to avoid the negative effect of occlusion.

Activity recognition has been researched for many years and some review papers (Chen *et al.*, 2013b; Lillo *et al.*, 2017; Vrigkas *et al.*, 2015; Ziaeefard & Bergevin, 2015) suggested the features for effectively representing motions play a key role in this area. Hassaballah (Hassaballah *et al.*, 2016) provided an overview of image feature range from detection, description to feature matching which are fundamental components for handling computer vision issues. The interesting point and local image features contributed to represent object patterns in a static image but failed to represent features for a dynamic image sequence. Space-time interest point was raised as a response. Such a method (Laptev, 2005) performed well for some simple motions such as walking and running.

To improve the human pose estimation on a single image, one way is to extending a static recognition method with utilising a regularisation on the body parts over time by using a probabilistic graphical model (Cherian *et al.*, 2014). This method typically

represents human body parts corresponding to different major body parts such as head, shoulders, elbows and hands. By forming these node parts into a graph, a kinematic method is usually adopted to capture the inter-part relationships. The Pictorial structure model (PSM) (Johnson & Everingham, 2011) allowed the inference to estimate the possible poses over the pose space.

The ordinal pattern is normally seen as the low-level feature. A middle-level feature, which was integrated into an orderlet (Yu *et al.*, 2015) character, was proposed to represent the relationships among joints and shape information respectively on both skeletons and depth maps. While High-level pose features (HLPF) were introduced for encoding spatial and temporal relations of human skeleton joints (Jhuang *et al.*, 2013). The performance of dense trajectories features have proved to be excellent in some activity recognition datasets (Wang *et al.*, 2013). In addition, spatiotemporal features have been applied in the activity recognition for representing the action with a dense feature set, while Shi (Shi *et al.*, 2013) introduced a fast random sampling method on a local part model to speed up the computational efficiency.

Cheron argued that the representation of human pose dominates the performance of the action recognition and introduced a pose-based scheme, which aggregated the descriptor based on the human pose for tracking human body parts (Cheron *et al.*, 2016). This supervised method extremely relied on the annotation of the human body parts and hand-crafted feature extraction, which needs considerable relative skills and lots of restless work, thus puts lots of burden on the human.

There is little previous work with enough annotations contribute for pose estimation, Johnson and his colleagues (Johnson & Everingham, 2011) proposed a method to estimate the human pose with only inaccurate annotation. Some computer vision related work had proved that the approach is useful, for example, the collaborative LabelMe object annotation system (Russell *et al.*, 2008) and utility data annotation (Sorokin & Forsyth, 2008) still benefit when obtaining data from some inexperienced annotators.

Apart from the methods for estimating poses in a single image and the spatiotemporal feature representing methods, a scheme, for continuous motion recognition, based on the static image feature is also important. With accurate skeleton information such as the position and the angles, a skeletal representation is needed for encoding the features with a dynamic time warping. Vemulapalli (Vemulapalli *et al.*, 2014) explored

a method through modelling the 3D geometric relationships among body parts with 3D space rotations and translations. To estimate the 3D human pose by optimising the joints over the set of the manifold with a particle-based optimisation algorithm, the low-dimensional manifold (Gall *et al.*, 2010) was analysed to emphasize the importance of a successful scheme for pose estimation in videos and handle the temporal coupling across time.

3.2 Deformable mixture-of-parts model

Our model is mainly inspired by three most recent papers (Belagiannis & Zisserman, 2016), (Zhou *et al.*, 2016) and (He & Chen, 2017), the first introduced an end-to-end model to train a recurrent human pose estimation, the second introduced a 3D human pose estimation from monocular video, the last introduced a visualizing method which given insight into both the function of intermediate feature layers and the operation of the classifier. In this section, we will first introduce the method with sampling strategy used for estimating the key joints information, then we introduce a skeletal descriptor method to represent the joints, to make the scheme work more effectively, a Gaussian Process is adopted for mapping between different dimensional space, then an on-line method is utilised for recognizing the real-time activities of the child. Human pose feature, especially the body joints, is essential for activity recognition. A deformable mixture-of-parts model is used to represent the body parts for a single image because of the computational efficiency and considerable property (Yang & Ramanan, 2013). The upper body part is modelled as a set of major joints which are the head, neck, two shoulders, two elbows, and two wrists (or hands). Theses joints contribute significantly the performance of the upper body motions. A pictorial structure model which uses the tree-type graph with nodes is introduced to represent each joint position and orientation. For some specific camera angles, self-occlusion could happen. To handle this issue, a clustering method is used to classify each body part with annotated ground truth T_i for one of the n training images. The problem is formulated as a maximum-likelihood problem through calculating the highest probability:

$$\hat{\Theta} = \underset{\Theta}{argmax} \prod_{i=1}^{N} \max_{j=1}^{k} P(T_i|\Theta_j) \tag{3.1}$$

There are k pose clusters in total, $P(T_i|\Theta_j)$ is the posterior probability of a particular pose for an image I, which is defined as:

$$P(p|I) \propto f(I|p)f(p) = \prod_i f(r_i|l_i) \prod_{(l_i,l_j \in E)} (l_i|l_j) \tag{3.2}$$

l_i denotes the 2D position and orientation, which is one element of the set $p = \{l_1, l_2, ...l_n\}$,r_i is the corresponding image region, the prior term defines the prior probability of a configuration. This has two main advantages: on one side, it can help to overcome the ambiguous image data, on the other side, it limits the model from the plausible human configurations when the kinematic limits of the body are learned.

In addition, a linear SVM classifier is used for each body parts, the classifier is bootstrapped with some negative samples of other body regions and non-body regions for training. The responses can be computed for each body part is:

$$p(r_i|l_i, \Theta_i) \propto \max_{j=1...n} w_j \Phi(r_i) \tag{3.3}$$

In which w_j is the weight vector for component j, $Phi(r_i)$ is the feature vector from the image region r_i. The maximum value allows us to determine the appearances mode with the highest confidence.

3.2.1 Dense Sampling

For a more efficient computation, we use a random sampling strategy for the denser patches, let us look an image with size $n \times m$ for instance, the number of possible sampled patches is n^4 which is explained in (Lampert *et al.*, 2008), besides, it is proved that the performance could be improved with randomly sampled patches for each image (Nowak *et al.*, 2006). Based on this, reducing the number of sampled points for an individual frame and still maintain an efficient sampling density for representing the features.

3.2.2 Joints Descriptor

With the skeleton information obtained, we use a skeletal representation method to represent the body part. The method was proposed in a previous paper(Vemulapalli

et al., 2014), which mainly considered a whole body parts, we slightly change the method for represent only the upper body. When a pair of body parts is given, their relative geometry is described as e_m and e_n, which denote the eight joints and oriented rigid body parts respectively, the starting point ($e^n_{m1}(t)$) and end point ($e^n_{m2}(t)$) of each part can be represented in a local coordinate system at time instance t.

$$\begin{bmatrix} e^n_{m1}(t) & e^n_{m2}(t) \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} R_{m,n}(t) & \overrightarrow{d}_{m,n}(t) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & l_m \\ 0 & 0 \\ 0 & 0 \\ 1 & 1 \end{bmatrix} \tag{3.4}$$

$$\begin{bmatrix} e^m_{n1}(t) & e^m_{n2}(t) \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} R_{n,m}(t) & \overrightarrow{d}_{n,m}(t) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & l_n \\ 0 & 0 \\ 0 & 0 \\ 1 & 1 \end{bmatrix} \tag{3.5}$$

where $R_{m,n}(t)$ and $R_{n,m}(t)$ are the rotations, $\overrightarrow{d}_{m,n}(t)$ and $\overrightarrow{d}_{n,m}(t)$ are the translations, these are measured in the local coordinate system. More detailed information for representing the joints we refer the (Vemulapalli *et al.*, 2014).

3.2.3 Gaussian Process

From the joint information in motion capture data, we use a Gaussian Process regressions, which is a straightforward extension of Gaussian Mixture Model, to map a low-dimensional space from a high-dimensional space.The equation 3.7 indicates the back process of the mapping.

$$x = f_a \sim GP(m(y), k(y, y^{'})) \tag{3.6}$$

$$y = g_a \sim GP(m(x), k(x, x^{'})) \tag{3.7}$$

f_a denotes the mapping from high-dimensional to low-dimensional space, while g_a denotes the inverse process, where m represents the mean and k denotes the covariance functions. M_a is learned to model the temporal transitions between effective motions for an action-specific manifold.

$$x_t = M_a(x_{t-1}) \sim GP(m(x-1), k(x_{t-1}, x^{'}_{t-1})) \tag{3.8}$$

Instead of using a single state space, a set of action-specific manifolds is considered, A_c is defined as a set $\{a_1, a_2, ..., a_{|A|}\}$,which denotes the action classes, where we consider to learn an action-specific manifold for all the classes. As the manifolds only utilised the joint space, the representation of a body pose is determined by $y_a = (r, t, \Theta_a)$, (r, t) is a vector indicates the global orientation and position, Θ denotes the joint angles.

3.2.4 SVM Classification

As our aim is to recognise both static motions and dynamic motions, we introduce a scheme which can estimate both motions, for the single image, a pose set regards to a tree-graph which including the 2D coordinates for representing the body parts is defined as:

$$P_s = p^i = (x^i, y^i) \tag{3.9}$$

Then we formulate the estimation issue as a minimization problem with the cost $C(I, P_s)$:

$$C(I, P_s) := \Sigma_i \phi_i(I, p^i) + \Sigma_{i,u} \varphi_{i,u}(p^i - p^u) \tag{3.10}$$

3.2.5 Motion Recognition

For a real-time activity recognition system, it needs to predict a continuous video sequences with reliable scores of different classes. The frame-level score is defined as:

$$R(I_t) = \sum_{a_1=1}^{a_{|A|}} \alpha_m R_m(I_t) \tag{3.11}$$

$R_m(I_t)$ denotes the response of a orderlet on the frame I_t,while α_m is the corresponding weight, which decides the balance between the positive and negative votes. It is clear that different types of actions have various properties such as the action speed and the duration. These make it difficult to determine the size of a fixed-length window. The temporal smoothness with adaptive smoothing window length is introduced for a reasonable result. The main concept is to maintain a reliable voting score for $t - th$ frame.

$$S(V_t) = max(0, S(V)_{t-1} + R(T_t)) \tag{3.12}$$

$S(V_t)$ denotes the score at time t, if the value is greater than 0, it means the current action is continuing, on the contrast, if the value is less than 0 or equal to 0, there is no action is happening. Then the value will be reset to 0 and forecasts that a new action will start.

3.3 Results

The main aim of this work is to recognise 11 activities and verify the efficiency of the proposed activity recognition framework. Figure 6.4 shows the previous results when training our network.

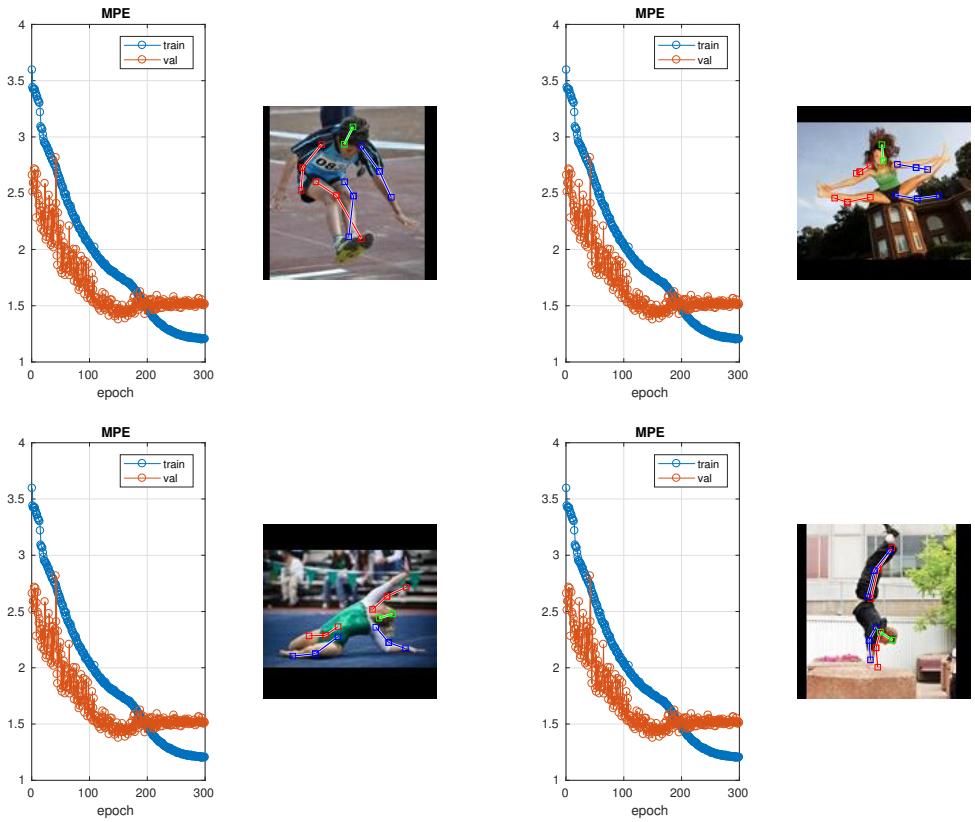


Figure 3.1: Network result

In this section, we report the results on some popular dataset within only our method as the method is designed only for the specific purpose. Figure 3.2 shows some result samples from the dataset. The joints information is estimated with our

Figure 3.2: The result samples for representing the joints

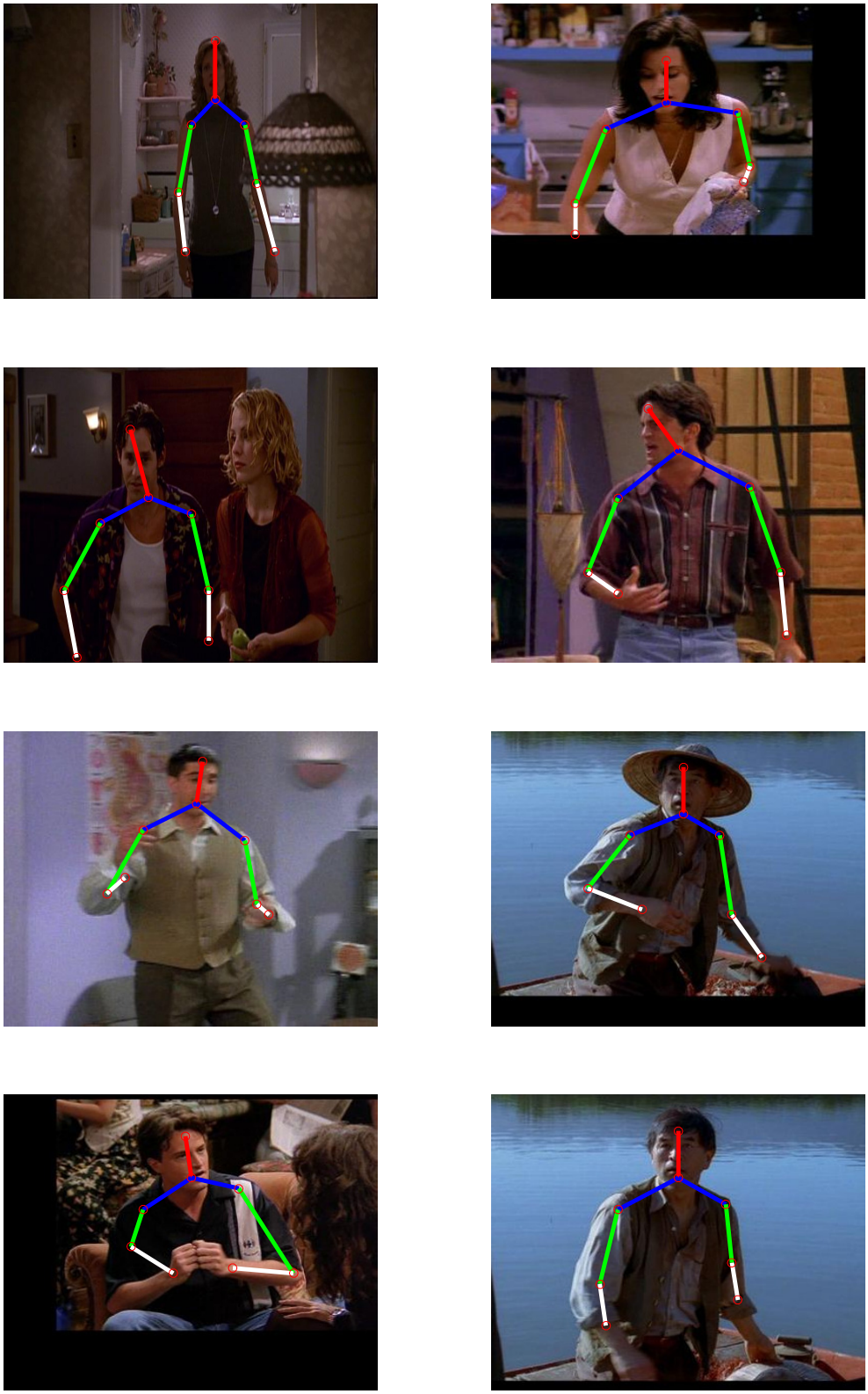


Figure 3.3: The result samples supplement for representing the joints

method and the table 3.1 shows the average accuracy of the estimating results for the upper body major joints.

Our dataset is extremely more challenging than the existing dataset as there are more unpredicted factors when recording the data as similar as behaviour dataset. The average accuracy is still kept at 84.7%.

Table 3.1: The average accuracy of the joints(%)

joints	accuracy
neck	87.9
shoulders	84.6
elbows	76.6
wrists	78.2
upper body	96.3

The figure 3.4 indicates the confusion matrix for estimating the motions on both our datasets. The average accuracy for predicting the motions is 85.9%, which can be seen as an acceptable result.

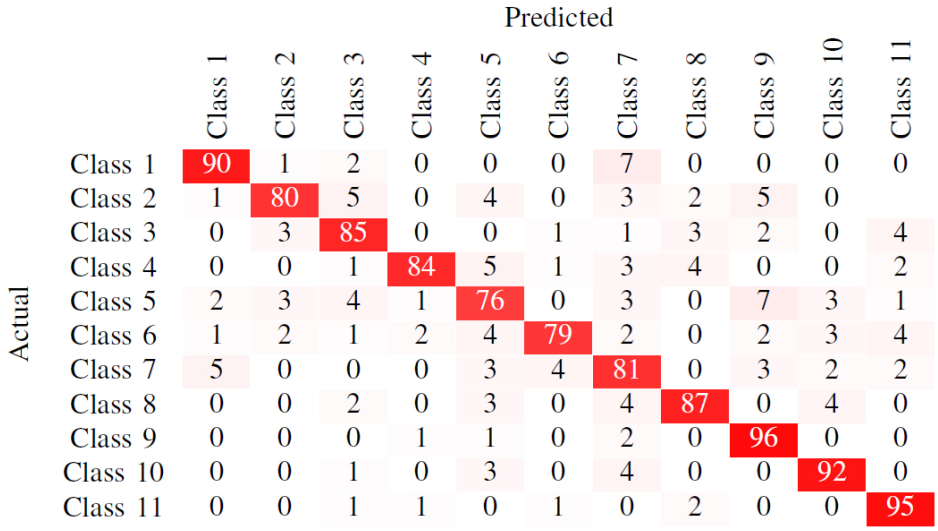


Figure 3.4: The recognition results

In this paper, we mainly focus on the atomic motions including namely waving the hand, drinking, and moving a toy etc., which are defined by the therapist, these move-

ments indicate a stable coordination pattern among the skeleton joints, each activity normally contains a joint set order (Yu *et al.*, 2015).For example, when the child is doing the drinking motion, the child first hold the cup from the table, move the cup to his/her mouth, hold on for some seconds and put the cup back to the table. We believe that if skeleton joints especially the wrist, elbow and shoulder are estimated accurately, they will provide us with an effective feature for modelling the motion and recognizing the motion. Thus the accuracy of joints information comes from the very first step for estimating both the continuous motions and some static motions.

We have presented the joint estimating results and motion estimation results. The accurate estimation of joint could provide an excellent classification result even using a linear SVM classification method, which implies the importance of the joints estimation for our datasets.

3.4 Summary

We propose a novel activity recognition method which is designed especially for recognising the upper body motion, we run our algorithm on both the upper body motion dataset and the full body motion dataset to verify the effectiveness of the proposed method. The experimental results show that our approach performs well on the datasets. The research confirms that the correct classification of the body parts leads to a significant improvement in estimating the human joints, and the pose estimation can benefit from the accurate human joints.

