Proceedings of the 33rd Chinese Control Conference
July 28-30, 2014, Nanjing, China

Cheating Behavior Detection based-on Pictorial Structure Model

Le Lv[1], Dongbin Zhao[1], Zhijiang Fan[2]

1. The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China.

2. USTECH(Beijing) Technology Co. Ltd.
E-mails: iamlvle@126.com, dongbin.zhao@ia.ac.cn,

Abstract: The video surveillance system is widely used in various areas. However, it is an intractable problem to process a large number of video data manually. There are many intelligent systems developed to analyze human behavior and detect unusual events, however, few concerns the cheating behavior in examinations. In this paper, we use video surveillance system to detect cheating behavior in the examination. The pose estimation method is presented to detect the cheating behavior, based on the analysis of the implementation environment. More in detail, the pictorial structure is adopted to model the human body and a required pose and skin color information is utilized to build the human appearance model. Finally, the belief propagation algorithm is used to infer the maximum a posterior pose. The experiment shows the effectiveness of our method.

Key Words: pose estimation, cheating behavior detection, pictorial structure model, belief propagation

1 Introduction

Automatically detect cheating behavior during an examination will release the human labor greatly. Actually, it is usually thought of as a human behavior analysis task based on human pose estimation. There are many research works focus on human behavior analysis. According to different cognitive levels, these researches can be classified into three categories [1]. The simplest approaches usually use centroid or bounding box to present human targets. They detect the occurrence of human and track their trajectory, which help to analyze human behaviors [8]. More complex approaches track the changing contour of human to extract some key points which are used to identify behaviors [17]. The highest level methods model human as an articulated object and evaluate the configuration of each body part. The dynamics of human body is used for recognition [12]. The higher cognitive level the approach can achieve, the higher resolution and stronger computing power are required.

The pictorial structure model proposed by Felzenszwalb in [7] can be used to describe articulated objects. Actually, it is a probabilistic graph model. For human pose estimation, the configuration and observation of body part are denoted by nodes, while the connected relationship of body parts and the observation process are described by edges in the graph. Pose estimation is achieved by evaluating the maximum a posterior configuration. There are many ways raised to improve this model. The appearance model is a part of the observation process. It is a key factor to estimate the human pose accurately, and can be learned through different methods. In [10], the appearance model is learned from annotated images. Based on roughly annotated images and image segmentation method [9], we can accurately obtain the image region of human body parts. The Method proposed in [13] sequentially learns better features tuned to a particular image. In [14], a stylized pose is detected in every single frame then the appearance model is learned. The graph structure and inference method can also be improved in different ways. If the human pose is estimated in a single image, a tree graph can be used and the belief propagation (BP) algorithm [11] obtains optimal inference. But applying the BP algorithm to approximate the continuous distribution is still a challenging task. Hence, Sudderth et al. propose the nonparametric belief propagation (NBP) in [15]. Particle filter [6] and hybrid Monte Carlo filter [5] can also be used to estimate human pose. In [2], the complete graph is adopted and the improved A*-search is used for inference.

During an examination, examinees will keep sitting posture. Their behaviors can be identified by their upper body motion. Hence, a color surveillance camera is placed on every table. In this way, one camera captures only one person's action with the front view. The pictorial structure model is adopted to estimate the pose of human upper body. This task is performed in the indoor environment. Hence, we assume the lighting condition is invariable. The arrangement of classroom is often as simple as possible, because complex environment will make monitoring difficult. If the background is too cluttered or the color of background is similar to the color of examinees' appearance, the monitoring task is hard even for supervisors. In this case, foreground and background is easy to discriminate. However, the appearance of people is unknown and unconstrained, as people can wear any kind of clothing. Some general feature such as edge can be used to describe the appearance. However, the accuracy and reliability of pose estimation is low with these features. Methods learning the appearance model from annotated images are also unsuitable for our application, because the annotation task will cost much time and manual labor if there are too many examinees.

Our work is based on the method proposed by Ramanan in [14]. We require examinees to act a specific pose at the beginning of an examination. Then their upper body parts are segmented for appearance modeling. In our application, human are captured with front view. Hence, the skin color can be a very useful clue. We establish a skin color histogram

This work is supported in part by National Natural Science Foundation of China under Grant Nos. 61273136 and 61034002, and Beijing Natural Science Foundation under Grant No. 4122083, and Visiting Professorship of Chinese Academy of Sciences.

and use back projection to locate the head and hands. According to the required human pose and the position of head and hands, we can approximately determine the region of examinee's torso and arms. After that edge detection and chamfer matching will be processed only in the interested image region. In this way, the configuration search space of body parts is greatly reduced and other foreground objects which have similar shape with body parts can be easily excluded. The accuracy of detecting the individual body part is improved. On this basis, we will obtain more reliable appearance model of human body. During the examination, we will use this appearance model to estimate their pose. We set an allowed range of body motion. If any body part of human is out of range, we will mark the body part to call the supervisors' attention.