Chapter 4

Pictorial Structure Feature for Pose Estimation

4.1 Introduction

Pictorial structure model is used in many pose estimation methods, as they benefit from the linear time model for searching over the full pose space when forming a part-dependencies tree. However, it is too large for an individual part state space when evaluating complex appearance models densely. Therefore, many simple linear filters, such as edges, colour and locations, are used for assisting modelling process. In addition, suffered from the quadratic state-space complexity, it is difficult to speed up inference as of the cost of image-based deformation on part-part relationships when computing convolution or distance transform. As a result, weak appearance cues will be inferred from poor localization of parts, what is more, they are more sensitive to background clutter. The accuracy of lower arms in a whole human figure could be lower compared with the other parts such as torso or head. So it is extremely important to use a more robust model for localizing these tricky parts, this will require a combination of some meaningful appearance parts such as contour continuation and segmentation cues and requires richer models of individual part shape and then modelling a joint part-part appearance for densely computing.

Based on this, an improved pictorial structure is proposed. The advantage of the proposed method can learn the pictorial structures even the pose resolution is increasing while keeping the pose state space. Compared with some classic methods, our

model can handle each level on a certain spatial and angular resolution. When we referred the level, it clarifies the candidates from previous feature layer for choosing proper poses with an inference process. The model can choose pose for each part according to the computed max-marginal score under the computational budget. The difference between our model and conventional pruning heuristics is that a simpler model is inferred when considering the pruning process to handle the output.

Overview of APS: At left, the original spring pictorial structure model. At right, the standard PS model for a 2D human pose. The states are shown as unit vectors indicating the position of joints and their direction. The mean displacement between joints is shown as solid black circles, connected by solid black lines to show the kinematic tree structure. The displacement from mean positions are shown as springs stretching.

It is obvious that a much smaller hypothesis set will be concerned for our model at the final level. This will contribute a more powerful model as it makes it easy to combine more valuable features. Unlike the traditional geometric features and part detectors, we use the proposed model to combine richer features and make the object boundary continuity and smoothness. The mid-level and bottom-up cues will be complementary to the traditional HoG-based part models. This is shown in the overview figure and more discussion will be found in the experimental part.

For our model, which is a general linear MRF and considers the part configurations based on unary terms and arbitrary pairwise:

$$s(x, y) = w \cdot f(x, y) = \sum_{i \in \gamma} w_i \cdot f_i(x, y_i) + \sum_{ij \in \gamma} w_{ij} \cdot f_{ij}(x, y_i, y_j) \tag{4.1}$$

In the equation, the pairwise and unary weight vectors parameters w_{ij} and w_i are corresponding to the pairwise and unary feature vectors $f_{ij}(x, y_i, y_j)$ and $f_i(x, y_i)$, the $\gamma = (V_\gamma, \epsilon_\gamma)$ indicates the tree-structured graph of part interactions. The main differences compared with the previous pose estimation model are lied on two aspects, firstly, data-independent terms are allowed in our pairwise cost function, secondly, there is no need to fit the constrained parameters with other parametric distribution, i.e. a Gaussian distribution. More specifically, in a positive semi-definite covariance matrix, we do not need the corresponding weights to be combined into the pairwise features which are $y_i \cdot y_i$, $y_j \cdot y_j$, and $y_i \cdot y_j$ in the general models.

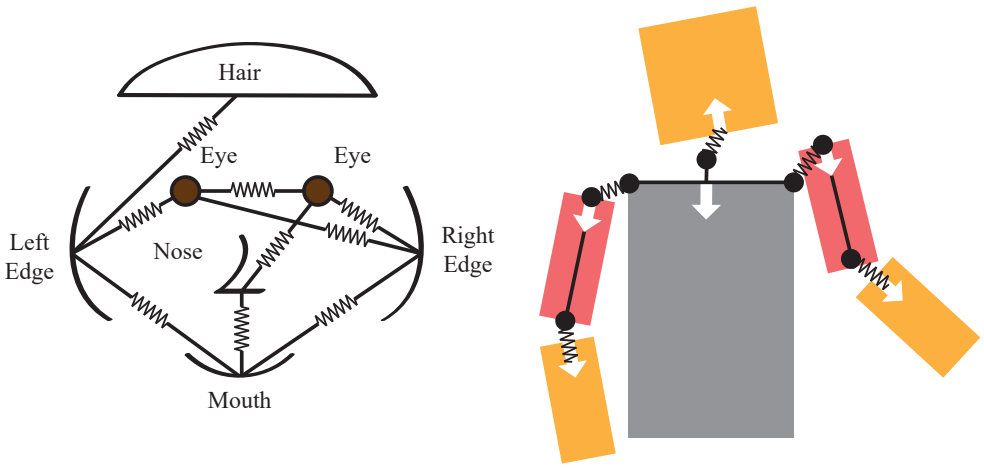


Figure 4.1: A Spring model of pose.

To get the advanced scoring assignment $argmax_y s(x, y)$, a standard $O(nk^2)$ dynamic programming technique is needed as the general form inference might fail to transform distance with a satisfied performance. That is to say, in practice, the combination of some meaningful features could remain sub-quadratic and efficient.

For unstructured and binary classification, combined of classifiers have been quite successful for reducing the computation. Fleuret & Geman (2001) proposed a coarse-to-fine sequence of binary tests to detect the presence and pose of objects in an image. The learned sequence of tests is trained to minimize expected computational cost. The extremely popular Viola-Jones classifier (Viola & Jones, 2002) implemented a combined model of boosting aggregate, with earlier stages using fewer features to quickly reject large portions of the state space.

It is quite possible to reduce computation by combining the precious classifiers, unstructured and binary classification, Our combined model is inspired by these binary classification combined. In natural language parsing, several works (Carreras *et al.*, 2008; Petrov, 2009) use a coarse-to-fine idea closely related to ours and Fleuret & Geman (2001): the marginals of a simple context-free grammar or dependency model are used to prune the parse chart for a more complex grammar.

Recently, P. Felzenszwalb (2010) proposed a combined for a structured parts-based model. Their combined works by early stopping while evaluating individual parts, if the combined part scores are less than fixed thresholds. While the a form of this combined can be posted in our more general framework (a combined of models with an increasing number of parts), we differ from P. Felzenszwalb (2010) in that our pruning is based on thresholds that adapt based on inference in each test example, and we explicitly learn parameters in order to prune safely and efficiently. In Fleuret & Geman (2001); P. Felzenszwalb (2010); Viola & Jones (2002), the focus is on preserving established levels of accuracy while increasing speed. The focus in this paper is instead of developing more complex models—previously infeasible due to the original intractable complexity—to improve the state-of-the-art performance.

4.2 Structured Prediction

The recently introduced Structured Prediction Combined framework (Chu *et al.*, 2016) provides a principled way to prune the state space of a structured prediction problem

via a sequence of increasingly complex models.

For a structured prediction problem, increasing a sequence of complex models is an effective way to reduce the state space. Many methods can make it come true. For example, a coarse to a fine method, which means a simple start process with a complex end, is one of the effective ways. Another option is to follow a special order, from unary and pairwise to ternary. This higher-order cliques then be introduced into successive stages. Then it is needed to consider the prune and refine the process.

The first scheme is used in our model with simple features, we will allow enough time to make the reasonable fine stage finished with a right resolution, then more complex feature can be introduced. To perform robust yet quick inference at the beginning state space, we use the geometric features and some standard pictorial structures with unary detector scores.

4.3 Algorithm Inference

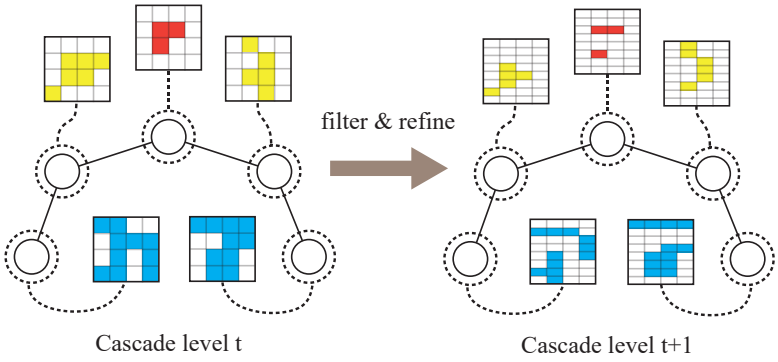


Figure 4.2: Intermediate combined filtering/refinement step.

The figure 4.2 indicates the two consecutive stages of a combined, showing a model with a sparse set of states in a coarsened state space (top) filtering out states, and then passing them on to the next model (bottom) which works on a finer, up-sampled version of the state space.

The algorithm procedure is demonstrated as follows.

- For an input x, initializing a basic state space $S_0 = Y_0$ by spatially pooling states in the start space (downsampling the original state space volume).
- Repeat for each combined level $t = 0, \ldots, T-1$:
  - Run sparse, exact inference over S_t using the t^{th} combined model, computing max-marginal scores.
  - Filter states based on max-marginal scores to obtain $\tilde{S}_t$:
    For each i, filter y_i if $mmi < t_x$, a data-dependent threshold.
  - Refine the state space of $\tilde{S}_t$ to obtain S_t for the next combined model.
- Predict with the final level: $y^\star = argmax_{y \in Y_T} s(x, y)$.

The max-marginal scores mmi is obvious one of the main factors for the combined model, which is used to reduce the space state and can be computed with a dynamic programming technique. In a pose estimation model, intuitively, it can be explained as a notion: for a part i at location y_i, the max-marginal highest scoring poses with part i fixed or "pinned" to location y_i. It is important to note that the max-marginal is a global quantity of a complete pose instead of a local pose. If most of the model are convinced that the high max-marginal score is possibly the right location of the part, then there is no difference whether a part has strong individual image evidence or not for the location y_i.

4.4 Threshold

It is essential that we always need to balance the accuracy and efficiency, especially when combining the model. Both minimized error for combining each level and maximized filtered max-marginals are considered during the whole stage. We can use a

common strategy to prune away the lowest ranked states with respect the max-marginal score. To make it clear, we roughly use a ranking threshold according to a data-specific threshold t_x: y_i , which is pruned if $mmi < t_x$, by reducing the max-marginal states which is much lower. The highest score $s_x^\star = \max_y s(x, y)$ and the mean max-marginal score, defined as:

$$\bar{s}_x^\star = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{|Y_i|} \sum_{y_i \in Y_i} mmi. \tag{4.2}$$

it also defines the threshold in this convex combination, the average mmi over all parts and states for each part are also defined. The threshold function is introduced as:

$$t_x(s, \alpha) = \alpha s_x^\star + (1 - \alpha)\bar{s}_x^\star \tag{4.3}$$

We use the $\alpha \in [0, 1]$ as a parameter to determine how to prune aggressively. When the best state is kept, the parameter $\alpha = 1$, which means the best, unconstrained assignment is found. Otherwise, if the median of max-marginals is equal to the mean, or say, if $\alpha = 0$, then about half of the states are reduced. There are two reasons why we choose the particular form of $t_x(s, \alpha)$, the first one is that the image x function can make the threshold fit the problem without difficulties, the second one is that it guarantees a convex learning formulation when max-marginals is sortied and the cut-off is chosen.

With the advantage of using $t_x(s, \alpha)$, which is convex in $s(x, y)$, estimating parameters function will be convex thus will remove the incorrect proportion and reduce the proportion state which is not pruned. The α has the function for controlling the efficiency, thus focusing on learning the parameters θ could minimize errors when the filtering level α is provided.

4.5 Parameters

The objective of learning is to accurately and efficiently prune the states, then fitting parameters and optimizing them to the model w, which are the most important in this task. The classic supervised learning tries to recognise the right from wrong when calculating the w. To some extent, this goal could be easier as the highest score of the correct answer is not necessary for the filtering learning, it is satisfied as long as the

score is higher than the threshold value. In contrast, we assume that the training set as follows for the hard-constraint learning objective: $\{(x^{(j)}, y^{(j)})\}_{j=1}^{m}$:

$$minimize_w \frac{1}{2}||w||_2^2 \tag{4.4}$$

$$subjecttos(x^{(j)}, y^{(j)}) \geq t_{x^{(j)}}(s, \alpha) + 1, \quad \forall\, j \tag{4.5}$$

General speaking, the aim of the equation is to find a regularized set of weights w, after that, it is secured that the score of the correct pose is above the image-adaptive threshold in every training example. By using the following equations:

$$minimize_w \frac{\lambda}{2}||w||_2^2 + \frac{1}{m}\sum_{j=1}^{m}\left[t_{x^{(j)}}(s, \alpha) - s(x^{(j)}, y^{(j)}) + 1\right]_+ \tag{4.6}$$

The hard constraint objective then can be seen as a max-margin structured learning problem with an unconstrained hinge-loss form.

To minimize the errors of the combined model, the definition of $t_x(s, \alpha)$ and the max-marginals are used for the learning process: if $s(x, y) > t_x(s, \alpha)$, then for all i, $mmi > t_x(s, \alpha)$, so there is no part state of y is pruned. To ensure the correct part are not pruned, the condition $s(x, y) > t_x(s, \alpha)$ is necessary when an example (x, y) is given, we use the pruning measure with the max-marginals for justification. According to the probabilistic property, and for arbitrary i, it is not guaranteed that$p(y_i|x)$ is above a threshold with the condition of $p(y|x)$ being above a threshold.

To solve the above problem, a stochastic sub-gradient descent method is introduced. As an example, we use the following equation $\left[t_x(s, \alpha) - s(x, y) + 1\right]_+$ when (x, y) is given, this term need to be non-zero when calculating the sub-gradient.

$$w \leftarrow w + \eta\left[-\lambda w + f(y, x) - \alpha f(y^\star, x) - (1 - \alpha)\bar{f}^\star(x)\right]. \tag{4.7}$$

Where $y^\star = argmax_{y'} s(x, y')$ is the highest scoring assignment,η is a learning rate parameter, and $\bar{f}^\star(x)$ are the average features used by all max-marginal witnesses, $y^\star(y_i)$ will be explained in the max-marginals:

$$\bar{f}^\star(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|Y_i|}\sum_{y_i \in Y_i} f(x, y^\star(y_i)). \tag{4.8}$$

In all max-marginal assignments $y^\star(y_i)$, the last term minus this features will be updated according to the structured perception, which is seen as the primary feature.

The process for combining the model is from coarse to fine, learning sequentially with individual parameters set w and $\mathcal{Y}_i$, as well as α. After refining the unprocessed states, we can get the next level state as we showed before.

The final stage left us nearly nothing, what we need to do is only playing the ground-truth states or similar states with a small and sparse state set. Actually, each part has about 500 states to be handled. The good news is that we have the freedom to adjust the complexity of the model including the pairwise part interaction features, such as appearance and geometry. In addition, the perceptual grouping principles compatibility need to be considered and we are left with about 500 states per part.

4.6 Summary

To sum up, it is worth to indicate that pre-filtering part locations might fail when estimating the human pose because of the primary evaluation for a dense pictorial structure model. So the combination of sparse feature detection and non-max suppression is proposed under the settled threshold output.By reducing the max-marginal values, our method can estimate articulated human pose with a robust first-order model.The individual detection scores might be reduced as there are enough evidence to indicate that strong evidence has been obtained.

Chapter 5

Aggregate Model for Pose Estimation

5.1 Introduction

In the previous chapter, we have known that only using a detection based method to estimate human pose remains some problems, especially the fluctuate of each joint. Such problems will severally affect the estimation in videos. In this chapter, we will focus on not only the task of estimating articulated human pose in videos but also the task of tracking based on some track-able features. The video is basically with some properties of single-view, classic film. These tasks could affect by many factors such as background clutter, motion blur, fast movement, camera resolutions and pose variation, to name a few. To improve the pose estimation and refine the action, we need parse human motion with a reliable semantic retrial. In the literature part, we have shown some related research on articulated pose estimation. Although there are many localization methods for detecting the human joints, the interesting parts such as arms and hands are not stable as these are an area in an image instead of only one coordinate point. So it is more difficult when handling a motion in a video. In addition, current tracking methods heavily rely on the manual initialization for the beginning frames. Therefore, how to combine the tracking and pose estimation to make the novel system can recognise pose accurately without the manual initialization will be the key work. Intuitively, it could be easy for finding a detectable canonical limb when we see a video with a human striking the easy canonical pose. However, in practice, the recorded video with motions is usually short and the poses in the video are challenging. Then it is difficult to improve the pose estimation with single frame parsing. Not only that, when the time is

considered, it is more challenging to estimate the pose and evaluate the accuracy. This is mainly because the joint parsing of articulated body parts over time need to take the intractable inference and learning into account. In addition, the body location variables and the models are in the spaces with high-level state and highly inter-connected So the poorly bias error was introduced when using approximate inference method. Meanwhile, for the learning issue, the computational complexity, always restricts the ability when learning meaningful features, thus lead to a simple location-persistence coupling. It is also noted that it is impossible to recover from poor choices in the initial frames of the video when handling the intractability of preserving a belief distribution. Furthermore, the computation issue could also limit the assumption for geometric features such as limb length. As there is some public dataset provided the videos with many different assumptions and considered the variation of the body type.So we use some public dataset as described in the previous chapter. Both computational and modelling restriction can be overcome after the introduction of the aggregate of tractable models, as we couple the points of body joints with meaningful features over a frame. In addition, it is difficult to use all points of the body joints, which actually are an area, we only use the joint centre instead. We separate the model into several sub-models to let them track each joint in a single frame and model the spatial arrangement. Benefit from the tree structure of these sub-models, more efficient inference and good temporal features can be obtained. This also relies on the image appearance when the colour tracking and optical flow are considered. With the introducing of the large-margin loss, the trained model is discriminative. Compared with the dual decomposition, a single variable could intriguingly enforce an agreement. The experiments could time-costing with a series of inference techniques if the coupling between models is increased.

5.2 Tracking method for Aggregate Model

5.2.1 Introduction

Object tracking is known as locating positions of interest area over time in every frame of a video. Many targets with various features have been researched for different applications. For example, Xiang (Xiang *et al.*, 2016) utilized optical flow and sampled points within the Markov Decision Process framework for tracking pedestrians. Single

person and multiple people could be tracked effectively under the proposed tracking framework. Li (Li *et al.*, 2016) adopted the discriminative feature into a convolutional neural network (CNN) framework for tracking an arbitrary single object after learning its feature online. In addition, the trajectory and the state of the tracking target could also contribute the tracking in many application areas. Besides, object tracking can be applied in many application areas such as human-computer interaction, self-driving vehicles and surveillance system. Many researchers contribute their time for getting a more robust and effective tracking result with a focus on object appearance modelling, model updating, optimizing algorithm and recent hot topic deep learning. Although many different kinds of object tracking algorithms have been studied for several decades, and much progress has been made in recent years, there are still many challenging problems such as fast movement, illumination variation, occlusion, background clutters and proceeding time. Generally speaking, current tracking algorithms are categorized as two methods: generative tracker and discriminative tracker. The noticeable difference between them is how to build an appearance model of the tracking target (Heber *et al.*, 2013). A generative tracker usually focuses on the appearance of a moving object and tries to find a model to represent it. It is unnecessary to consider the background information, which makes the tracker works faster. Online updating method is often used in a case that the appearance changed. However, the change of the object appearance caused by some factors such as occlusion and pose variation makes it more difficult for modelling. Some generative tracking examples can be found from a benchmark paper (Wu *et al.*, 2013). A discriminative method mainly emphasizes how to separate the target from the background in a video scene. Finding a decision boundary between the object and the background is the key issue for a discriminative method. It is well known that the discriminative method works better when enough training samples were given. This tracker works well even though dramatic changes of an object, but it needs more sophisticated calculation, which makes it fail to use in a real-time system as the higher process speed is needed. Due to a large number of features is necessary, the offline feature selection procedure and trained classifier make it difficult to get an arbitrary object type for tracking approaches which need online boosting. Some methods try to combine the generative and discriminative models, which can be often treated as a semi-supervised problem. A common approach learns an online appearance model which can select features from an arbitrary object. The

core idea of a combination method is to predict a classifier with the aim of enlarging the training data after obtaining two independent conditional classifiers from the same data. More detail information about the comparison of discriminative and generative models can be found in (Zilka *et al.*, 2013). Therefore, various representative models become an important research topic. The algorithm framework will also be researched for computing different models. For example, the methods (Avidan, 2007; Brendel & Todorovic, 2009) used the Bayes theorem as a basic framework in this chapter. However, a novel appearance model and the solution method for a model are quite different. Motivated by literature about Bayes framework, which assumes tracking issue as a prediction problem, a MMOT (mixture model of object tracking) has been adopted in this chapter, which combines the colour information and context information when modelling the appearance of an object. In addition, Fourier transform would be also called to compute the object model to make the algorithm run in a real-time system. Also, the algorithm could be integrated into a robot system for motion recognition or behaviour understanding. A typical tracking algorithm consists of four steps: object representation, search mechanism, model solving and model updating. For recent generative trackers and discriminative trackers, both of their key steps is how to acquire a better appearance of an object. There are many papers focus on object information to find the target. Recently, there are several methods utilized context information to handle object tracking which locates the target through finding consistent information of an object. To do so, related data mining method should be introduced for extracting both object and its surrounding region as supplement information, although satisfied results have been obtained, the computational cost is still needed. Not only that, templates and subspace models also contribute to robust performance. Dong et al (Wang *et al.*, 2010) utilized the subspace model, which can handle appearance change while online learning model can learn appearance model in IVT methods. To solve this kind of model, the optimized algorithms (Ross *et al.*, 2007) have been proposed to meet the real-time performance such as proximal gradient approach and the l1-norm related minimization method Yao *et al.* (2013). These methods seem sensitive to partial occlusion according to many experiments. Although some algorithms (Wu *et al.*, 2011) were proposed to manage occlusion while drift might occur because of the offline update of template or offline subspace model. Many researchers have developed the online

updating model which can deal with drift well. However, the scale of an object sometimes changes which poses another challenge for these trackers. For different trackers (Lasserre *et al.*, 2006), scale updating should be considered separately. Compressive tracking (Zhang *et al.*, 2012) method cannot handle scale variance well but introduce a multi-scale information in fast compressive tracking (FCT) (Zhang *et al.*, 2014). However, there is no colour information included for FCT which might fail when the colour of the object and background are similar. Lots of researchers who exploited colour information have achieved an excellent performance in object detection. This method not only handles the similar colour problem but can also locate the first location of an object (Liu *et al.*, 2011). Martin et al. (Danelljan *et al.*, 2014) analysed how the colour information contributes the performance of tracking and the experiments proved the effectiveness compared with CSK tracker and VTD tracker (Kwon & Lee, 2010; Wright *et al.*, 2009). The accessing algorithm standard is used to judge the effectiveness of an object tracking algorithm. The centre error rate and overlap rate are the most common factors for analyzing the tracking method. In our method, advantages of both colour information and Fourier transform are utilized for effectiveness and efficiency.

5.2.2 MMOT Algorithm

Recently, object tracking problem has been treated as a predictive problem which can be solved by the particle filter framework based on the Bayes theorem. The main difference compared with previous traditional particle filter framework is that the number of particles is not needed for solving the model while using a kernel function to obtain the probability needed. When estimating the object location, the object location likelihood is used which is shown as follows:

$$p(x) = p(x|o) \tag{5.1}$$

x is the output vector which includes the predicting object information and represents the current object feature in an image sequence. p(x) can be computed according to the Bayes theory.

$$\begin{aligned} p(x) = p(x|o) &= \sum_{f(\mathbf{z}) \in X^f} p(x, f(\mathbf{z}, o)) \\ &= \sum_{f(\mathbf{z}) \in X^f} p(x|f(\mathbf{z})|o) p(f(\mathbf{z})|o) \end{aligned} \tag{5.2}$$

Then, the problem can be transferred to compute the joint probability. $p(x)$ represents the context feature, $f(z)$ denotes image information including the location and the feature of a target, it can be represented as eq.5.11.

$$M(z) = (V(\mathbf{z}), \mathbf{z}) \tag{5.3}$$

denotes the colour information which adopted the HSV (Hue, Saturation, Value) colour space at location $z(m, n)$,especially the value of V channel (the use of V channel makes the algorithm work well for both colour images and gray-scale images), z belongs to the neighbourhood of location X that includes target object. The target model is defined as z which includes the vectorized image patches centred at pixel position c, the distance between the surrounding pixel and the centre is assigned by applying an isotropic kernel $k(c)$, and then the target model is obtained by computing the value of the colour model histogram, in which the $j - th$ value is:

$$q_j = N_c \sum_{i=1}^{N} k(||c||^2)|\alpha_f| \tag{5.4}$$

where N_c is the normalisation constant to make sure the summation is 1, and α_f is the coefficient of the image patch. α_f is the learning rate, C_f is the covariance matrix of the current frame appearance, and Λ_j is the a $D_1 \times D_2$ diagonal matrix. Then we select a mapping matrix B_1 according to normalised eigenvectors of R_f, which denotes the largest eigenvalue. The mapping matrix is found by the dimensionality reduction technique to get a projection $D_1 \times D_2$ with orthogonal column vectors. As the colour attributes normally have high-dimensional colour features, a dimensionality reduction method is used to make the algorithm preserve useful information after the colour dimensions are reduced dramatically, then the computational time will be decreased. The problem of dimension reduction is formulated to find a mapping for the current frame f, by performing an eigenvalue decomposition of the matrix in eq. 5.12. The framework of the algorithm is described in the following table.

To solve the eq.5.2, two conditional probability should be computed separately.

$$p(x|f(\mathbf{z}), o) = h(x - \mathbf{z}) \tag{5.5}$$

Where h can be seen as a kernel function with respect to the relationship between the centre location of object and its surrounding region. The object location likelihood can

Algorithm 1 The framework of the MMOT method

1. Compute the target appearance with q_j
2. Integrate the appearance into a Bayes framework
3. Compute the condition probability
4. Solve the equation with FFT
5. Update the learning parameters
6. Update the appearance model

be computed through the confidence map:

$$C_m(x) = P(f(z)|o) = ae^{\frac{|z-x^*|^\beta}{\sigma}} \tag{5.6}$$

In eq.5.6, a denotes the normalization constant, σ represents a scale parameter and β define the shape parameter. The confidence map in the eq. 5.7 considers the colour information of the tracking target which improves the challenging problem effectively. The STC method guides us about how to set the parameters of β with some experimental results. Then, take eq. 5.11,5.12,5.5,5.6 into account, the eq. 5.2 can be formulated as:

$$\begin{aligned} p(x) &= \sum_{f(\mathbf{z})\in X^f} h(x-\mathbf{z})V(\mathbf{z})\omega_\sigma(\mathbf{z}-x^*) \\ &= h(x) \bigotimes V(x)\omega_\sigma(x-x^*) \end{aligned} \tag{5.7}$$

As the $\bigotimes$ is a convolution operator, so the Fast Fourier Transform (FFT) can be applied for ensuring the computing speed fast, the location of an object can be determined by

the maximum value of $p(x)$ at the $(t+1)$th frame, which can be represented as:

$$F(be^{|\frac{(x-x^*)}{\alpha}|^\beta}) = F(h(x)) \bigodot F(I(x)\omega_\sigma(x-x^*)) \tag{5.8}$$

Therefore, the appearance model can be obtained by:

$$h(x) = F^{-1}(\frac{F(be^{-|\frac{x-x^*}{\alpha}|^\beta})}{F(I(x)\omega_\sigma(x-x^*))}) \tag{5.9}$$

In addition, it is well known that the visual tracking could fail when the target appearance changes. So it is necessary to update the target model over time. For the MMOT tracker, the appearance model considers the learned target x and the transformed classifier coefficient A computed using the current appearance, and then we use a simple linear interpolation method to update the classifier coefficients:

$$A^t = (1-\rho)A^{t-1} + \rho A \tag{5.10}$$

where t indicates the current frame and ρ means the learning rate parameter, thus a suboptimal problem is introduced. A scheme, allowing the model to be updated without storing the previous target appearances, is introduced to ensure the fast computing speed. Then not all previous frames are considered when computing the current model.

$$A_C^t = (1-\rho)A_D^{t-1} + \rho O^t(O^t + \rho) \tag{5.11}$$

$$x_C^t = (1-\rho)x_C^{t-1} + \rho x_C^t \tag{5.12}$$

O^t is the output of the Fourier transformed kernel, the weight is set with a learning rate ρ, x^t denotes the learned target appearance to calculate the detection scores for the next frame appearance. Therefore, only A_C^t and x^t need to be stored with updating method in above equations.

5.2.3 Tracking Results

We have successfully integrated our method into a real-time system which is used for the task of not only tracking the Autism children but also need tracking some objects that children are grasping. To prove the efficiency of the algorithm, we evaluate our

method on eight challenging image sequences and compare its performance with some other methods which represent the most common tracking framework. For the convenience of comparison, the algorithm is implemented in Matlab and achieves at least 25 frames per second on a PC with Intel E7500 CPU (2.93GHz).

Figure 5.1: The identification of different methods

For the colour information, the tracker normalises the scale values to [-0.5,0.5], which can counter the distortion as an effect of the window operation, thus avoid to affect the kernel. The kernel is introduced as we extend the colour feature to multi-dimensional features, which are extracted from an image patch and it is set to 6 to get the best result. For the map function, we set the parameters β as 1. The learning rate of the algorithm is initially set to 0.05. In order to illustrate the qualitative comparison more clearly, some methods most used to be compared are introduced as they used different object representing methods to locate the very first object appeared in the first frame, and various computing methods are used to solve their models. These methods are described briefly here. The Visual Tracking Decomposition (VTD) method used the observation model, which is decomposed into multiple basic observation models that are constructed by sparse principal component analysis (SPCA) of a set of feature templates. The MIL method put all ambiguous positive and negative samples into bags to learn a discriminative model for tracking. The $L1$ method adopted the holistic representation of the object as the appearance model and then track the object by solving the $L1$ minimization problem. The assessment of several methods above in different situations are shown as below: a. Qualitative and quantitative evaluation Our results show the tracking results of the proposed method and three different algorithms including L1, VTD and MIL in eight diverse images sequences for tracking. These images sequences are extremely challenging because they contain various difficulties for tracking such as occlusion, scale change, similar objects, illumination change, fast motion, camera angles and cluttered background. The tracking rectangle with different colours represents the compared methods which are shown in Figure 5.1. In the sequence of Cliffbar, the L1 and VTD trackers drift away from the object and could not

track the target again when the object is on the top of the book shown in Figure 5.2, the major challenge is the object and background share the similar appearance sometimes. The results show our method performs well even the background information is similar to the target. There is a huge illumination change in the sequence of DavidIndoor which is supposed to be one of the main challenges to track. However, from the Figure 5.3, it is clear to see that all these four trackers can handle the challenges but the MIL method seems more sensitive to the scale change. Our method is adaptive when the light is changing, the camera is moving and the appearance is changing because of the glasses and the face angle.

To the DavidOutdoor in Figure 5.4, the L1 tracker performs the worst after the person appears in the back of a tree and could not track it again. Even though the VTD and MIL tracker fails to track the person when the occlusion occurs but they can track it afterwards. Our method can track the person from the beginning to the end even though the occlusion occurs. The four trackers try to track a girl's face in Figure 5.5, only

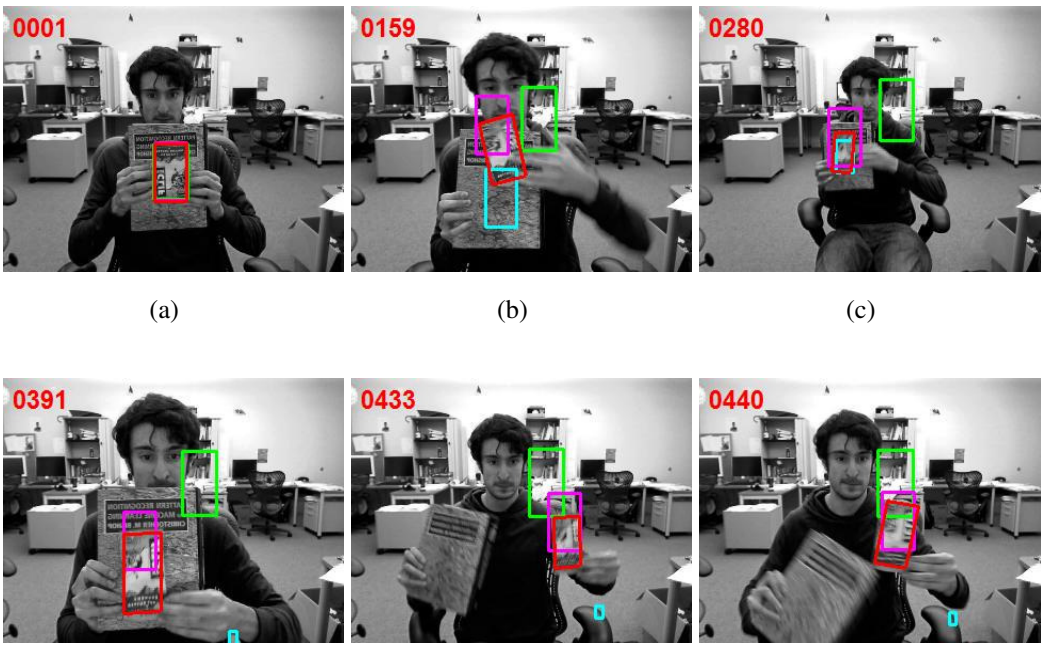


Figure 5.2: The sequences of Cliffbar

our method can track the face from the beginning to the end though the similar face

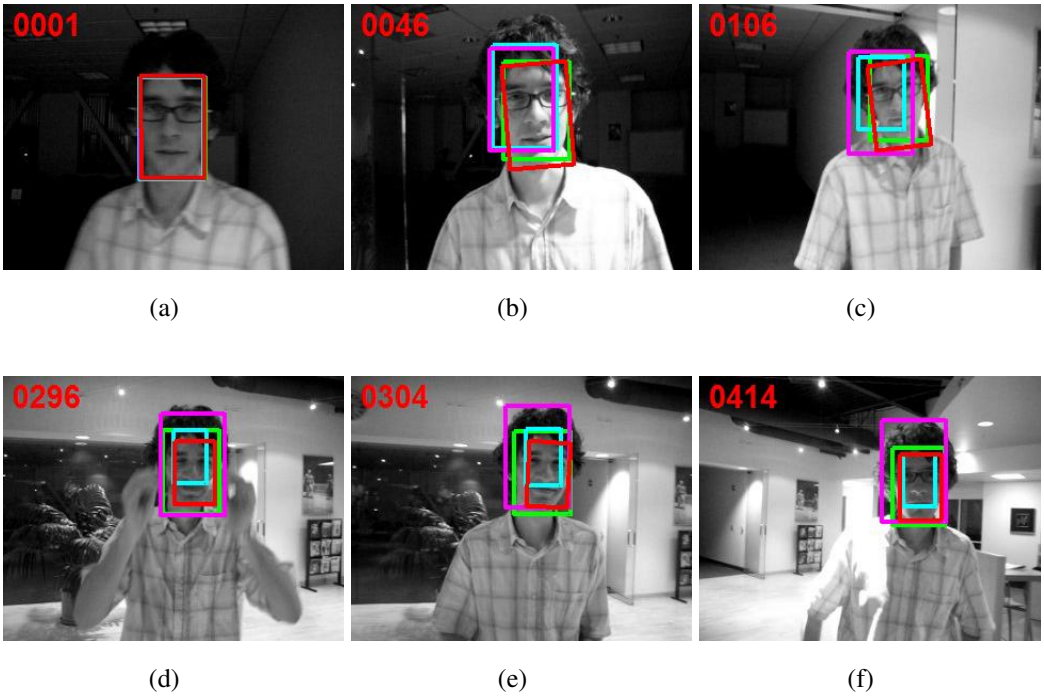
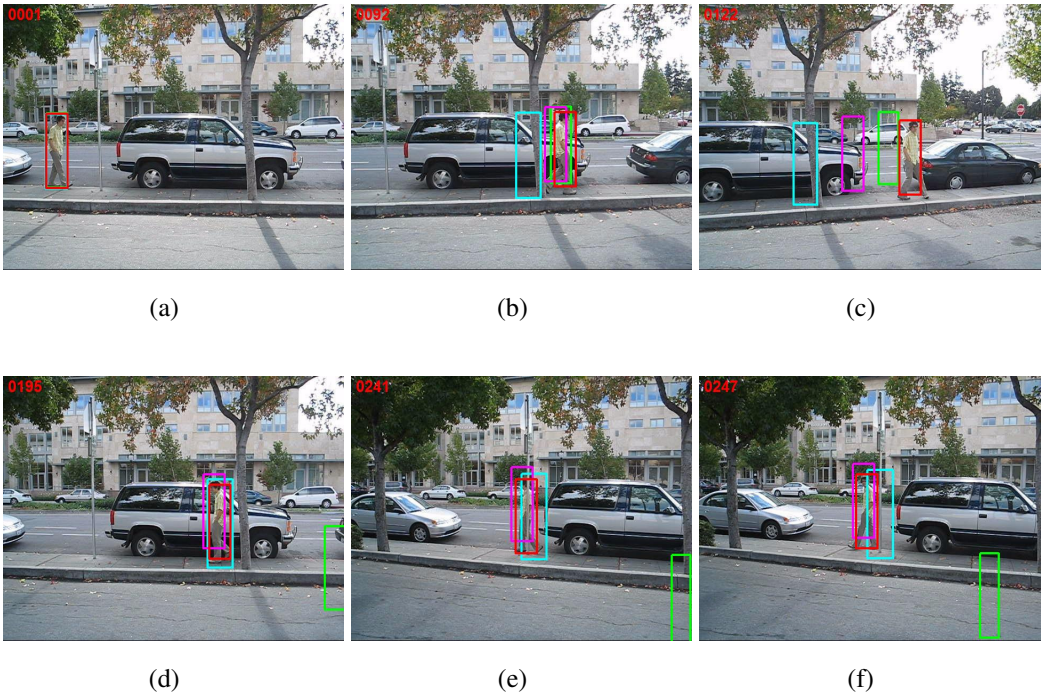


Figure 5.3: The sequences of DavidInDoor

appears and blocks the target. The other three cannot handle this problem and the scale change. But they perform well when only occlusion occurs. The only problem is that they could not locate the target particular accurate when the target changes the angles. Fast motion is an extremely difficult problem in object tracking, both our method and VTD method achieve a satisfied performance all along as shown in Figure 5.10. All these three trackers except ours could not track an indicated object when the object is quite similar with the background in Figure 5.11. Our method has the ability to handle different tracking difficulties no matter they appear individually or in distinct combinations. b. Discussion Although these methods can track the object both for the sequences Davidindoor and Occlusion1 as some occlusion occurs if there is some rotation for the Occlusion2 sequences occurs, only our method performs well. L1 and VTD could not handle the severe occlusion like Girl sequences, while only the VTD the method could not track the object again if there is a drift when tracking, other methods could keep tracking after temporally drift. For the sequences of Cliffbar, the colour of moving object is nearly same with its surrounding region, only our method can keep



(a) (b) (c)

(d) (e) (f)

Figure 5.4: The sequences of DavidOutdoor

tracking over the time, as both colour information and context information was adopted when modelling appearance. So the experimental results show our method is robust to the current tracking challenges including the occlusion and rotation and performs best compares with other methods. In addition, as both centre error evaluation and overlap evaluation, which is defined by the PASCAL VOC, have been used to evaluate the performance of the proposed algorithms. We use the same evaluation criterion. Table 5.1 and Table 5.2 summarizes the experimental results in terms of the average centre error and the average tracking overlap. It is clear to see that our method achieves the lowest tracking errors compared with the others in Table 1, and the highest overlap rate in Table 5.2. The overlap rate is one of the evaluations to verify the tracking success. According to the PASCAL VOC criterion, given the tracking result of each frame R_T and the corresponding ground truth R_G, the $score = \frac{area(R_T \cap R_G)}{area(R_T \cup R_G)}$, indicates the tracking performance. The tracking results are regarded as being valid when the score is over 0.5. The average overlap rate of our tracker is 0.75 while the highest is 0.50 at present. Therefore, our method is valid and better than the other compared