The paper is organized as follows. In section 2, we describe the basic concept of pictorial structure. In section 3, the method to initialize the appearance model is discussed in detail. In section 4, we present the belief propagation algorithm to inference the human body pose. Section 5 shows the results of our method. Section 6 concludes our work.

2 Pictorial Structure Model

The pictorial structure model is an undirected probabilistic graph model to represent articulated objects.

2.1 Undirected Probabilistic Graph Model

Graph is a set of nodes and edges. A node is denoted by v and the set of nodes is denoted by V. Each edge connects two nodes. Hence, the edge is indicated by a pair of nodes $e=(v_1,v_2)$. The set of edges is E. In probabilistic graph model, nodes are associated with random variables. There are two kinds of variables. One is the latent variable; the other is the observed variable. The variables in graph satisfy the pairwise Markov property. We can use Hammersley - Clifford theorem to derive the joint probability distribution [3]. The chain and tree graph is frequently used in practical problems. In these two structures, two nodes connected by an edge will form the maximum cliques. Hence, each edge is associated with a potential function. Every observed variable must be connected with its corresponding latent variable and their edge is associated with a potential function of observation measurement. The edges connecting latent variables are associated with potential functions that describe the mutual relationships between latent variables.

2.2 Pictorial Structure Model

In general pose estimation task, human motions are very complex. It makes the perspective relation between human body part and camera changing and the shape of body parts' projection inconstant. However, in our application, the surveillance video is taken with the front view and the human motion is limited by sitting posture. Hence, we assume that the projection of body parts is a fixed-size rectangular and the action of body parts will only result in the translation and rotation of rectangular. Under this assumption, human body can be simplified into an articulated object. The pictorial structure model can be used to describe human body. A configuration vector (p,q,θ) represents the state of rectangular. (p,q) is the pixel coordinates of rectangular center. θ is the rotation angle of rectangular. The latent variable nodes denoted by x_s represent the configuration of rectangular. The corresponding observed variable nodes denoted by y_s represent the image of body parts. The observation potential function is denoted by $\psi_s(x_s,y_s)$. We will interpret it more detail in section 3. Human body part is connected by joints. The connection is described by a potential function $\psi_{st}(x_s,x_t)$. According to the configuration and size of body part we can calculate the coordinate of joint point. We require that the joint points of the two connected body parts can't be far from each other. The angle between two connected body parts can be unconstraint. But some angle will cause occlusion, for example, the upper arm will be occluded by the lower arm when their angle is π. If an occlusion happens, the observation potential function will fail to represent the observation process. Therefore, we limit the range of angles. When the angles are out of range, we will not estimate the pose of occluded body part. We show these in Fig. 1. The potential function between the upper arm and the lower arm is defined as follows.

$$\psi_{st}(x_s,x_t)=\begin{cases}\exp(-d^2/2\sigma^2),\Delta\theta<\dfrac{5}{6}\pi\\ 0,others\end{cases} \tag{1}$$

where d is the Euclidean distance between two joint points and $\Delta\theta=|\theta_1-\theta_2|$ is the angle between two body parts. σ is a parameter that control the strength of connection. If we require the joint point must be very close, then we can set σ to a small value.

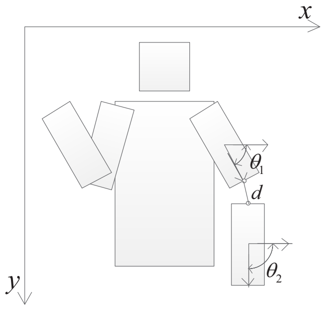


Fig. 1: The connection between upper arm and lower arm.

The human upper body has 6 parts including head, torso, left/right upper/lower arm and their connected relation can be described by a tree structure, as shown in Fig. 2. The white nodes denote latent variable and the shade nodes denote the observed nodes.

Let $x=\{x_s\}$ be the set of latent variables and $y=\{y_s\}$ be the observed variables set. In the tree graph, the maximum clique is the set of two adjacent nodes. Therefore, the joint probability distribution of x and y is expressed as below.

$$p(x \mid y) \propto p(x,y) \propto \prod_{(s,t)\in E} \psi_{st}(x_s,x_t) \prod_{s} \psi_s(x_s,y_s) \tag{2}$$

Pose estimation is achieved by evaluating the maximum a posterior configuration. The inference method will be discussed in section 4.

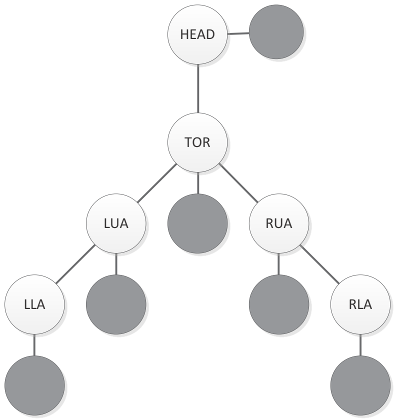


Fig. 2: The tree structure of human upper body.

3 Observation Measurement

In the pose estimation task, the projection of human body part will occupy a region of foreground image. This is the observation of body part. Obviously, the observation is determined by the appearance and configuration of real body part. However, even given the configuration and known the appearance of body part, the observation image can be different. This is because our model simplifies the appearance of real body part. Other causes such as motion blur and deformation of clothes can also result in the uncertainty of observation. Additionally, body part in different configuration or other objects having similar appearance with the body part may generate "identical" observation image. This is the uncertainty of observation process. In order to estimate human pose in an image, we must measure the uncertainty of observation image given the appearance and configuration of body part. We have given an important assumption about the shape of body part in section 2. We assume that the body part can be described as a rectangular. Hence, we usually use the similarity between the image region of body part and the rectangular to describe the uncertainty. There are many methods using different image features such as edge, color and histogram of oriented gradient [18] to measure the uncertainty. No matter which method is adopted, we need to find the image feature of body part in the image. If we have acquired the feature of body part, then according to the assumption about the shape of body part in section 2, we only need to estimate the uncertainty of generating this image region under different configuration of rectangular. There are many methods to extract the feature. In [14], color is used to calculate the polynomial feature and then the logistic regression model [3] is used to segment the image into human body part and background. In this way we will get a labeled image, the labeled region is compared with the rectangular to measure the uncertainty. But this will cause a new problem. In general, the color of appearance is unknown. If we use the logistic regression model, we must have a training set which has the color data of body part and background. The simplest method to solve this problem is to annotate the image manually. But when there are too many people, this method is inefficient. A stylized pose method is proposed in [14]. Actually, this method is also a pose estimation process based on the pictorial structure model. But the image edge which is a more general feature is used to describe the appearance of body parts and a specific human pose is required. The segmentations of human body parts are obtained and better feature training set can be easy to obtain.

Our work is based on [14]. We require examinees to initially sit straight up and raise their arms sidely to the shoulder level, as shown in Fig. 3. In this pose, human body parts will not occlude each other. Hence, using edge feature and chamfer matching [16] can segment body parts and background accurately and reliably. Besides the image edge, the skin color is also a general feature. We use the skin color to detect human head and hands. This will give us the image range of the upper body and reduce the search space of human body part. The process is shown in Fig. 4.

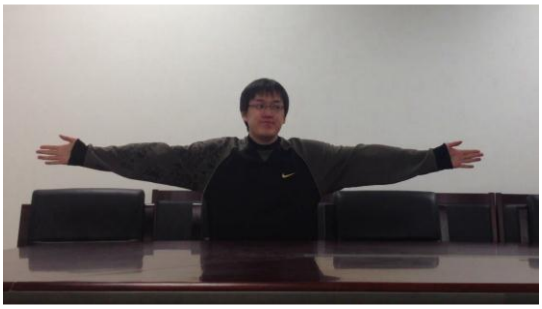
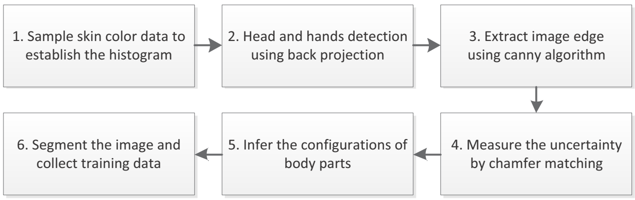


Fig. 3: The required initial pose.



1. Sample skin color data to establish the histogram
2. Head and hands detection using back projection
3. Extract image edge using canny algorithm
4. Measure the uncertainty by chamfer matching
5. Infer the configurations of body parts
6. Segment the image and collect training data

Fig. 4: The process of collecting appearance data.

First of all, we must establish a skin color histogram. We sample the pixel which represents skin color manually. Then the back projection method is used to find the image regions of head and hands, shown in Fig. 5.