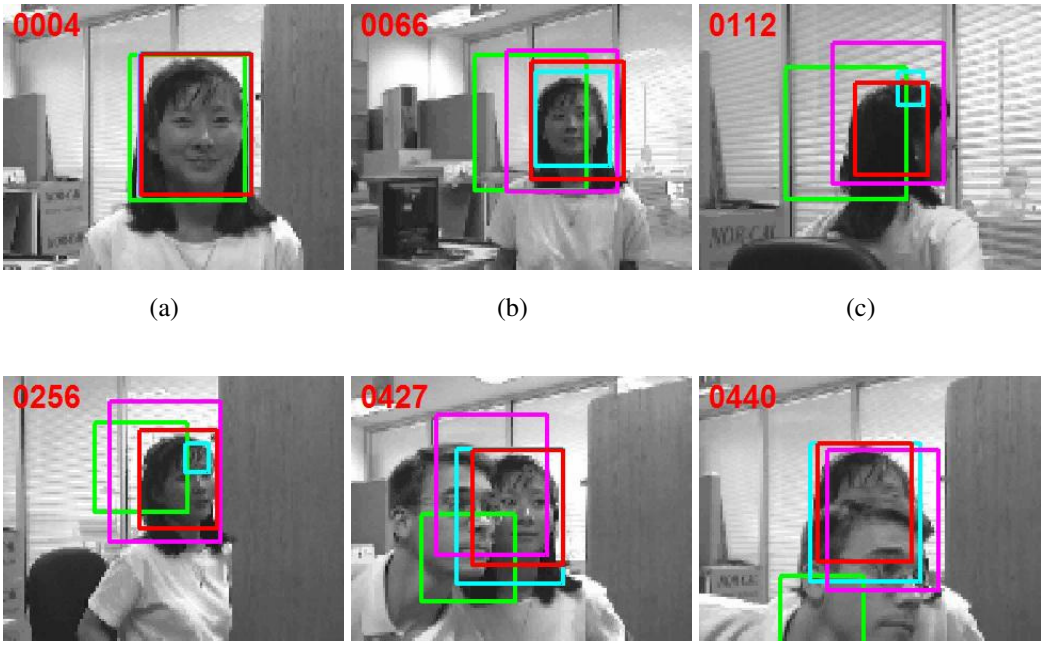
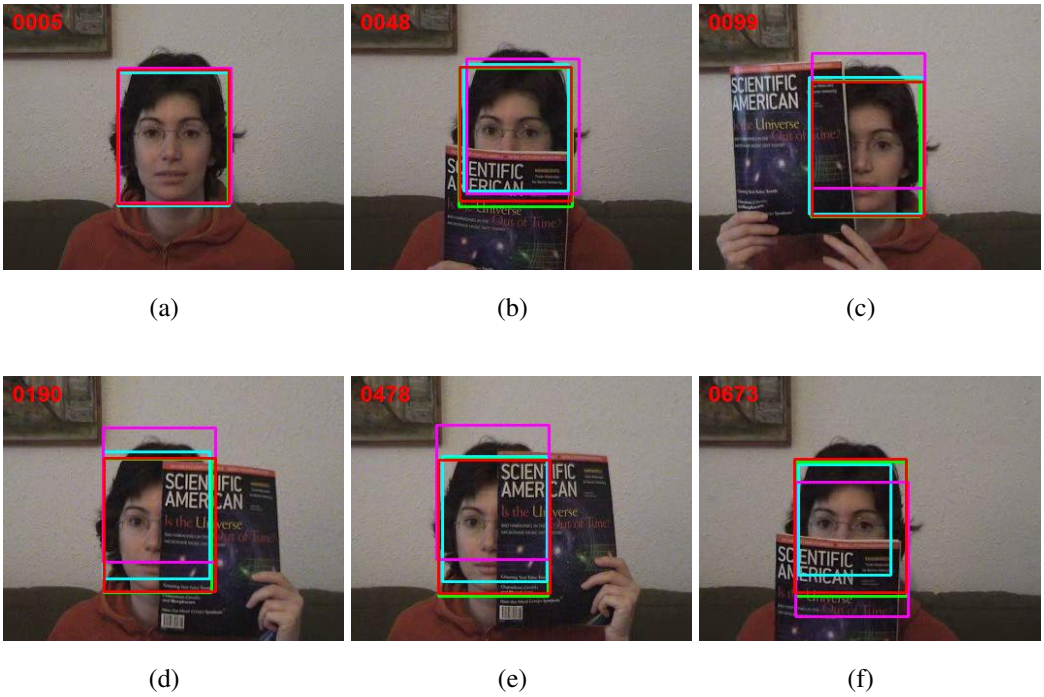


Figure 5.5: The sequences of Girl

methods.

5.3 Tracking-based Modelling

As the proposed method need to track multiple joints over time, some assumptions for representing the pose interaction need to be done before implementing. It is believed that the intractability of joint tracking when modelling the foreshortening should be handled. For example, the interactions between different body parts need to be considered to capture rich and image-dependent relationships.



(a) (b) (c)

(d) (e) (f)

Figure 5.6: The sequences of Occlusion1

5.4 Aggregate Model

5.4.1 Stretchable models of human pose

Among many big limitations, one stand-out drawback of current models of human pose is the "cardboard people" assumption (Ju *et al.*, 1996), which is the fixed size rigid patches up to a global scale body parts. In our previous introduction to the pictorial structure framework. Because of the prohibitive increase in the state space, the human body is completely determined by the partly body pose, which can be represented as fixed length parts collection along with the angle and position of each joint. Thus each part has a certain length, that is to say, the elbows are always a fixed distance away from the shoulder in the posed model. However, because of foreshortening and variation in body type, the assumption is always violated in a realistic video. Based on the pixel coordinates of each joint, we introduce a model directly rather than model human limbs as an orientation and position according to a theory that joints are greater than one limb. Although more variables have been introduced,the state space can be

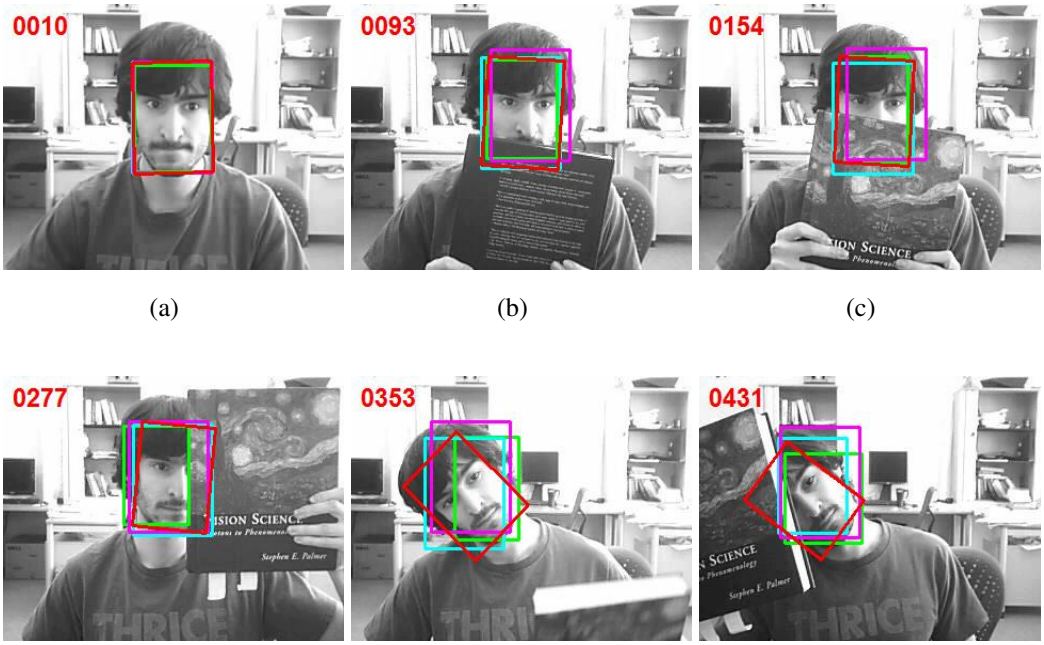
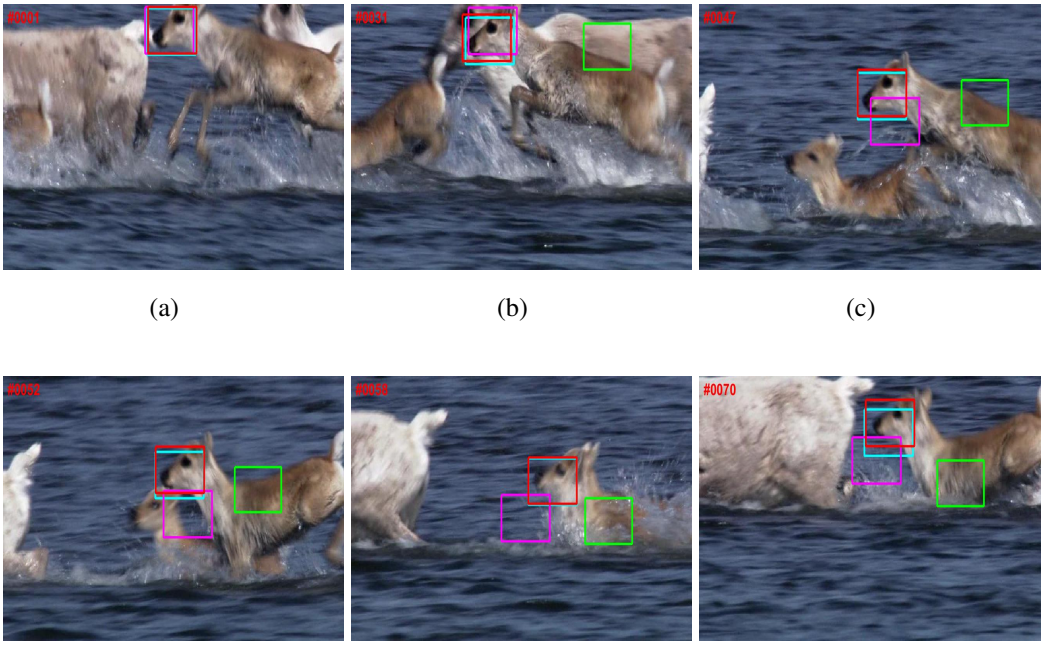


Figure 5.7: The sequences of Occlusion2

reduced dramatically for each variable, this is because in the state space, there is no need to reason over a cautiously discretized set of angles. At the same time, any angle between parts can be represented implicitly. Compared with the PS implementations, a 24 times reduction has been made for each variable in the state space. Moreover, the model lends itself naturally to capture large variability in the part length as the length of the limb could be determined implicitly with the pixel distance between connected joints. Different with the typical rigid and rectangle based representation, the stretchable model has the ability to represent finely discretized limb lengths. However, one of the disadvantages of switching from a limb centric to a joint centric model of body pose is that unary attributes of the lib centric model are now pairwise attributes of a joint centric model. Moreover, we avoid using the pairwise attributes as in a joint-centric model, pairwise attributes in a limb-centric model need to match ternary attributes. However, they are only image-independent functions of geometry in a standard PS model. While in our model, they potential all incorporate image information. Therefore we can benefit an more expressive model compared with a standard PS model.



(a) (b) (c)

(d) (e) (f)

Figure 5.8: The sequences of Deer

5.4.2 Aggregate of stretchable models (ASM)

The relationships between all correlated parts are important when modelling a human motion. This is also important for a recoded data. All these parts are connected kinematically and the parts are left symmetric or right symmetric (i.e. the left elbow is connected with the left shoulder). In addition, the instantiations of the same part appear in continuous frames. For example, the right shoulder at time t and $t+1$ respectively. It is obvious that modelling all relationships together would result in cyclic dependencies within each frame and also consecutive frames. This is because of the three symmetry edges and the tracking edges respectively. In general, however, it is not always impossible to express the score of a given state assignment in an intractable and full model because of the sum of scores under a set of tree sub-models that cover all edge collectively in this full model. It is the core insight which allows us to apply all the rich relationships as expected. Based on the all the interesting relationships of parsing human motion, our model is decomposed of all aggregate sub-models. So they are tractable with the benefit of tree framework. Because the tree sub-model is respon-

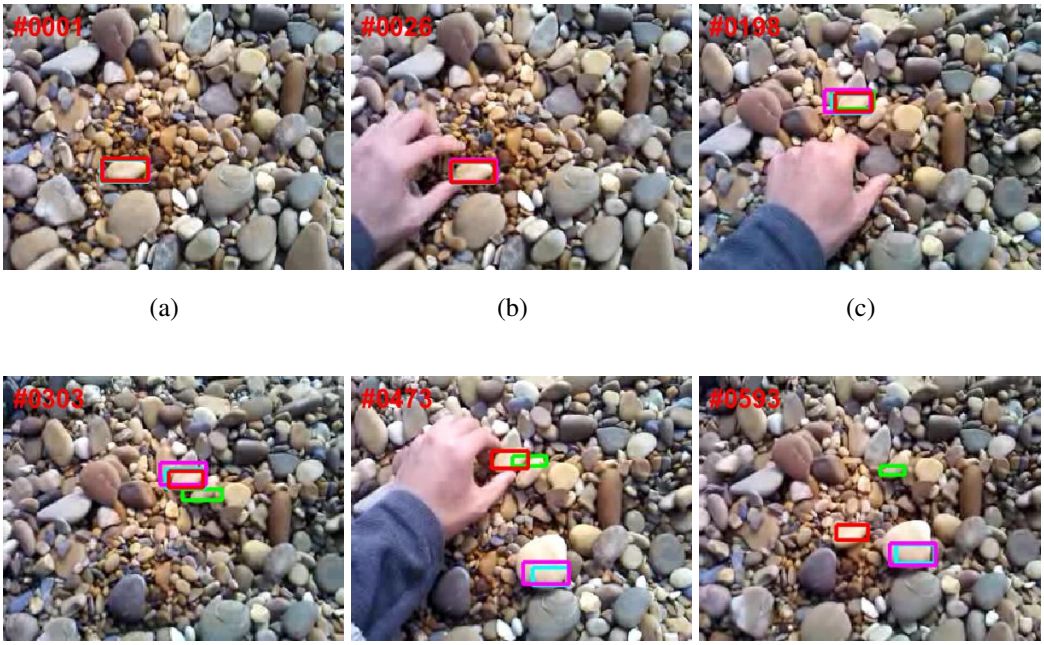


Figure 5.9: The sequences of Stone

sible for tracking a single joint over time, at the same time it models the corresponding set of pairwise interactions between joints for every single frame. Based on above, we form the problem to a structured prediction task, where we use a video sequence with ℓ images as the input x and the output y is a sequence of $n\ell$ variables, each output y_i is the 2D coordinate of some part joint in some frame. The shortcut notation $y_{\mathbf{t}} = \{y_i \mid y_i \text{ is in frame } t\}$ is also introduced to indicate all n joint variables in frame t, in the model, the variable n is the number of joint locations part included, it is noticed that in the joint can be defined as 80×80 in the pixel space. Instead of using the full 80×80 state space for every joint, we make use of Combined Pictorial Structures, which is trained to prune unlikely portions of the state space away in a single frame. This gives us with $|Y_i| \leq 500$ possibilities for each joint in practice. To represent the full graphical pose model, which is defined as $G = (V, E)$, an assumption need to be made first: a general pairwise MRF model can decompose over the vertices V and

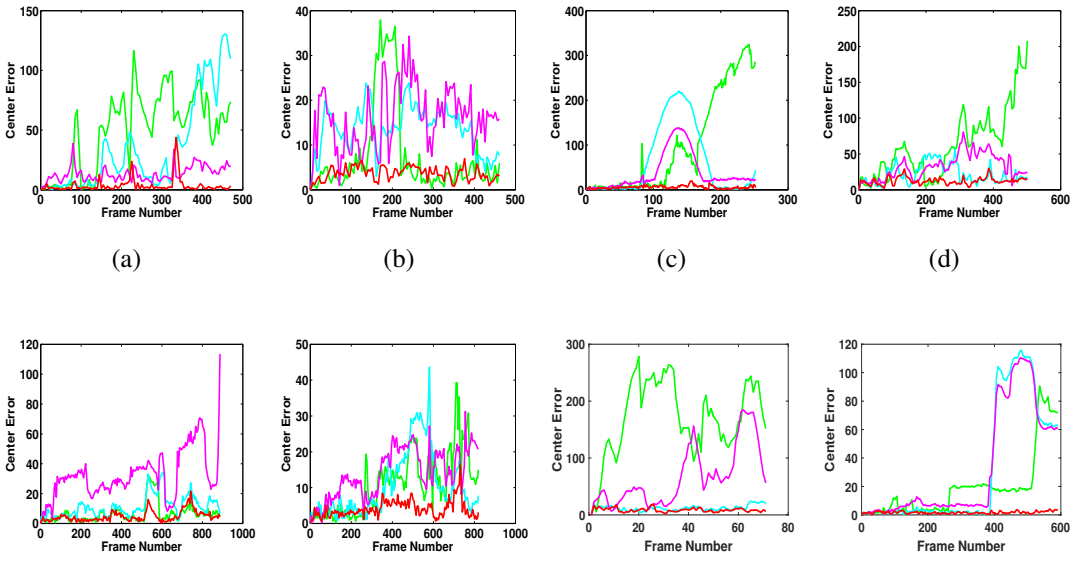


Figure 5.10: The center error result

edges E, then we have:

$$score = \sum_{i \in V} w_i \cdot f_i(x, y_i) + \sum_{(i,j) \in E} w_{ij} \cdot f_{ij}(x, y_i, y_j). \tag{5.13}$$

In the equation, (i, j) is used to represent the edges, which may connect variables between consecutive frames. For representing the model, we use $G_p = (V, E_p)$ as the sub-graph of G corresponding to the p'th one, P to denote the tree sub-models, then the score $score$ is decomposed into the sum of the scores of the P constituent sub-models: $score = \sum_{p=1}^{P} s^p(x, y)$. With the restricted to the edges E_p, the score of the p'th model is can be represented as the following equation:

$$s^p(x, y) = \sum_{i \in V} w_i^p \cdot f_i(x, y_i) + \sum_{(i,j) \in E_p} w_{ij}^p \cdot f_{ij}(x, y_i, y_j). \tag{5.14}$$

Note that parameters across different models w^p will not be coupled so that different models can learn different parameters and affect different behaviours according to strengths and weaknesses dictated by their graph structures.