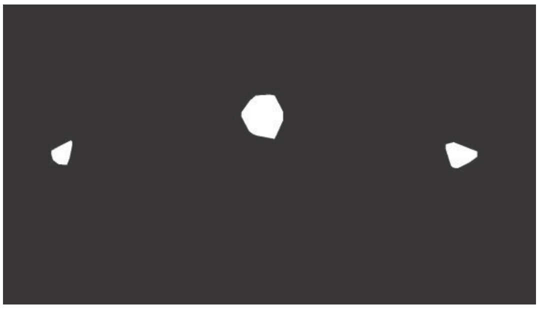


Fig. 5: The region of head and hands.

Because we require examinees to raise their arm, the hands and arms will be about the same height. Therefore, the image region of hands can be used to limit the search space of arms. Similarly, the position of the head will limit the range of torso. Fig. 6 shows the reduced search space. The red region is the search space of the torso and the green region is the search space of arms.

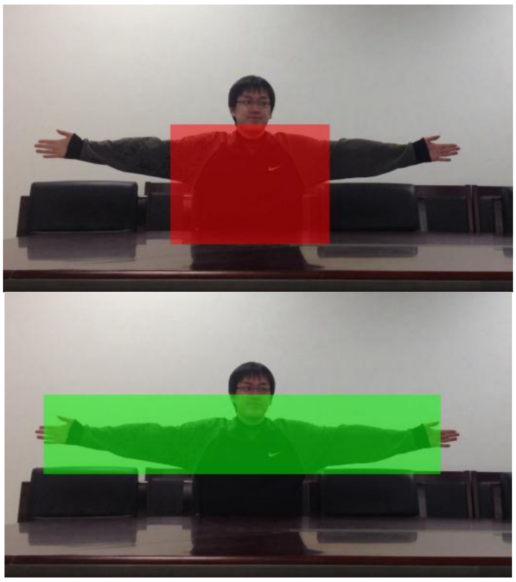


Fig. 6: The reduce search space.

In the third step, canny algorithm [4] is adopted to extract the image edge. According to the shape assumption, the edges of the human body part are two parallel lines. The template of arm's edge is shown in Fig. 7. The chamfer matching is used to measure the uncertainty under various human body configurations.

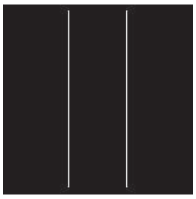


Fig. 7: The template of arm's edge.

In the fifth step, the human pose is estimated by belief propagation. According to the shape assumption and the configuration of human body parts, the regions of human body parts can be labeled. At last, we collect the color data of the body part in the labeled region and the surrounding color data of background, shown in Fig. 8. The green region is the arm and the blue region is the surrounding background. We extract the second-order polynomial feature vector of the color vector, and then train the logistic regression model.

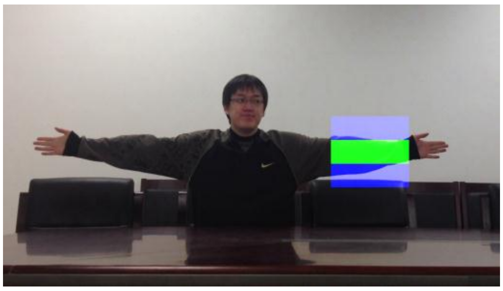


Fig. 8: The region of arm and the surrounding background

After the appearance modeling, the logistic regression model is used to segment the image. The feature vector of a pixel is extracted from its color. The process of pose estimation is essentially the same as the appearance modeling. The difference is the uncertainty measurement. The region which may be the body part is marked. The region of arms is shown in Fig. 9. We can see except the real arm there are many other regions of background. This is because the color of these noisy regions is similar to the arm. But the shape of these noisy regions is not the same as the arm.

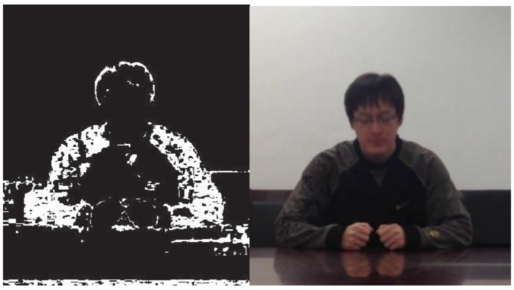


Fig. 9: The region of arms and the kernel of arms.

We use a kernel which encodes the shape and configuration information to measure the uncertainty. The kernel of arm is shown in Fig. 10.

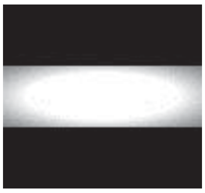


Fig. 10: The kernel of arms.