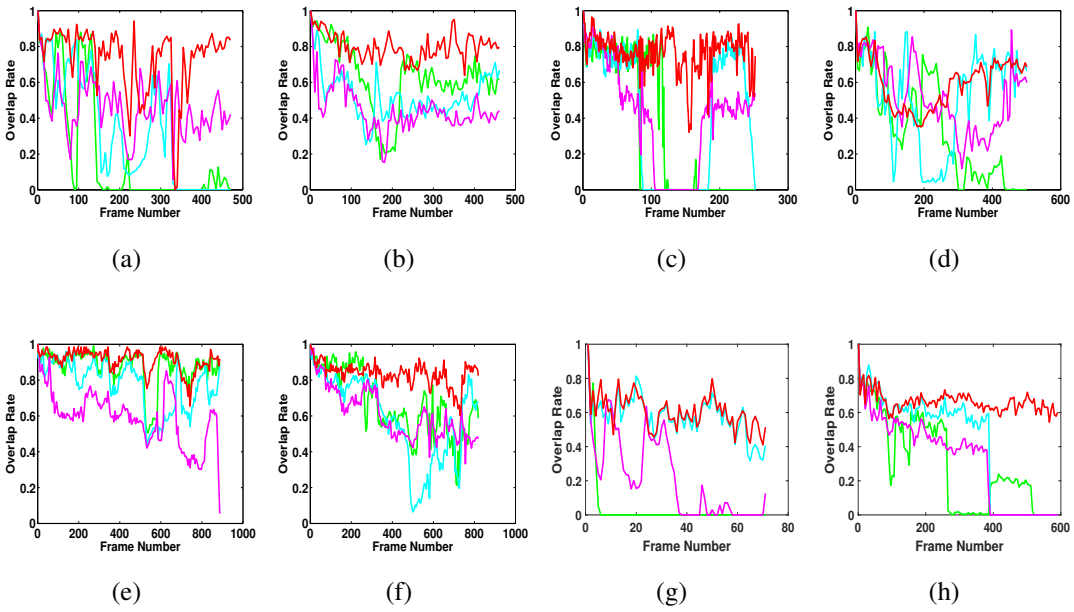


Figure 5.11: The overlap result

5.4.3 Algorithm Inference

To make an effective final decision, the P independent models need to be combined during test time with our explored several methods. A hierarchy of agreement criteria between sub-models can be formed by these methods: the constraint is enforced that all sub-models must agree on the maximizing chore to all variables at one extreme, so the inference process is completely decoupled across sub-models. It should be noted that the existence of the inherent trade-off between the degree of agreement makes the models and the computational cost of the corresponding inference imposed. According to the order of decreasing agreement, the inference method is explained as below:

Full Agreement after Dual Decomposition. With the aggregate of stretchable models, what we want to do is to calculate the argmax by decoding joint locations via the whole sequence of frames. However, because of the high tree width of the cyclic graph, it is prohibitively expensive to solve the argmax decoding problem. To avoid this, the Dual Decomposition (DD) was used to solve a linear programming relaxation of the decoding problem (Bertsekas, 1999; Komodakis *et al.*, 2007). With a global equality

overlap	L1	VTD	MIL	Ours
Cliffbar	24.8	34.6	13.4	3.3
David Indoor	7.6	13.6	16.2	3.2
DavidOutdoor	100.3	61.9	38.3	4.6
Girl	62.4	21.4	32.3	10.2
Occlusion1	6.5	11.1	32.3	3.8
Occlusion2	11.1	10.4	14.1	3.9
Deer	171.5	11.9	66.5	8.3
Stone	19.2	31.3	32.3	1.3
Average	50.4	24.5	30.7	4.8

Table 5.1: The average center error

center error	L1	VTD	MIL	Ours
Cliffbar	0.2	0.3	0.5	0.7
David Indoor	0.7	0.6	0.5	0.8
DavidOutdoor	0.3	0.4	0.4	0.7
Girl	0.3	0.5	0.5	0.6
Occlusion1	0.8	0.7	0.6	0.9
Occlusion2	0.6	0.6	0.6	0.8
Deer	0.04	0.5	0.2	0.7
Stone	0.3	0.4	0.3	0.8
Average	0.37	0.50	0.45	0.75

Table 5.2: The average overlap

constraint, therefore, these problems are coupled and the argmax decoding problem of the full model can be decomposed into the P sub-problems. This can be represented as below:

$$argmax_{y,y^1,\ldots,y^P} \sum_{p=1}^{P} s^p(x, y^p) \text{ s.t. } y^p = y \qquad (DD) \tag{5.15}$$

Now, only the optimization is still intractable due to the integral constraints. Fortunately, the dual optimizing of dual-decompose is tractable if the integrality requirement can be dropped if the return optimal integral solution can be guaranteed. So the

dual problem with sub-gradient descent needs to be solved beforehand. Then the inference can be converged to the exact solution of dual-decompose until all agreement between all y^p is reached. However, it is not guaranteed that if we can have a convergence result. During our experiments, the maximum scoring primal variables found for descent iteration are used only if dual decomposition does not converge with the condition that 500 iterations have finished while all are required in all P models. In addition, the Lagrange multipliers ν^p_{ik} for every possible state assignment $y_i = k$ in every part model are introduced for solving the dual problem with sub-gradient. We then alternate between updating the dual variables ν^p and the primal variables y^p:

$$y^p \leftarrow argmax_y \left(s^p(x, y) + \sum_{i,k} \nu^p_{ik} 1[y_i = k^p] \right) \tag{5.16}$$

$$\nu^p_{ik} \leftarrow \nu_{ik} - \alpha \left(1[y^p_i = k] - \frac{1}{P} \sum_{p=1}^{P} 1[y^p_i = k] \right) \tag{5.17}$$

where α is a rate parameter chosen according to the scheme given in Komodakis *et al.* (2007).

Single Frame Agreement. The inference could be considerably simpler if a subset of the variables in an only single frame is restricted with respect the computing of the MAP decoding, which is also an effective way for finding the argmax solution if the model agreement is constrained. Then for each frame t, the below equation is used to represent the joint configuration for the current frame.

$$argmax_{y'_{ft}} \sum_{p=1}^{P} \max_{y: y_{ft} = y'_{ft}} s^p(x, y) \qquad (SF) \tag{5.18}$$

In the equation, y_{ft} indicates the *set* of n joint variables over frame t. With respect to the constraint, all variables in frame t need to be fixed at positions y'_{ft}, the highest scoring sequence y can be found with the inner maximization. The outer $argmax$ is greater than all possibilities of single frame configurations y'_{ft}. So it can extend the notion of a max-marginal of a variable max-marginals to a max-marginal to many related

variables. For accomplishing this, computing the max-sum messages in a forward-backwards message passing algorithm is the first requirement that incident to the variables in frame t for every sub-models. After that, the argmax decoding of y_{ft} need to be found which is equivalent to inference in a grid with n variables, by the way, in our experiment, we use $n = 6$. For solving this problem further, we form cliques of size to 3, then the passing message can form a clique tree chain. There are at most pairwise potentials in each clique as the state space of each part is relatively small according to the first running AMPE model $|Y_i| \leq 500$. In practice, we can see that the cost of this inference $O(\sum_i |Y_i|^3)$ takes less than a second for each frame. Overall, out experiments took about as twice long as performing inference in all tree sub-model P.

Single Variable Agreement. Based on the idea of (Sapp *et al.*, 2010), the subset of interest could be further down to a single variable at a time. For the model agreement, this seems a weaker criteria, but it yields simpler and cheaper inference, which leaves us the inference problem for the i^{th} variable as follows:

$$argmax_{y'_i} \sum_{p=1}^{P} max_{y:y_i=y'_i} s^p(x, y) \qquad (SV) \tag{5.19}$$

With standard forward-backwards message passing, the question can be solved if the max-marginals are computed for each model. The highest scoring sum is the result what we expect by summing the P max-marginal.It should be noted that in a full model, the result is actually equal to the best assignment decoding when all sub-models comply on the argmax. However, the loopy model occurs very rarely in practise.

Independent / No Agreement The single model is used for comparing the above methods to predict each joint in the aggregate model, which is incorporated temporal dependencies for any specific part, this is the Independent decoding scheme we used.

5.4.4 Parameters

For each model, parameters w^p can be learned separately with decoupled inference method. By doing so, a convex hinge-loss is optimized for each model under the objective. After the agreement was enforced during the inference process was aggregated

at test time, the use of part of a learning procedure becomes prohibitively expensive when at least two variables are trained with the coupling inference.The leaning parameter w^p can be decoupled when we separately use the inference for each model. Therefore, we use an optimization method for a convex thing-loss objective for each model p separately to lean parameters.

$$min_{\theta_m}\frac{\lambda}{2}||w^p||^2+\frac{1}{n}\sum_{j=1}^{m}\Big[max_y s^p(x^{(j)},y)-s^p(x^{(j)},y^{(j)})+1\Big]_+ \tag{5.20}$$

This is optimized through a stochastic sub-gradient descent and for the purpose of minimizing error, the λ is used a held-out development set is also used as the number of training epochs.

5.5 Summary

The method in this chapter combines the colour information and context feature for tracking, which makes it have the robustness to appearance changes of the object. It can work well even though occlusion and similar colour occur. Not only that, scale update information and online update are considered to make it perform better. In addition, it can run in a real-time system as the algorithm computed in frequency domain through Fourier Transform. Qualitative and quantitative experiments prove the effectiveness and efficiency of MMOT algorithm compared with existing methods. The next step of this work will try to compare the results on benchmark sequences with the VOT. Based on the tracking method, we also show a summary depiction of the Stretchable Aggregate process. Because we decompose our cyclic graph, we are able to use a variety of rich image-dependent features for tracking. However, this also requires us to reconcile different tree submodels' beliefs on the final predicted pose. We will demonstrate our results in the next chapter.

Chapter 6

Experiments and Discussion

In this chapter, we first describe the datasets we used, followed by introducing how the datasets have been collected and annotated. In addition, we present the common used evaluation for these dataset. Then we detail the information of how we conduct the experiments introduced in chapter 4 and 5. Meanwhile, the experimental results are demonstrated and evaluated. After that, an overall analysis of experimental results and methods are discussed.

6.1 Dataset

6.1.1 Buffy Stickmen

The buffy stickmen version 2.1 dataset is used in our experiments. This dataset was contributed by (Ferrari *et al.*, 2008), which includes 748 frames from the TV show Buffy the Vampire Slayer, from episodes 2, 4, 5 and 6 from season 5. It is believed that over 50% overlap rate compared with the ground-truth will be acceptable for an upper-body detector.

6.1.2 PASCAL Stickmen

Eichner & Ferrari (2009) contributes this dataset as a part of the PASCAL VOC 2008 challenge. This dataset contains 360 examples (version 1.0) obtained from amateur

photographs. It is used for test only and shares the same protocol with the Buffy stickmen dataset.

6.1.3 MoviePose

In practice, it is necessary to use video dataset to obtain enough information. Few examples of general datasets are far from enough, Even though the normal datasets are bigger and bigger, from hundreds to thousands, the pose distribution is also wider and wider, even the canonical sports poses are added. We still face the shortcomings issue, fortunately, a video-based dataset was contributed from popular Hollywood movies called MoviePose. The dataset was made by automatically running a state-of-the-art people detector algorithm on 30 movies. The labelled ground-truth was obtained from crowdsourcing marketplace Amazon Mechanical Turk with the high confidence when using the detected method. About 5 Turk users labelled 10 upper body joints for the same image. After that, the labelled data was double-checked for getting rid of the bad samples such as occluded examples or images with severely non-frontal. Normally, 80% images are used for training, and the rest is used for testing.

6.1.4 VideoPose

The VideoPose 2.0 was introduced in the background that even some state-of-the-art methods cannot achieve a satisfying result on the previous dataset because of the challenges of a significant portion of frames (30%) with foreshortened lower arms, rapid gesticulation and a highly varied range of poses. The new version, consists of 44 short clips, was hand-selected them to emphasize the natural settings. It is necessary to mention that for each clip, the length is about two to three seconds. Manual annotation is used to fix the issues such as the global scale and translation.

6.1.5 Evaluation

It seems that all these datasets have their own parameters such as the biases. One of the main factors for a dataset is the joint locations distribution, we demonstrate these according to the complexity. That is to say, the Buffy has the least while the

VideoPose has the most. More specifically, the most spread of wrist locations goes to the VideoPose dataset. In addition, it has more foreshortening, the wrist is more likely lie on the elbow because of the rising hand situation occurs more frequently. These issues are obviously caused by the big-capability of the MoviePose, which is five to ten times bigger compared with the other dataset.

For the 2D model, some common assumptions need to keep in mind as the dataset was collected using the pictorial structure for both Buffy and Pascal, where could pluck frames individually. For the video based dataset, the duration, viewpoint, and scale are the main considerations when collecting the data. Each frame has been applied a people detector automatically. For the rising hand situation, especially the hand is over the shoulder, less than samples can be found in the Buffy and Pascal dataset while there is no sample for the VideoPose. Because this motion is not an usual pose for amateur photographs and sitcoms.

The background sometimes could affect the average image bias. For example, the unique background setting makes the Buffy biases plain. The primarily interior settings also contribute this. In contrast, the bias could be higher for a diverse background such as coffee shop, apartment and etcetera, to name a few.

6.1.5.1 Evaluation Measures

To evaluate whether the performance of a method is good or not normally respects the ground-truth, how often the predicted pose matches the ground-truth is the performance index. We also wish to highlight the degree of recovering human joints. However, it is impossible to achieve the pixel-level precision, which can be only achieve based on the ground-truth. In addition, the ground-truth is not the only accurate points which can represent the real joints information. Because the real joint has many pixels which differ according to the resolution and image size. General speaking, the information that using 1 pixel to represent could less than a person's pupil information. Not only that, the uncertainty of human labelling and the variety of human models also make the evaluation tricky. Furthermore, the translation invariant features that all existing methods could not fit the pixel extremely accurate.

To approximately estimate the performance the pose estimation methods, some commonly used strategies are introduced:

6.1.5.2 Root-Mean-Square Error (RMSE)

The root-mean-square error is normally used to measure the accuracy, which can be seen as meaningful in these pose estimation dataset as it can be scale normalized and the ground-truth is made by using an upper-body detection method and should achieve the pixel-level accuracy to some extent.

In our thesis, we also use this method to compare the predicted joints and the ground-truth labelled data on the test data. It should be noted that there is no meaning if the predicted result is far way from the ground-truth. The prediction result could be any value on the image. In another word, if the algorithm tells that the result is wrong, the result could be arbitrarily far away from the true value of the test data based on the ground-truth, it is also likely to skew the RMSE arbitrarily.

6.1.5.3 Pixel error threshold

To avoid the skewing problem the RMSE evaluation method, the Euclidean distance between the predicted result and the ground-truth value is computed for every given threshold in terms of percentage of the prediction of the test joints.The result is scaled according to the torso size in the ground-truth.

Many accuracy results with the thresholds from a wide range are explored, they could be accurate with reasonable distance from the central joint in the ground-truth, or to some extent, a little far from the upper-body centre in pixels accuracy.

From the figure, we can have a good understanding of the accuracy at different processing points based on the performance curve of the result. It is extremely useful when evaluating different systems with their own application purposes. For example, if the system is for the purpose of the activity recognition, then the normal pose joints is required, if the system is designed for the sign language understanding, the near-certainty will be required.

6.1.5.4 Percentage of Correct Parts (PCP)

Apart from the previous two common measurement methods, limb-base evaluation is also used in the pose estimation area. This method judges the correctness by measuring the percentage of correct parts. Specifically, the predicted limb is believed to be right

if the endpoints are in the ground-truth endpoints area, which is normally within the radii, the radii are set as half the length of the limbs from the ground-truth.

The distinctiveness of the common public implementation for measuring the error is that they implemented in a coordinate space with a max-norm, which is aligned to the limbs in the ground-truth. At the same time, arbitrary matching of endpoints could happen, for example, the elbow is predicted as the head. Of course, the matching result could be right which matched the elbow as the elbow on the true limb. It needs to be guaranteed that the length of the true limb is longer than the predicted length.

There are several reasons that this evaluation is more convincing: Firstly,

We prefer not to use PCP because (1) its criteria for a matched guess is too relaxed (2)it is discontinuous (3) it only considers one operating point instead of a range and (4) there are different interpretations of the metric leading to discrepancies in implementations (see the code release of Eichner *et al.* (2010)). However, for historical reasons, many works have reported results in terms of PCP, and here we do the same for compatibility.

6.1.5.5 Competitor Methods

The field of 2D human pose estimation has exploded in the last 5 years. The public datasets we report numbers on are highly competitive, with absolute performance rising each year since 2008. We compare to several competing models. Whenever possible, we report numbers from the publicly available implementations of competitors' code; these numbers are typically different than numbers reported in papers. When the public code is not available, we include PCP measures reported by the authors.

Note that the numbers reported here for our models mean any practitioner should be able to replicate results exactly.

- Mean pose baseline

  One reasonable sanity check is to compare against guessing the average pose in every frame. There is some centrality to the data, such that guessing the mean joint position will get some fraction of joints correct. The mean pose is obtained by empirical averages of joint locations on the training sets; each joint estimated independently.

- Andriluka *et al.* (2009b)

  This is a classical dense PS method. The unary limb detectors are trained Adaboost aggregates of Shape Context on top of Canny edges. The pairwise potentials are standard unimodal geometric displacement costs, computed efficiently with distance transforms.

- Eichner & Ferrari (2009)

  This is a complex system built off of Ramanan & Sminchisescu (2006). The initial unary term is linear filters on image edges. It then iteratively re-parses using colour estimated from the initial parse. It also uses graphcut (Boykov *et al.*, 2001) to rule out some of the background clutter, and post-processing of probabilistic marginals to obtain final limb segments.

- Yang & Ramanan (2011a)

  This recent work uses HoG as its only image cue. Like our models, it is trained jointly in a discriminative learning framework. Contemporary with our proposed Stretchable Aggregate model, Yang & Ramanan (2011a) also use joints as a basic unit of inference. Their work focuses on modelling several appearance modes for each part, which they treat as latent variables to be estimated during training.

6.2 Implementation Details

With the advantage of the coarse-to-fine cascade of the fine-level state space, which has size $80 \times 80 \times 24$, the first level of our cascade coarsens model can down to $10 \times 10 \times 12 = 1200$ states per part from the state-space, we can do exhaustive inference efficiently. Then the $\alpha = 0$ is used for training and pruning. For each stage, half of the states can be throw away with effective learning. In practice, to prune as much as possible while retaining 95% of the ground truth validation hypotheses, we adjust α's per part after a cascade stage is learned via cross-validation error.

One of the dimensions such as angle, width and height, is doubled after pruning, this will be repeated. The standard PS features are only used in the coarse-to-fine stages. After the original state space, the HoG part detectors are run and their outputs

are resized for the coarser state spaces features via max-pooling. The standard relative geometric cues are also introduced in our model. The values of each feature are bin uniformly, thus can add the flexibility to the standard PS model instead of learning a covariance and mean, so we can lean the multi-modal pairwise costs.

We use the Ncut to obtain segments in the final stage. And we use 30 segments for the contour features and 125 segments for the region moments. The result shows that the about 500 hypotheses per part are left by the coarse-to-fine cascade. Then we generate all features for these hypotheses. Features are not evaluated for pairs of part hypotheses which are father than 20 % of the image dimensions from the mean connection location, an additional feature indicates that it is added to the feature set. All unary and pairwise feature for part-pairs are concatenated into a feature vector and learned to boost aggregates. These leverage us the pairwise clique potentials. There are several advantages over stochastic sub-gradient learning for the clique potentials learning. This can determine better thresholds on features than uniform binning with faster training, thus can combine different features in a tree to learn non-linear and complex interactions.

The aggregate of tree models is a collection of six models that captures time persistence of each of the six joints and the left/right symmetric joints edges for both elbows, wrists and shoulders respectively. The decomposition covers all reasonable connections that conceived of modelling, which gives us an opportunity to incorporate all features.

The locations of potential wrist and elbow generated by the coarse-to-fine cascade of AMPE are the input of out method, they are independently for each frame. Typically, 300 to 500 possible elbow and shoulder locations per image are yielded. We project possible joints locations at 4 different lengths for each of these discrete joints orientations predicted by AMPE chose from the 5th, 25th, 50th, and 75th lower arm length quantiles on the training set. Formerly, the top 500 wrist locations scored according to the foreground colour features for every single frame. Therefore, the result is a sparse set of locations for all joints with the higher recall.

—	MoviePose		Buffy v2.1		Pascal Stickmen		all
method	uarms	larms	uarms	larms	uarms	larms	mean
Andriluka *et al.* (2009b)	—	—	79.3	41.2	—	—	60.3
Eichner *et al.* (2010)	90.40	52.85	92.77	53.40	67.92	30.56	57.19
Yang & Ramanan (2011a)	94.05	67.08	92.77	65.53	65.83	37.92	70.53
APS	93.95	52.12	95.11	67.02	86.39	59.17	75.62
LPPS	95.57	71.85	88.94	70.64	68.61	44.58	73.37
mean pose	78.25	33.22	93.19	43.83	72.92	34.03	59.24
mean cluster prediction	95.13	63.73	96.81	70.85	85.83	53.75	77.68

Table 6.1: PCP Evaluation of single frame pose estimation.

6.3 Results

In this section, we go through quantitative results on the datasets Buffy, Pascal and VideoPose, analyzing behaviour and design choices of our methods. We also compare our models—AMPE, ASM and LLPS—against competing models. For much of the analysis, we focus on upper and lower arms only—in particular, elbow and wrist localization accuracy. The reasons for this are that (1) torso and head localization are near-perfect given a detected person (Yang & Ramanan, 2011a), (2) arms are the most interesting parts, involved in actions, hand-held objects and object-person interactions.

In this section we analyze end-to-end system results, using the publicly available code for our systems and competitors, for reproducibility's sake.

6.3.0.1 Single frame pose estimation

The performance of all single-frame models are shown on the MoviePose, Buffy and Pascal datasets in 6.1. The LLPS model outperforms the rest across the three datasets. Yang & Ramanan (2011a) is the closest competitor overall, but AMPE outperforms the others slightly on the most difficult Pascal dataset. Eichner *et al.* (2010) (and Andriluka

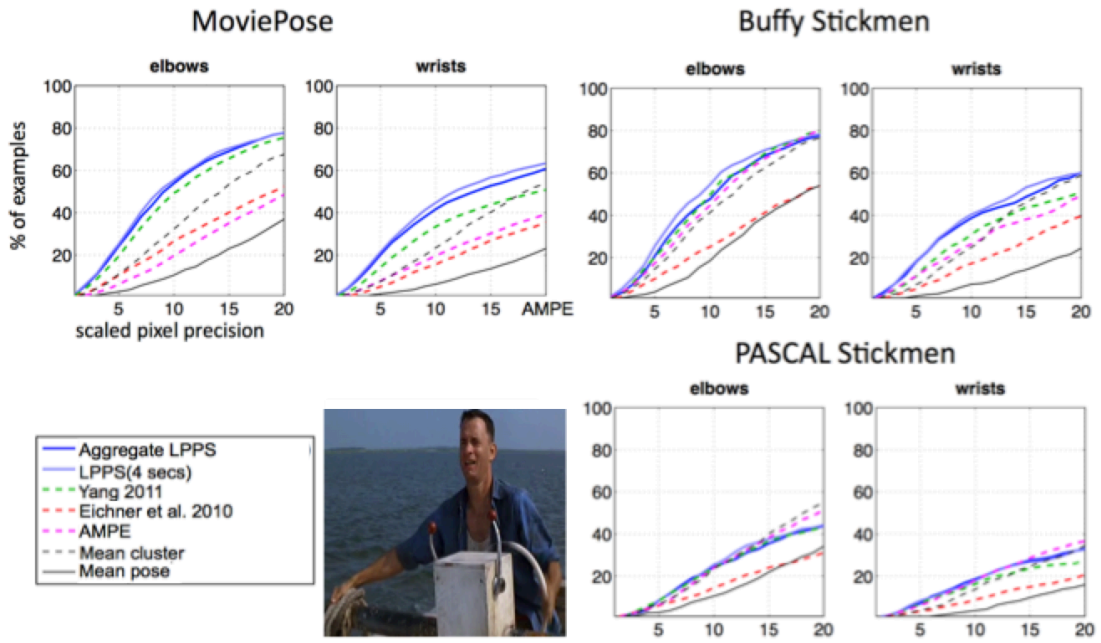


Figure 6.1: Single frame pose estimation results.

.

et al. (2009b), whose code we were unable to run; so it is uniformly worse than the other models, most likely due to the lack of discriminative training and the unimodal modeling. Localization accuracy is not the only way to measure the quality of a model (eg speed)—see 6 for more discussion.

Surprisingly, the two simple prior pose baselines perform comparatively well. The "mean pose" baseline is a lower bound on performance but is competitive with Eichner *et al.* (2010) in some cases. The "mean cluster prediction" baseline actually outperforms or is close to AMPE and Eichner *et al.* (2010) on the three datasets, at the very low computational cost per image. This surprising result shows the importance of multimodal modelling in even the simplest form. The decent performance of these mean pose baseline is also an indication of either the difficulty or lack of pose variation in these datasets, or a combination of both. In fact, the scatterplots do show that most

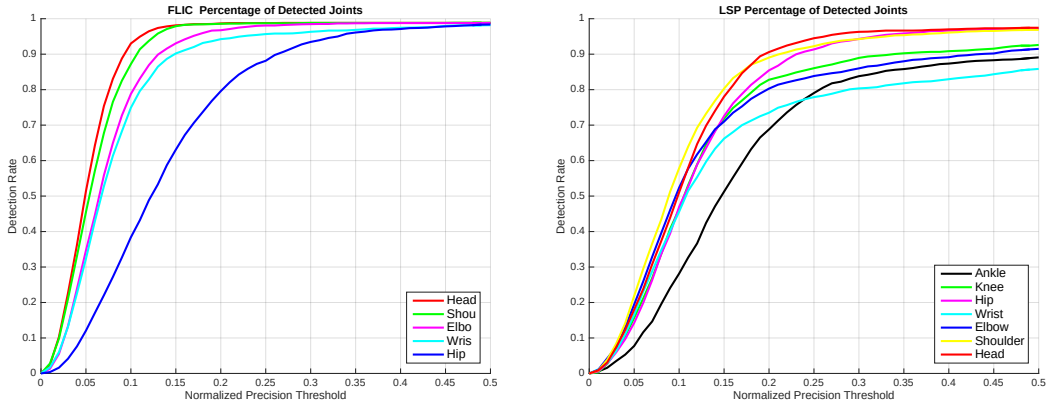


Figure 6.2: upper body pose estimation (left) and whole body pose estimation (right)

elbows are very tightly grouped in these single frame datasets.

The figure 6.2 shows our methods' performance on both upper body and whole body datasets.

Quantitative results for pose estimation in video are shown in figure 6.2 . All three methods we explore for inference in our pose estimation video model (Aggregate of Stretchable Models) outperform the state-of-the-art single-frame methods by a significant margin. Using just a single one of our Stretchable Model trees already does significantly better than single frame models. This shows the usefulness of our stretchable model (joint-centric) representation of pose, as well as some of the rich pairwise interactions we use that other model do not. It is important to note that previous work has found that incorporating time persistence into models actually *hurt*

performance (Ferrari *et al.*, 2009; Sapp *et al.*, 2010)—hence single frame models are the most competitive models for which to compare.

6.3.1 Evaluation

We show the progress of our AMPE model's coarse-to-fine combined, in the Buffy dataset. As explained in 6.2, we start with a small state space and continue pruning and refining until we reach a somewhat fine $80 \times 80 \times 24$ grid. At the end of the combined, we are left with on average 492 states for each part, 99.67% fewer states than the original $80 \times 80 \times 24$ state space. We see that after one level of the combined, we have already pruned more than half of the full state space away. This is intuitive because there are many easy decisions of states to reject based on even geometry alone, eg the left elbow does not ever appear in the upper right corner of the person's bounding box.

As the combined progress, we do lose arm hypotheses close to the ground truth arms, seen in the last column of tab c2f. However, the percent of hypotheses close to the ground truth after the combined process is still higher than any current system's accuracy on lower arms (see table res-table). Thus this number (68.4%) is an upper bound on how well we could do with our small, pruned set of states.

To verify that our pruning is better than heuristic pruning, we compare to heuristic pruning in tab c2f, last row. The heuristic is to sample states proportional to their unary potential scores (ie, HoG limb detectors), with non-max suppression. At the same number of states sampled as we obtain from the combined, the heuristic pruning misses 10% lower arm hypotheses.

Finally, it could be the case that the benefits of our rich features in the last stage of AMPE make discrepancies in the accuracy of pruning strategies negligible. In other words, even the pruning heuristic retains 58.6% of lower arms in its hypothesis set, and it is possible that it could perform equally well at final-level prediction when using the same features as AMPE. We see that this is not the case. AMPE performs 5 to 10% better than simple detector pruning coupled with the rich features we use in AMPE. This makes a strong case that the AMPE state filtering strategy is important.

Figure 6.3: Successful estimation result (left) and failed estimation result (right)

Figure 6.4: Result demonstration

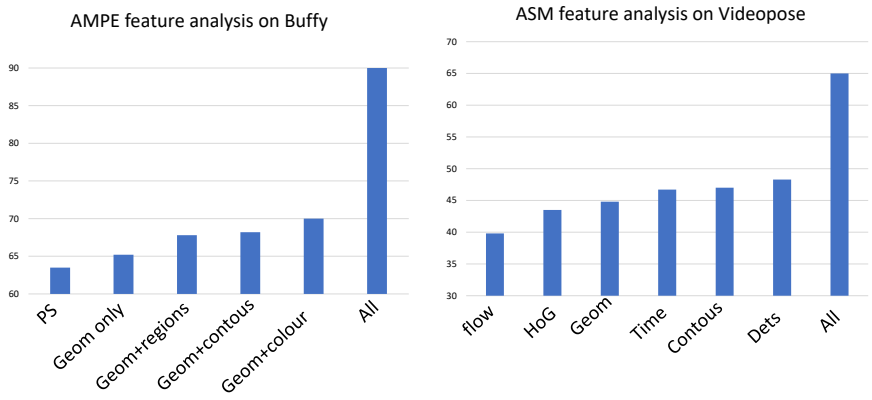


Figure 6.5: Feature analysis. On the left, we observe how adding features contributes to final system performance of AMPE on Buffy, measuring the Area Under the Curve of pixel error distance. On the right, we observe how removing single feature modalities from our Aggregate of Stretchable Models affects performance.

6.3.2 Analysis

One of the main contributions of this thesis is technical innovations that allow us to include "everything and the kitchen sink" (in the house of vision features). At this point, we wish to verify that the work is done implementing, computing and learning parameters for features make a difference in performance. This is actually quite difficult to do in general because it is computationally infeasible to explore the combinatorially many possibilities of feature sets. Non-additive interactions between feature types may occur. We analyze the importance of features by grouping them by modality, and adding or removing them from our full systems, in turn, measuring the change in performance.

In figure 6.5 we do a feature analysis for AMPE (left) and ASM (right), grouping features by coarse modalities. The first important thing to note is that the rich features

we include over standard edge template and geometry information lead to better performance. For AMPE, including rich feature types individually in conjunction with geometry features helps performance, in particular, pairwise cues such as colour and contours that were infeasible to compute without our combined approach. In the bar marked "baseline PS" we evaluate a classical unimodal PS, whereas the geometric parameters in our combined models ("geom only") are learned bin weights that can achieve non-linearity. All features together do significantly better than anyone feature modality in isolation.

On the right side of figure 6.5, we do an ablative analysis of our ASM model. The most important individual feature modality is optical flow, which gives us a fairly good estimate of foreground/background separation in many video frames. Importantly, many of these feature modalities are not used in pose estimation models because they require joint interactions which lead to loopy, cyclic models.

6.4 Discussion

In this chapter, both advantages and disadvantages of the different models developed are discussed. After that, some other design choices which we believe of importance are reported such as possible criticisms and other trade-offs. The results show the performance of different models in terms of part localization. in 6.3.

6.4.1 Image-dependent interactions

From the previous description, we can find that a set of useful features are significantly better than single feature pose estimation approaches. The use of image-dependent pairwise cues contributes the performance of the pose estimation. Most of the existing systems used pre-processing methods for reducing the state space by estimating foreground colour, all of these models benefit from the edge-based potentials or some general geometric pairwise cues. One of the main reasons that our method performs well is the coarser nature of our features, it can find the near contour or the foreground colour accurately compared with the normal features such as HoG cell.

From some experiments, image-dependent interaction feature could significantly advance the model when it is combined with some classical model, such as contour continuity. Therefore, we have sufficient evidence to say that image-dependent interactions are useful for improving the model performance.

The image-dependent interactions feature, on one hand, can improve the performance of the pose estimation model, on another hand, it can be combined with some aggregate models and further reinforced with fewer interactions when exploiting the variety of body or video clip.

We believe using our combined approach will help researchers design models with larger image-dependent interactions. It is a key component to all models presented here to achieve a high degree of accuracy and/or efficiency. Our proposed framework could contribute more robust performance for building rich models in the future. Generally speaking, The freedom from unlimited pairwise potentials will advance the pose estimation model, this advantage will provide much more flexibility when incorporating more higher-level features.

6.4.2 The balance of features and models

In general, two main approaches for improving the performance of a system are features and models. Both adding more features or building more complex models can significantly advance the human pose estimation system.

LLPS and Yang & Ramanan (2011a) are both multimodal models that outperform the feature- and computation-heavy AMPE. It is unclear given the current state-of-the-art if the multimodal HoG model approach is yet saturated, or more data and more nuanced mode definitions will continue to yield increased performance in coming years. Zhu *et al.* (2012) has an excellent study of whether these models are near saturation; see 6.6 for our own trend analysis. Both suggest that we are not near saturation yet, in terms of number of parts, modes and training data. There is also no limit to the performance gains to be had by adding more and better features, only computational hurdles.

Ultimately, these research directions are complementary, and an ideal model would use a combination of rich features and robust models. Each additional feature type (eg, segments, contours, optical flow, depth) incurs an additional cost to obtain but adds to

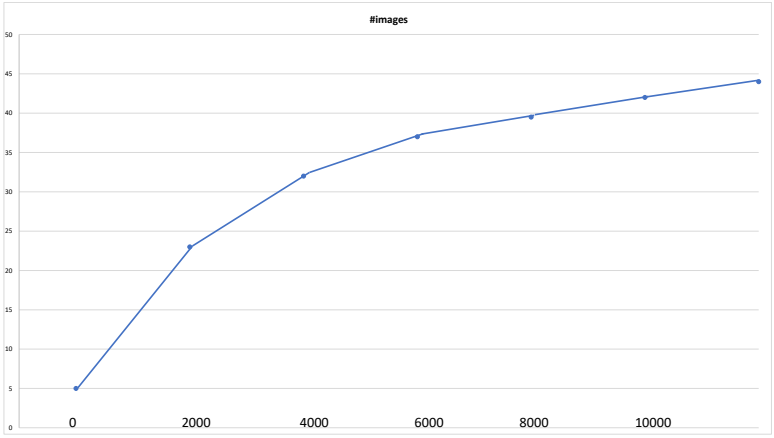


Figure 6.6: LLPS learning curve: test set accuracy for wrists and elbows combined for LLPS models trained with different amounts of data. The mode definitions, number of modes, and cascaded mode prediction step were held fixed and used for training.

the generalization capabilities of the model. Additional modes allow for more specific modelling of different scenarios, but require more training data to estimate parameters accurately. This leaves a large space of possible models combining the two approaches.

6.4.3 Feature selection

The ASM model introduced a joint-based 2D representation of pose. At the same time, Yang & Ramanan (2011a) also introduced a model based on joints and limb midpoints as basic units of inference. This approach has clear benefits for easily capturing foreshortening and scale. It has the seeming disadvantage of not being able to capture limb-pair features with a pairwise model. However, this is not a fundamental limitation. Especially using cascaded inference techniques we should not shy away from describing higher order cliques in the future. In general, the scope of the basic atomic unit for inference (the inference variables) need not be dictated by the scope of the largest clique we capture in our model. Joint-based models are worth exploring further. In particular, we expect an Aggregate of Stretchable Models approach applied to *single frame* pose estimation to work well.

6.4.4 Model selection

Some pose estimation systems advertise that they work well for both person detection *and* pose estimation—in particular Andriluka *et al.* (2009b) and Yang & Ramanan (2011a). One system that does both is beneficial in its simplicity—one function to both find a person and find their body parts. Our approach, on the other hand, requires first detecting a person with a dedicated person detector and then running our pose models on the detected person.

We believe that detection and localization are fundamentally different tasks and should be decoupled. In detection, we wish to generalize over all poses and determine how to discriminate any pose from background clutter—a detection is correct even when a pose is incorrect, and a detector must also have some notion of global confidence to determine, over all possible image patches, whether it is a person or not. A pose estimator works under the assumption that a pose is present, and is correct only if it predicts the right pose versus combinatorially many wrong poses.

One model that attempts to perform both tasks is bound to perform only as well as it could be tuned to each task independently, and probably worse. It may be that PS models are the right *family* of models for both detection and localization—they have attractive benefits for generalizing over poses with deformations and obstructions in addition to localizing pose—and they should be used for both. However, models should be trained evaluated specifically to a single task.

One of the selling points of our models is the ability to include a multitude of features. The goal is to include as many feature modalities as possible in our AMPE and ASM models. Having so many features makes it difficult to determine exactly what is contributing to the success of our model.

From a machine learning standpoint, this is an attractive aspect of our system: given training data, we can try everything and see what works. From a computer visionist's (or perceptual scientist's) perspective, this is a disadvantage—it is difficult to gain insight into why the model is performing well.

We take a functional, application-driven approach towards computer vision, and consider our problem one of engineering rather than perceptual science. The inability to measure the individual performance of components in any complex system is

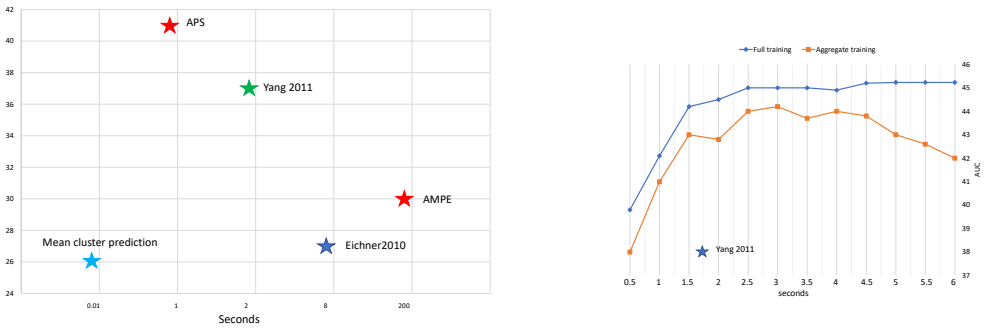


Figure 6.7: The comparison between the test time speed and accuracy

inevitable—the whole is greater than the sum of its parts. We provide individual feature analysis and make convincing arguments that the features and interactions we include are beneficial. We make no statement as to which features are the "best", in any sense other than their contribution to final system performance.

When developing our AMPE and ASM models, we focused our attention on obtaining the best performing, computationally tractable system. Besides raw performance, practitioners care as much about *speed*—does the system run quickly?—and *simplicity*—how long does it take to download, compile, understand, run and/or reimplement? One of the motivations of LLPS and attractiveness of Yang & Ramanan (2011a)'s model is its speed and simplicity, in terms of image features (only HoG) and lines of code.

LLPS is strictly better than other models in terms of speed and simplicity, according to 6.4.4. Yang & Ramanan (2011a) is strictly better than all but LLPS. Among the other models, there is no clear winner—AMPE is more accurate but slower. Predicting cluster means is extremely fast but less accurate.

Chapter 7

Conclusion and Future direction

7.1 Conclusion

This thesis mainly focuses on the performance of human pose estimation with improved features and models. Some previously proposed methods have their own advantages but fail to achieve reasonable results on our specific datasets, which motivates the research on exploring the balance between different features and models, namely the pairwise method. It has the advantage of combining different advantages and produces better performance.

We drive the past induction boundary in a few different ways, enabling us to incorporate more extravagant picture subordinate communications. To start with, we proposed the Aggregate Pictorial Structures, a succession of organized models that proficiently prune the state space of conceivable stances down to a reasonable number. This enables us to perform productive correct surmising without confinements. We abuse this by joining an assortment of rich highlights from integral sources, enhancing the best in class PS approaches in the single casing present estimation.

The proposed method can also be extended deal with the estimation tasks in a video. Keeping up a rich arrangement of variable collaborations in a video makes a cyclic system, which is known to require deduction exponential in the number of casings of a video. We keep up tractability using a course venture as in APS, furthermore,

an inexact surmising technique which disintegrates the cyclic structure of associations into a total of tree charts, which catch all the associations of the cyclic system with redundancies. With this Aggregate of Stretchable Models (**ASM**), a rough derivation is just direct in various casings of a video. Moreover, to deal with the fine-grained explanation also, foreshortening impacts regularly present in genuine video cuts, we utilize a joint-based portrayal of posture, instead of an appendage based portrayal, subsequently, our model is ”stretchable”.

Finally, a complementary approach is investigated to the line of research that motivated our methods of APS and ASM. These methods focused on novel computational techniques that allowed us to add more different features and rich interactions into our pose models with their own advantages kept. Experiments have proven that more features would lead to a better performance accuracy on our used datasets and also performs well on other tasks. Different from previously proposed Pictorial Structures, more features are introduced during our experiments which enable us to focus on the nonlinear problem with multimodal nature. Instead, each local neighbourhood is modelled for the PS model.

We show experimentally that our models are best in class on focused open datasets, confirming the value of our demonstrating developments: (1) falls of organized models (2) total of three models (3) joint-based portrayals (4) neighbourhood direct displaying. These thoughts are significant commitments to the field of posture estimation, with the possibility to help in different areas including organized issues too.

Human posture estimation in the wild in its most broad setting is still a long way from a tackled issue. In spite of the fact that we have made huge advances through this exploration. Pushing ahead, we anticipate that further advances should be made with (1) bigger datasets, (2) the computational abilities to scale current ways to deal with a request of extent more information, and (3) accommodating appraisals of posture inside the bigger set of scene understanding.

Future enhancements in posture exactness appear to be encouraging and the fantasy of understanding human posture for an assortment of utilization is progressively convincing given the headways in mechanical autonomy and the inescapability of cameras in our lives. Taking everything into account, the fate of posture estimation is brilliant.

7.2 Future directions

In this chapter, we suggest some future research directioins, inclduing not only pose estimation areas but also more general field such as machine learning and computer vision. Some of them are theoretical, our expectation is a basic building block for those research on related topics.

7.2.1 Potential improvements for Pose estimation

What might it take to announce the present estimation illuminated? Unquestionably with current correctnesses - getting wrists directly about a fraction of the time - we can't guarantee that the cutting edge "works," from a layman's point of view. Taking a useful viewpoint, we consider to call present estimation fathomed when it is prepared to be included in a buyer item in the wild, the way confront location is in cameras and the Kinect now in computer games. We conjecture that precisely restricting elbows and wrists 90% of the time in datasets like MoviePose would be adequate for this dimension of expansive use. Be that as it may, this would at present leave huge opportunity to get better - MoviePose and our present models don't consider dealing with various individuals and their communications or thinking about impediment or very non-frontal non-upstanding stances.

As a side note, our present techniques in posture estimation might just be prepared to be utilized as a second-level strategy in certifiable applications. As in the posture yield isn't depended upon, yet is treated as a non-vital however supportive loud sensor. Wang *et al.* (2011), for instance, use it as a descriptor for activity recognition. The equivalent is improved the situation picture quality(Ferrari *et al.*, 2009) and scene geometry estimation (Delaitre *et al.*, 2012; Gupta *et al.*, 2011).

In what manner may we accomplish higher precision on MoviePose? The externally uninteresting yet most encouraging route forward is utilizing similar models, yet with essentially more information. Our current research about proposes that notwithstanding settling the present arrangement of modes, test set precision has still not immersed as we increment the number of precedents, as appeared in 6.6. In its present frame, LLPS has involved 32 mode models, which cover the scope of human posture in the video great. The most widely recognized mode (arms very still) has 800 preparing

models at its expendable; the least normal modes (arms raised above head and hand held up to confront) have under 25 precedents each. Quickly we see there is space to tissue out a portion of the rarer modes and gauge them better.

As a back-of-the-envelope computation, to gather enough information to prepare the least incessant mode with 500 models, regarding the characteristic mode appropriation gathered from motion pictures, we would require twenty fold the amount of information. Furthermore, despite the fact that 32 modes cover the varieties in the abdominal area present well, we hope to get increasingly precise demonstrating by additionally thinking about methods of appearance. From instinct, it appears to be sensible that every one of our current 32 modes could be part into no less than 3 submodels dependent on appearance. This would raise the count of information expected to multiple times the present dimensions, for the least successive modes to be displayed with 3 submodules, each with 500 preparing models.

It is practically saucy to recommend basically utilizing more information will take care of our issues. To start with, getting and naming this information is no little accomplishment. Second, astute strategies to prepare on such huge datasets should be structured. Innocently utilizing current preparing strategies for multiple times more information would result in 13 days to prepare and would require 1.4TB of memory; only marginally out of the domain of attainability for current standard servers.

The last issue is that scaling up to a bigger number of classes constantly tends towards more class perplexity. This is a substantially more contemplated issue in huge order assignments, for example, the ImageNet challenge (Deng *et al.*, 2010), in which 10,000 distinct classes are to be anticipated. A typical method to manage class perplexity everywhere scale is the various levelled arrangement, eg first anticipate noncreature from the creature, at that point a canine from different creatures, at that point Corgi from Bernese Mountain Dog. The progressive choices are simpler to make and there are less of them than looking at all fine-level classes. This proposes a comparable to way to deal with multimodal present demonstrating: aggregate the 32 modes recursively into 16, 8, 4 and 2 coarse supermodes. The fell expectation could likewise be successfully connected here.

In outline, to push the current LLPS structure as far as possible, we require something like a request of greatness or increasingly extra information, cleverer preparing

calculations (or persistence), and a more extravagant order of modes, some progressively explicit, some broader than our present gathering, in light of both posture and appearance. It is our conviction that such enhancements to our model could appreciate staggering jumps in execution in the coming years. A fundamental report on driving the best in a class of multimodal models with expanding information and modes in different spaces have achieved comparative ends (Zhu *et al.*, 2012).

7.2.2 Limitations

We think pushing multimodal models as far as possible in the coming years will convey a high level of precision to upper body pose estimation. All things being equal, these models fail to impress anyone. Critically, they can't reason about occlusion and numerous individuals.

Some previous work has shown that modelling occlusion probabilistically is limited by the lack of aapearance evidence because of the threshold difficulties. (Wang & Mori, 2008). As we have appeared all through this proposition, the choice to pronounce an arm missing versus simply being hard to identify is very troublesome with current models. In thinking about different individuals, Eichner & Ferrari (2010) investigate the combinatorially numerous conceivable outcomes of recognized individuals in a scene, Kulesza & Taskar (2010) give a system to inspecting a high calibre, assorted arrangement of postures in a picture, and Andriluka & Sigal (2012) show collaborating individuals by associating them together in one tree-organized PS display.

All current models assume that a key missing setting is a context around the person. Pose models only reason the separate part between the pose and the background without understanding the detail of the background. How to explain the scene while lacking the lable information of people is quite difficult issue and it is similar to the task of standard scene understanding. If the detail of the scene can be understood, then both the pose and the backgound should be clear for the task of pose estimation.

References

AANÆS, H., DAHL, A.L. & PEDERSEN, K.S. (2012). Interesting interest points. **97**, 18–35. 20

AGARWAL, A. & TRIGGS, B. (2006). Recovering 3d human pose from monocular images. **28**, 44–58. 9

AGGARWAL, J. & CAI, Q. (1999). Human motion analysis: A review. **73**, 428 – 440. 13

AGGARWAL, J.K. & RYOO, M.S. (2011). Human activity analysis. *ACM Comput. Surv.*, **43**, 1–43. 23

AGGARWAL, J.K. & XIA, L. (2014). Human activity recognition from 3D data: A review. 6

ANDRILUKA, M. & SIGAL, L. (2012). Human context: Modeling human-human interactions for monocular 3d pose estimation. In *Proc. AMDO*. 91

ANDRILUKA, M., ROTH, S. & SCHIELE, B. (2009a). Pictorial structures revisited: People detection and articulated pose estimation. In *2009*, 1014–1021. 14

ANDRILUKA, M., ROTH, S. & SCHIELE, B. (2009b). Pictorial structures revisited: People detection and articulated pose estimation. In *Proc. CVPR*. 73, 75, 85

ANDRILUKA, M., ROTH, S. & SCHIELE, B. (2010). Monocular 3d pose estimation and tracking by detection. In *2010*, 623–630. 15

ARIE-NACHIMSON, M. & BASRI, R. (2009). Constructing implicit 3d shape models for pose estimation. In *12th*, 1341–1348. 13

AVIDAN, S. (2007). Ensemble Tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **29**, 261–271. 48

BAAK, A., MULLER, M., BHARAJ, G., SEIDEL, H.P. & THEOBALT, C. (2011). A data-driven approach for real-time full body pose reconstruction from a depth camera. In *13th*, 1092–1099. 10

BARINOVA, O., LEMPITSKY, V. & KOHLI, P. (2010). On detection of multiple object instances using hough transforms. In *2010*, 2233–2240. 12

BAY, H., ESS, A., TUYTELAARS, T. & GOOL, L.V. (2008). Speeded-up robust features (surf). **110**, 346 – 359. 19, 21

BELAGIANNIS, V. & ZISSERMAN, A. (2016). Recurrent Human Pose Estimation. *Arxiv*, 1–16. 26

BEN-ARIE, J., WANG, Z., PANDIT, P. & RAJARAM, S. (2002). Human activity recognition using multidimensional indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**, 1091–1104. 8

BERTSEKAS, D.P. (1999). *Nonlinear Programming*. Athena Scientific, 2nd edn. 63

BISSACCO, A., YANG, M.H. & SOATTO, S. (2007). Fast human pose estimation using appearance and motion via multi-dimensional boosting regression. In *2007*, 1–8. 1, 9

BOURDEV, L. & MALIK, J. (2009). Poselets: Body part detectors trained using 3d human pose annotations. In *12th*, 1365–1372. 15

BOYKOV, Y., VEKSLER, O. & ZABIH, R. (2001). Fast approximate energy minimization via graph cuts. *PAMI*. 73

BRENDEL, W. & TODOROVIC, S. (2009). Video object segmentation by tracking regions. *ICCV*, 833–840. 48

BURL, M.C., WEBER, M. & PERONA, P. (1998). A probabilistic approach to object recognition using local photometry and global geometry. In *5th*, 628–641, Springer-Verlag. 11

CARRERAS, X., COLLINS, M. & KOO, T. (2008). Tag, dynamic programming, and the perceptron for efficient, feature-rich parsing. In *Proc. CoNLL*. 38

CHEN, L.L., WEI, H. & FERRYMAN, J. (2013a). A survey of human motion analysis using depth imagery. *Pattern Recognition Letters*, **34**, 1995–2006. 7

CHEN, Y., LIU, Z. & ZHANG, Z. (2013b). Tensor-based human body modeling. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 105–112. 24

CHERIAN, A., MAIRAL, J., ALAHARI, K. & SCHMID, C. (2014). Mixing Body-Part Sequences for Human Pose Estimation. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2361–2368. 24

CHERON, G., LAPTEV, I. & SCHMID, C. (2016). P-CNN: Pose-based CNN features for action recognition. *Proceedings of the IEEE International Conference on Computer Vision*, **11-18-Dece**, 3218–3226. 25

CHU, X., OUYANG, W., LI, H. & WANG, X. (2016). Structured feature learning for pose estimation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, **2016-Decem**, 4715–4723. 38

CRIMINISI, A., SHOTTON, J., ROBERTSON, D. & KONUKOGLU, E. (2011). Regression forests for efficient anatomy detection and localization in ct studies. In *Proceedings of the 2010 International MICCAI Conference on Medical Computer Vision: Recognition Techniques and Applications in Medical Imaging*, vol. 6533, 106–117. 15

DALVI, R., HACIHALILOGLU, I. & ABUGHARBIEH, R. (2010). 3D ultrasound volume stitching using phase symmetry and harris corner detection for orthopaedic applications. In *SPIE Medical Imaging*, 762330:1–8. 20, 21

DANELLJAN, M., KHAN, F.S., FELSBERG, M. & VAN DE WEIJER, J. (2014). Adaptive Color Attributes for Real-Time Visual Tracking. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 1090–1097, IEEE. 49

DELAITRE, V., FOUHEY, D., LAPTEV, I., SIVIC, J., GUPTA, A. & EFROS, A. (2012). Scene semantics from long-term observation of people. In *Proc. ECCV*. 89

DENG, J., BERG, A., LI, K. & FEI-FEI, L. (2010). What does classifying more than 10,000 image categories tell us? In *Proc. ECCV*. 90

DONNER, R., BIRNGRUBER, E., STEINER, H., BISCHOF, H. & LANGS, G. (2011). Localization of 3D anatomical structures using random forests and discrete optimization. In *Medical Computer Vision. Recognition Techniques and Applications in Medical Imaging*, vol. 6533, 86–95, Springer Berlin / Heidelberg. 15

DONOSER, M. & BISCHOF, H. (2006). 3d segmentation by maximally stable volumes (msvs). In *2006*, vol. 1, 63–66. 16

DUTAGACI, H., CHEUNG, C.P. & GODIL, A. (2011). Evaluation of 3d interest point detection techniques. In *2011*, 57–64, Eurographics Association. 20

EICHNER, M. & FERRARI, V. (2009). Better appearance models for pictorial structures. In *Proc. BMVC*. 68, 73

EICHNER, M. & FERRARI, V. (2010). We are family: Joint pose estimation of multiple persons. *Proc. ECCV*. 91

EICHNER, M., MARIN-JIMENEZ, M., ZISSERMAN, A. & FERRARI, V. (2010). Articulated human pose estimation and search in (almost) unconstrained still images. Tech. rep., ETH Zurich, D-ITET, BIWI. 72, 75, 76

EICHNER, M., MARIN-JIMENEZ, M., ZISSERMAN, A. & FERRARI, V. (2012). 2d articulated human pose estimation and retrieval in (almost) unconstrained still images. **99**, 190–214. 11, 14

FEI-FEI, L. & PERONA, P. (2005). A bayesian hierarchical model for learning natural scene categories. In *2005*, vol. 2, 524–531 vol. 2. 10, 12

FELZENSZWALB, P. & HUTTENLOCHER, D. (2000). Efficient matching of pictorial structures. In *2000*, vol. 2, 66–73. 14

FELZENSZWALB, P. & HUTTENLOCHER, D. (2005). Pictorial structures for object recognition. **61**, 55–79. 11, 14

FERGUS, R., FEI-FEI, L., PERONA, P. & ZISSERMAN, A. (2005). Learning object categories from google"s image search. In *10th*, ICCV '05, 1816–1823, IEEE Computer Society. 10, 12

FERGUS, R., PERONA, P. & ZISSERMAN, A. (2007). Weakly supervised scale-invariant learning of models for visual recognition. **71**, 273–303. 11

FERRARI, V., MARIN-JIMENEZ, M. & ZISSERMAN, A. (2008). Progressive search space reduction for human pose estimation. In *Proc. CVPR*. 68

FERRARI, V., MARIN-JIMENEZ, M. & ZISSERMAN, A. (2009). Pose search: retrieving people using their pose. In *Proc. CVPR*. 78, 89

FISCHLER, M.A. & BOLLES, R.C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, **24**, 381–395. 14

FISCHLER, M.A. & ELSCHLAGER, R. (1973). The representation and matching of pictorial structures. **C-22**, 67–92. 10, 13

FLEURET, G. & GEMAN, D. (2001). Coarse-to-Fine Face Detection. *IJCV*. 38

FLITTON, G., BRECKON, T. & MEGHERBI BOUALLAGU, N. (2010). Object recognition using 3d sift in complex ct volumes. 11.1–11.12. 12, 16, 20

GALL, J., YAO, A. & VAN GOOL, L. (2010). 2D action recognition serves 3D human pose estimation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6313 LNCS, 425–438, Springer Berlin Heidelberg. 24, 26

GLOMB, P. (2009). Detection of interest points on 3d data: Extending the harris operator. In M. Kurzynski & M. Wozniak, eds., *Computer Recognition Systems 3*, vol. 57, 103–111, Springer Berlin / Heidelberg. 20

GUPTA, A., SATKIN, S., EFROS, A. & HEBERT, M. (2011). From 3d scene geometry to human workspace. In *Proc. CVPR*. 89

HADFIELD, S. & BOWDEN, R. (2013). Hollywood 3D: Recognizing actions in 3D natural scenes. In *2013*, 3398 – 3405, IEEE, IEEE, Portland, Oregon. 20, 21

HARRIS, C. & STEPHENS, M. (1988). A Combined Corner and Edge Detection. In *Proceedings of the 4th Alvey Vision Conference*, 147–151. 17, 21

HASSABALLAH, M., ABDELMGEID, A.A. & ALSHAZLY, H.A. (2016). Image features detection, description and matching. *Studies in Computational Intelligence*, **630**, 11–45. 24

HE, Y. & CHEN, L. (2017). Fast fashion guided clothing image retrieval: Delving deeper into what feature makes fashion. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **10115 LNCS**, 134–149. 26

HEBER, M., GODEC, M., RÜTHER, M., ROTH, P.M. & BISCHOF, H. (2013). Segmentation-based tracking by support fusion. *Computer Vision and Image Understanding*, **117**, 573–586. 47

HOGG, D. (1983). Model-based vision: a program to see a walking person. **1**, 5–20. 13

HU, N., ENGLEBIENNE, G., LOU, Z. & KROSE, B. (2016). Learning to Recognize Human Activities using Soft Labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **XX**, 1–1. 7

HUA, G. & WU, Y. (2007). A decentralized probabilistic approach to articulated body tracking. **108**, 272–283. 14

IOFFE, S. & FORSYTH, D. (1999). Finding people by sampling. In *7th*, 1092–1097. 13

IONESCU, C., LI, F. & SMINCHISESCU, C. (2011). Latent structured models for human pose estimation. In *13th*, 2220–2227. 1, 9

JANOCH, A., KARAYEV, S., JIA, Y., BARRON, J., FRITZ, M., SAENKO, K. & DARRELL, T. (2011). A category-level 3-d object dataset: Putting the kinect to work. In *13th*, 1168–1174. 19

JHUANG, H., GALL, J., ZUFFI, S., SCHMID, C. & BLACK, M.J. (2013). Towards understanding action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, 3192–3199, IEEE. 25

JIANG, H. (2011). Human pose estimation using consistent max covering. **33**, 1911–1918. 9

JOHNSON, S. & EVERINGHAM, M. (2011). Learning effective human pose estimation from inaccurate annotation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1465–1472. 25

JU, S., BLACK, M. & YACOOB, Y. (1996). Cardboard people: A parameterized model of articulated motion. In *Automatic Face and Gesture Recognition*. 58

KADIR, T. & BRADY, M. (2001). Saliency, scale and image description. **45**, 83–105. 11

KNOPP, J., PRASAD, M., WILLEMS, G., TIMOFTE, R. & GOOL, L.V. (2010). Hough transform and 3D SURF for robust three dimensional classification. In *11th*, 589–602, Springer-Verlag, Berlin, Heidelberg. 15, 19, 20, 21

KOELSTRA, S. & PATRAS, I. (2009). The FAST-3D spatio-temporal interest region detector. In *Workshop on Image Analysis for Multimedia Interactive Services*, 242–245. 19, 20, 22

KOMODAKIS, N., PARAGIOS, N. & TZIRITAS, G. (2007). MRF optimization via dual decomposition: Message-passing revisited. In *Proc. ICCV*. 63, 65

KULESZA, A. & TASKAR, B. (2010). Structured determinantal point processes. In *NIPS*. 91

KUMARI, P., MATHEW, L. & SYAL, P. (2017). Increasing trend of wearables and multimodal interface for human activity monitoring: A review. *Biosensors and Bioelectronics*, **90**, 298–307. 24

KWON, J. & LEE, K.M. (2010). Visual tracking decomposition. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1269–1276, IEEE. 49

KYRIAZIS, N., OIKONOMIDIS, I., PANTELERIS, P., MICHEL, D., QAMMAZ, A., MAKRIS, A., TZEVANIDIS, K., DOUVANTZIS, P., RODITAKIS, K. & ARGYROS, A. (2016). A generative approach to tracking hands and their interaction with objects. In *Advances in Intelligent Systems and Computing*, vol. 391, 19–28, Springer. 23

LAI, K. & FOX, D. (2010). Object recognition in 3d point clouds using web data and domain adaptation. **29**, 1019–1037. 19

LAMPERT, C.H., BLASCHKO, M.B. & HOFMANN, T. (2008). Beyond sliding windows: Object localization by efficient subwindow search. In *26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. 27

LAPTEV, I. (2005). On space-time interest points. In *International Journal of Computer Vision*, vol. 64, 107–123. 15, 17, 21, 24

LASSERRE, J.A., BISHOP, C.M. & MINKA, T.P. (2006). Principled hybrids of generative and discriminative models. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, 87–94, IEEE. 49

LEARNED-MILLER, E. (2006). Data driven image models through continuous joint alignment. **28**, 236–250. 13

LEIBE, B., LEONARDIS, A. & SCHIELE, B. (2008). Robust object detection with interleaved categorization and segmentation. **77**, 259–289. 12

LI, H., LI, Y. & PORIKLI, F. (2016). DeepTrack: Learning Discriminative Feature Representations Online for Robust Visual Tracking. *IEEE Transactions on Image Processing*, **25**, 1834–1848. 47

LILLO, I., NIEBLES, J.C. & SOTO, A. (2017). Sparse composition of body poses and atomic actions for human activity recognition in RGB-D videos. *Image and Vision Computing*, **59**, 63–75. 24

LINDEBERG, T. (1998). Feature detection with automatic scale selection. **30**, 79–116. 16, 20

LIU, A.A., SU, Y.T., NIE, W.Z. & KANKANHALLI, M. (2017). Hierarchical Clustering Multi-Task Learning for Joint Human Action Grouping and Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**, 102–114. 7, 23

LIU, B., HUANG, J., YANG, L. & KULIKOWSK, C. (2011). Robust tracking using local sparse appearance model and K-selection. In *CVPR 2011*, 1313–1320, IEEE. 49

LOWE, D.G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. **60**, 91–110. 16, 17

MARR, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, vol. 8. W.H. Freeman. 10

MATAS, J., CHUM, O., URBAN, M. & PAJDLA, T. (2004). Robust wide-baseline stereo from maximally stable extremal regions. **22**, 761 – 767. 16, 21

MICHEL, D., PANAGIOTAKIS, C. & ARGYROS, A.A. (2014). Tracking the articulated motion of the human body with two RGBD cameras. *Machine Vision and Applications*, **26**, 41–54. 24

MICILOTTA, A., ONG, E.J. & BOWDEN, R. (2006). Real-time upper body detection and 3d pose estimation in monoscopic images. In *9th*, vol. 3953, 139–150. 13

MIKOLAJCZYK, K. & SCHMID, C. (2002). An affine invariant interest point detector. In *7th*, vol. 2350, 128–142, Springer Berlin Heidelberg. 22

MIKOLAJCZYK, K. & SCHMID, C. (2004). Scale & affine invariant interest point detectors. **60**, 63–86. 12, 18, 21

MOREELS, P. & PERONA, P. (2007). Evaluation of features detectors and descriptors based on 3d objects. **73**, 263–284. 13

NAVARATNAM, R., FITZGIBBON, A.W. & CIPOLLA, R. (2006). Semi-supervised learning of joint density models for human pose estimation. 70.1–70.10. 1, 13

NOWAK, E., JURIE, F. & TRIGGS, B. (2006). Sampling strategies for bag-of-features image classification. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 3954 LNCS, 490–503. 27

P. FELZENSZWALB, D.M., R. GIRSHICK (2010). Cascade Object Detection with Deformable Part Models. In *Proc. CVPR*. 38

PETROV, S. (2009). *Coarse-to-Fine Natural Language Processing*. Ph.D. thesis, University of California at Bekeley. 38

PHAM, M.T., WOODFORD, O., PERBET, F., MAKI, A., STENGER, B. & CIPOLLA, R. (2011). A new distance for scale-invariant 3D shape recognition and registration. In *13th*, 145–152. 12, 13, 20

PONS-MOLL, G., BAAK, A., GALL, J., LEAL-TAIXE, L., MULLER, M., SEIDEL, H. & ROSENHAHN, B. (2011). Outdoor human motion capture using inverse kinematics and von mises-fisher sampling. In *13th*, 1243–1250. 1, 10, 14

POPPE, R. (2007). Vision-based human motion analysis: An overview. **108**, 4–18. 13

PRASAD, M., KNOPP, J. & GOOL, L.V. (2011). Class-specific 3d localization using constellations of object parts. 34.1–34.11, BMVA Press. 11, 15

RAFI, U., LEIBE, B., GALL, J. & KOSTRIKOV, I. (2016). An Efficient Convolutional Network for Human Pose Estimation. In *Procedings of the British Machine Vision Conference 2016*, 109.1–109.11, British Machine Vision Association. 8

RAJA, K., LAPTEV, I., PEREZ, P. & OISEL, L. (2011). Joint pose estimation and action recognition in image graphs. In *18th*, 25–28. 8

RAMAKRISHNA, V., KANADE, T. & SHEIKH, Y. (2012). Reconstructing 3d human pose from 2d image landmarks. In *12th*, vol. 7575, 573–586. 15

RAMANAN, D. & SMINCHISESCU, C. (2006). Training deformable models for localization. In *Proc. CVPR*. 73

RAMANATHAN, M., YAU, W.Y. & TEOH, E.K. (2014). Human Action Recognition With Video Data: Research and Evaluation Challenges. *Ieee Transactions on Human-Machine Systems*, **44**, 650–663. 6

RIEMENSCHNEIDER, H., DONOSER, M. & BISCHOF, H. (2009). Bag of optical flow volumes for image sequence recognition. 1–11. 15, 21

ROGEZ, G., RIHAN, J., ORRITE-URUUELA, C. & TORR, P. (2012). Fast human pose detection using randomized hierarchical cascades of rejectors. **99**, 25–52. 1, 9

ROSS, D.A., LIM, J., LIN, R.S. & YANG, M.H. (2007). Incremental Learning for Robust Visual Tracking. *International Journal of Computer Vision*, **77**, 125–141. 48

ROSTEN, E., PORTER, R. & DRUMMOND, T. (2010). Faster and better: A machine learning approach to corner detection. **32**, 105–119. 18, 19, 22

ROUDPOSHTI, K.K., NUNES, U. & DIAS, J. (2016). Probabilistic social behavior analysis by exploring body motion-based patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **38**, 1679–1691. 8

RUIZ-ALZOLA, J., KIKINIS, R. & WESTIN, C.F. (2001). Detection of landmarks in multidimensional tensor data. *Signal Processing*, **81**, 2243–2247. 21

RUSSELL, B.C., TORRALBA, A., MURPHY, K.P. & FREEMAN, W.T. (2008). LabelMe: A database and web-based tool for image annotation. *International Journal of Computer Vision*, **77**, 157–173. 25

SALTI, S., TOMBARI, F. & STEFANO, L.D. (2011). A performance evaluation of 3d keypoint detectors. 3DIMPVT '11, 236–243, IEEE Computer Society. 20

SAPP, B., WEISS, D. & TASKAR, B. (2010). Sidestepping intractable inference with structured ensemble cascades. In *NIPS*. 66, 78

SAPP, B., WEISS, D. & TASKAR, B. (2011). Parsing human motion with stretchable models. In *2011*, 1281–1288. 14

SHAHROUDY, A., LIU, J., NG, T.T. & WANG, G. (2016). NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. In *Cvpr*, 1010–1019, IEEE. 24

SHI, F., PETRIU, E. & LAGANIERE, R. (2013). Sampling strategies for real-time action recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2595–2602. 25

SHOTTON, J., BLAKE, A. & CIPOLLA, R. (2008a). Multiscale categorical object recognition using contour fragments. **30**, 1270–1281. 12

SHOTTON, J., JOHNSON, M. & CIPOLLA, R. (2008b). Semantic texton forests for image categorization and segmentation. In *2008*, 1–8. 12

SIGALAS, M., PATERAKI, M. & TRAHANIAS, P. (2016). Full-Body Pose Tracking - The Top View Reprojection Approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **38**, 1569–1582. 7

SIMO-SERRA, E., RAMISA, A., ALENYA, G., TORRAS, C. & MORENO-NOGUER, F. (2012). Single image 3d human pose estimation from noisy observations. In *2012*, 2673–2680. 15

SIPIRAN, I. & BUSTOS, B. (2011). Harris 3D: a robust extension of the harris operator for interest point detection on 3D meshes. *The Visual Computer: International Journal of Computer Graphics - Special Issue on 3DOR 2010*, **27**, 963–976. 20

SIVIC, J., RUSSELL, B., EFROS, A., ZISSERMAN, A. & FREEMAN, W. (2005). Discovering objects and their location in images. In *10th*, vol. 1, 370–377 Vol. 1. 10, 12

SMITH, S.M. & BRADY, J.M. (1997). SUSAN - a new approach to low level image processing. **23**, 45–78. 21

SOROKIN, A. & FORSYTH, D. (2008). Utility data annotation with Amazon Mechanical Turk. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops*. 25

SUN, M., KOHLI, P. & SHOTTON, J. (2012a). Conditional regression forests for human pose estimation. In *2012*, 3394–3401. 14

SUN, M., TELAPROLU, M., LEE, H. & SAVARESE, S. (2012b). An efficient branch-and-bound algorithm for optimal human pose estimation. In *2012*, 1616–1623. 10

TAYLOR, G., HINTON, G.E. & ROWEIS, S. (2007). Modeling human motion using binary latent variables. *Advances in neural information processing systems*, **19**, 1345. 8

TAYLOR, J., SHOTTON, J., SHARP, T. & FITZGIBBON, A. (2012). The Vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 103–110. 10

TUYTELAARS, T. & MIKOLAJCZYK, K. (2008). Local invariant feature detectors: a survey. *Foundations and Trends in Computer Graphics and Vision*, **3**, 177–280. 15, 19

UNNIKRISHNAN, R. & HEBERT, M. (2008). Multi-scale interest regions from unorganized point clouds. In *2008*, 1–8. 20

VEMULAPALLI, R., ARRATE, F. & CHELLAPPA, R. (2014). Human action recognition by representing 3D skeletons as points in a lie group. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 588–595. 24, 25, 27, 28

VIOLA, P. & JONES, M. (2002). Robust real-time object detection. *IJCV*. 38

VRIGKAS, M., NIKOU, C. & KAKADIARIS, I.A. (2015). A Review of Human Activity Recognition Methods. *Frontiers in Robotics and AI*, **2**, 28. 7, 24

WANG, D., LU, H. & CHEN, Y.W. (2010). Incremental MPCA for color object tracking. In *Proceedings - International Conference on Pattern Recognition*, 1751–1754, IEEE. 48

WANG, H., KLÄSER, A., SCHMID, C. & LIU, C.L. (2013). Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, **103**, 60–79. 25

WANG, J., LIU, Z., CHOROWSKI, J., CHEN, Z. & WU, Y. (2012a). Robust 3D action recognition with random occupancy patterns. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7573 LNCS, 872–885. 8

WANG, J., LIU, Z., WU, Y. & YUAN, J. (2012b). Mining actionlet ensemble for action recognition with depth cameras. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1290–1297, IEEE. 7

WANG, Y. & MORI, G. (2008). Multiple tree models for occlusion and spatial constraints in human pose estimation. In *Proc. ECCV*. 91

WANG, Y., TRAN, D. & LIAO, Z. (2011). Learning hierarchical poselets for human parsing. In *Proc. CVPR*. 89

WEBER, M., WELLING, M. & PERONA, P. (2000). Unsupervised learning of models for recognition. In *6th*, ECCV '00, 18–32, Springer-Verlag. 11

WEI, P., ZHAO, Y., ZHENG, N. & ZHU, S.C. (2013). Modeling 4D human-object interactions for event and object recognition. *Proceedings of the IEEE International Conference on Computer Vision*, **XX**, 3272–3279. 7

WEI, X.K. & CHAI, J. (2009). Modeling 3d human poses from uncalibrated monocular images. In *12th*, 1873–1880. 15

WESSEL, R., NOVOTNI, M. & KLEIN, R. (2006). Correspondences between salient points on 3D shapes. In *Proceedings of Vision, Modeling, and Visualization Workshop 2006 (VMV 2006)*, 365–372, Akademische Verlagsgesellschaft Aka GmbH, Berlin. 20

WILLEMS, G., TUYTELAARS, T. & GOOL, L. (2008). An efficient dense and scale-invariant spatio-temporal interest point detector. In *10th*, 650–663, Springer-Verlag, Berlin, Heidelberg. 16, 19, 20, 21

WILLEMS, G., BECKER, J.H., TUYTELAARS, T. & GOOL, L.V. (2009). Exemplar-based action recognition in video. 90.1–90.11. 15

WISKOTT, L., FELLOUS, J.M., KRUGER, N. & VON DER MALSBURG, C. (1997). Face recognition by elastic bunch graph matching. In *1997*, vol. 1, 129–132 vol.1. 13

WOODFORD, O., PHAM, M.T., MAKI, A., PERBET, F. & STENGER, B. (2013). Demisting the hough transform for 3d shape recognition and registration. 1–10. 12

WRIGHT, J., YANG, A., GANESH, A., SASTRY, S. & YI MA (2009). Robust Face Recognition via Sparse Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **31**, 210–227. 49

WU, Y., LING, H., YU, J., LI, F., MEI, X. & CHENG, E. (2011). Blurred target tracking by blur-driven tracker. In *Proceedings of the IEEE International Conference on Computer Vision*, 1100–1107, IEEE. 48

WU, Y., LIM, J. & YANG, M.H. (2013). Online Object Tracking: A Benchmark. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2411–2418, IEEE. 47

XIANG, Y., ALAHI, A. & SAVARESE, S. (2016). Learning to track: Online multi-object tracking by decision making. In *Proceedings of the IEEE International Conference on Computer Vision*, vol. 11-18-Dece, 4705–4713. 46

XIAO, T., LI, S., WANG, B., LIN, L. & WANG, X. (2016). Joint Detection and Identification Feature Learning for Person Search. *arXiv cs.CV*, **4**, 01850. 7

YANG, X. & TIAN, Y.L. (2017). Super Normal Vector for Human Activity Recognition with Depth Cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**, 1028–1039. 7

YANG, Y. & RAMANAN, D. (2011a). Articulated pose estimation using flexible mixtures of parts. In *Proc. CVPR*. 73, 75, 83, 84, 85, 86

YANG, Y. & RAMANAN, D. (2011b). Articulated pose estimation with flexible mixtures-of-parts. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1385–1392. 11, 14

YANG, Y. & RAMANAN, D. (2013). Articulated human detection with flexible mixtures of parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**, 2878–2890. 26

YAO, A., GALL, J. & GOOL, L. (2012). Coupled action recognition and pose estimation from multiple views. **100**, 16–37. 9, 14

YAO, R., SHI, Q., SHEN, C., ZHANG, Y. & VAN DEN HENGEL, A. (2013). Part-based visual tracking with online latent structural learning. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2363–2370, IEEE. 48

YE, M., WANG, X., YANG, R., REN, L. & POLLEFEYS, M. (2011). Accurate 3d pose estimation from a single depth image. In *13th*, 731–738. 10

YU, G., YUAN, J. & LIU, Z. (2015). Propagative hough voting for human activity detection and recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, **25**, 87–98. 25, 34

YU, T.H., KIM, T.K. & CIPOLLA, R. (2010). Real-time action recognition by spatiotemporal semantic and structural forest. 52.1–52.12. 8, 15, 16, 18, 20, 22

YU, Y., MANN, G.K.I. & GOSINE, R.G. (2012). A SINGLE-OBJECT TRACKING METHOD FOR ROBOTS USING OBJECT-BASED VISUAL ATTENTION. *International Journal of Humanoid Robotics*, **09**, 1250030. 8

YUILLE, A., COHEN, D. & HALLINAN, P. (1989). Feature extraction from faces using deformable templates. In *1989*, 104–109. 11

ZAHARESCU, A., BOYER, E., VARANASI, K. & HORAUD, R. (2009). Surface feature detection and description with applications to mesh matching. In *2009*, 373–380. 20, 22

ZHANG, K., ZHANG, L. & YANG, M.H. (2012). Real-Time Compressive Tracking. vol. 7574 LNCS, 864–877. 49

ZHANG, K., ZHANG, L. & YANG, M.H. (2014). Fast Compressive Tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **36**, 2002–2015. 49

ZHENG, S., DONGANG, W. & SHIH-FU, C. (2016). Temporal Action Localization in Untrimmed Videos via Multi-stage CNNs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016*, 1049–1058, IEEE. 7

ZHOU, X., ZHU, M., LEONARDOS, S., DERPANIS, K.G. & DANIILIDIS, K. (2016). Sparseness Meets Deepness: 3D Human Pose Estimation from Monocular Video. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4966–4975, IEEE. 8, 26

ZHU, L., LIN, C., HUANG, H., CHEN, Y. & YUILLE, A.L. (2008). Unsupervised structure learning: Hierarchical recursive composition, suspicious coincidence and competitive exclusion. In *10th*, 759–773. 10

ZHU, X., VONDRICK, C., RAMANAN, D. & FOWLKES, C. (2012). Do we need more training data or better models for object detection? In *Proc. BMVC*. 83, 91

ZHU, Y. & LUCEY, S. (2015). Convolutional sparse coding for trajectory reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **37**, 529–540. 7

ZIAEEFARD, M. & BERGEVIN, R. (2015). Semantic human activity recognition: A literature review. *Pattern Recognition*, **48**, 2329–2345. 7, 24

ZILKA, L., MAREK, D., KORVAS, M. & JURCICEK, F. (2013). Comparison of bayesian discriminative and generative models for dialogue state tracking. *Proceedings of the SIGDIAL 2013 Conference*, 452–456. 48

Appendix A

Publications

A.1 Conference Papers

1. Gao, D., Ju, Z., Cao, J., & Liu, H. (2015, August). Real time object tracking via a mixture model. *24th IEEE International Symposium on In Robot and Human Interactive Communication (RO-MAN)*(pp. 112-116).

2. Gao, D., Ju, Z., Cao, J., & Liu, H. (2016, August). Towards Hand-Object Gesture Extraction from Depth Image. *2016 Joint 8th International Conference on In Soft Computing and Intelligent Systems (SCIS) and 17th International Symposium on Advanced Intelligent Systems* (pp. 311-316)

3. Gao, D., Ju, Z., Cao, J., & Liu, H. (2017). Activity recognition for ASD children based on joints estimation. *IEEE conference on System, Man, and Cybernetics.*

Appendix B

Research Ethics

University of Portsmouth

Certificate of Ethics Review

Project Title:	Image-based human pose estimation
User ID:	748196
Name:	Dongxu Gao
Application Date:	05/11/2017 20:16:36

You must download your certificate, print a copy and keep it as a record of this review.

It is your responsibility to adhere to the University Ethics Policy and any Department/School or professional guidelines in the conduct of your study including relevant guidelines regarding health and safety of researchers and University Health and Safety Policy.

It is also your responsibility to follow University guidance on Data Protection Policy:

- General guidance for all data protection issues
- University Data Protection Policy

You are reminded that as a University of Portsmouth Researcher you are bound by the UKRIO Code of Practice for Research; any breach of this code could lead to action being taken following the University's Procedure for the Investigation of Allegations of Misconduct in Research.

Any changes in the answers to the questions reflecting the design, management or conduct of the research over the course of the project must be notified to the Faculty Ethics Committee. **Any changes that affect the answers given in the questionnaire, not reported to the Faculty Ethics Committee, will invalidate this certificate.**

This ethical review should not be used to infer any comment on the academic merits or methodology of the project. If you have not already done so, you are advised to develop a clear protocol/proposal and ensure that it is independently reviewed by peers or others of appropriate standing. A favourable ethical opinion should not be perceived as permission to proceed with the research; there might be other matters of governance which require further consideration including the agreement of any organisation hosting the research.

GovernanceChecklist
A1-BriefDescriptionOfProject: Human pose estimation has become an active research topic in the field of computer vision. However, there are still some technical challenges because of the complexity of human motion. Although the depth sensors, such as Kinect and Xtion, open up new possibilities of handling with issues, they present some new challenges. In this thesis, we only address human pose estimation frameworks based on colour image and explore the possibility of the tradeoff between effective representing features and models on the public dataset.
A2-Faculty: Technology
A3-VoluntarilyReferToFEC: No

Certificate Code: 8BA9-754C-AB3E-0CC7-10A2-5C44-BF45-1FBD Page 1

A5-AlreadyExternallyReviewed: No
B1-HumanParticipants: No
HumanParticipantsDefinition
B2-HumanParticipantsConfirmation: Yes
C6-SafetyRisksBeyondAssessment: No
SafetyRisksBeyondAssessmentWarning
D2-PhysicalEcologicalDamage: No
PhysicalEcologicalDamageWarning
D4-HistoricalOrCulturalDamage: No
HistoricalOrCulturalDamageWarning
E1-ContentiousOrIllegal: No
ContentiousOrIllegalWarning
E2-SociallySensitiveIssues: No
SociallySensitiveWarning
F1-InvolvesAnimals: No
InvolvesAnimalsWarning
F2-HarmfulToThirdParties: No
HarmfulToThirdPartiesWarning
G1-ConfirmReadEthicsPolicy: Confirmed
G2-ConfirmReadUKRIOCodeOfPractice: Confirmed
G3-ConfirmReadConcordatToSupportResearchIntegrity: Confirmed
G4-ConfirmedCorrectInformation: Confirmed

FORM UPR16

Research Ethics Review Checklist

University of Portsmouth

Please include this completed form as an appendix to your thesis (see the Postgraduate Research Student Handbook for more information

Postgraduate Research Student (PGRS) Information			**Student ID:**	**748196**
Candidate Name:	**Dongxu Gao**			
Department:	**School of Computing**	**First Supervisor:**	**Zhaojie Ju**	
Start Date: (or progression date for Prof Doc students)	**1 February 2015**			

Study Mode and Route:	Part-time ❑	MPhil ❑	Integrated Doctorate (NewRoute) ❑
	Full-time x	MD ❑	Prof Doc (PD) ❑
		PhD x	

Title of Thesis:	**Image-based human pose estimation**
Thesis Word Count: (excluding ancillary data)	**32,171**

If you are unsure about any of the following, please contact the local representative on your Faculty Ethics Committee for advice. Please note that it is your responsibility to follow the University's Ethics Policy and any relevant University, academic or professional guidelines in the conduct of your study

Although the Ethics Committee may have given your study a favourable opinion, the final responsibility for the ethical conduct of this work lies with the researcher(s).

UKRIO Finished Research Checklist: (If you would like to know more about the checklist, please see your Faculty or Departmental Ethics Committee rep or see the online version of the full checklist at: http://www.ukrio.org/what-we-do/code-of-practice-for-research/)	
a) Have all of your research and findings been reported accurately, honestly and within a reasonable time frame?	**YES**
b) Have all contributions to knowledge been acknowledged?	**YES**
c) Have you complied with all agreements relating to intellectual property, publication and authorship?	**YES**
d) Has your research data been retained in a secure and accessible form and will it remain so for the required duration?	**YES**
e) Does your research comply with all legal, ethical, and contractual requirements?	**YES**

*Delete as appropriate

Candidate Statement:	
I have considered the ethical dimensions of the above named research project, and have successfully obtained the necessary ethical approval(s)	
Ethical review number(s) from Faculty Ethics Committee (or from NRES/SCREC):	
Signed: ***(Student)***	**Date:**
If you have *not* submitted your work for ethical review, and/or you have answered 'No' to one or more of questions a) to e), please explain why this is so:	
Signed: ***(Student)***	**Date:**