When the kernel is in some position, it will count the number of labeled pixels in the middle region and the number of unlabeled pixels in the side region. If the configuration of kernel is similar to the configuration of body part, then the number of labeled pixel in the middle will increase and the unlabeled pixel on the side will also increase. The middle region of the kernel is a Gaussian function. We choose the center of kernel as the origin, so that the kernel is calculated by

$$\omega(x,y)=\exp(-\frac{x^2}{2\sigma_x^2}-\frac{y^2}{2\sigma_y^2}) \tag{3}$$

The value of σ_x (σ_y) is half of the width (height) of rectangular. The Gaussian kernel is used to convolve with the region of arm shown in Fig. 9. In order to count the number of unlabeled pixels in the side region, we convolve the inverted image of Fig. 9 with the kernel shown in Fig. 11.

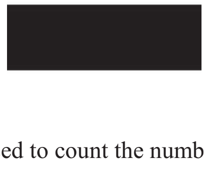


Fig. 11: The kernel used to count the number of unlabeled pixels.

The kernel is rotated by 24 discretized angles. Then the observation measurement under all configurations is obtained.

4 Belief Propagation

The belief propagation (BP) algorithm can be used to calculate the marginal probabilistic distribution of any node in graph. At each iteration of the BP algorithm, we will calculate belief messages $m_{ts}(x_s)$ which is sent from node t to its parent node s according to the message from the child nodes u of node t.

$$m_{ts}(x_s) = \alpha \sum_{x_t} \psi_{st}(x_s, x_t) \psi_t(x_t, y_t) \prod_{u \in \Gamma(t) \backslash s} m_{ut}(x_t) \tag{4}$$

The message passing process is shown in Fig. 12.

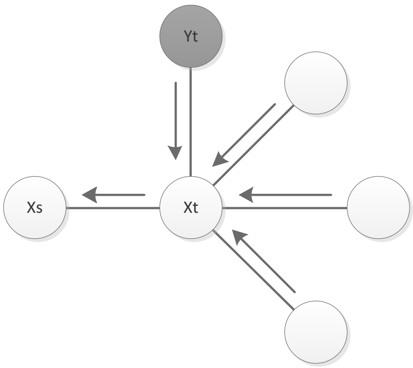


Fig. 12: The message passing of belief propagation algorithm.

In general, the appearance of human body is symmetric. So when we calculate the observation measurement, it is not necessary to discriminate which is the left arm and which is the right arm. We can treat this as a multiple targets problem. If the arms don't occlude each other, then the configuration distribution of arm will have two local maximum. The graph structure actually used is shown in Fig. 13.

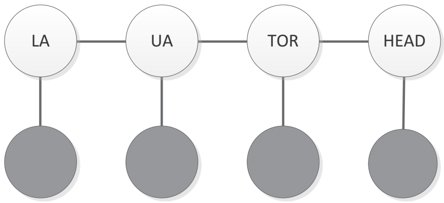


Fig. 13: The graph structure we actually used.

Hence, we will have one configuration of head and torso and two configurations of arm. This is the estimated pose of examinees.

The configurations of body parts are discretized. But the configuration space is still very large. Using the belief propagation algorithm directly is still difficult. Hence, a sampling method is adopted. We will sample some configurations of all human body parts from the observation measurement. Then we use belief propagation to calculate the marginal weight vector of root node. We use ancestral sampling to obtain a sample set of body pose to represent the uncertainty of pose. If we directly choose the maximum a posterior pose in the sample set, the human pose will be discontinuous because we discretize the configuration space. To solve this problem, we use mean-shift method to approximate the maximum a posterior pose. We transform the configurations of all body parts to a vector which consists of all joint points coordinates.

5 Experimental Results

To validate the approach, we use an image sequence to test. The resolution of images is 640×480 pixels. We set the size of arm is 30×10 pixels, the size of torso is 25×50 pixels and the size of head is 15×15 pixels. The number of sample in BP algorithm is 2000 for each body part. The algorithm is implemented with C++ and OpenCV. The program is running on a computer with Intel Core I5-2400 processor.

Evaluating the performance of pose estimation is difficult, because we don't have the ground truth of the body parts' configurations. We can compare the result of our method with manual annotated images. However, the ambiguity of partial occlusions is unavoidable. In this work, our goal is to detect the cheating behavior. We consider the examinee is cheating when any body part is out of range. We stretch out the arms and repeat this action 10 times. Fig. 14 shows the performance of our methods. In all experiments, our method can detect that the lower arm is out of range. But when people perform too complex action, the pose estimation will fail. This is because self-occlusion is very serious. To process a frame will cost 4-5 seconds.

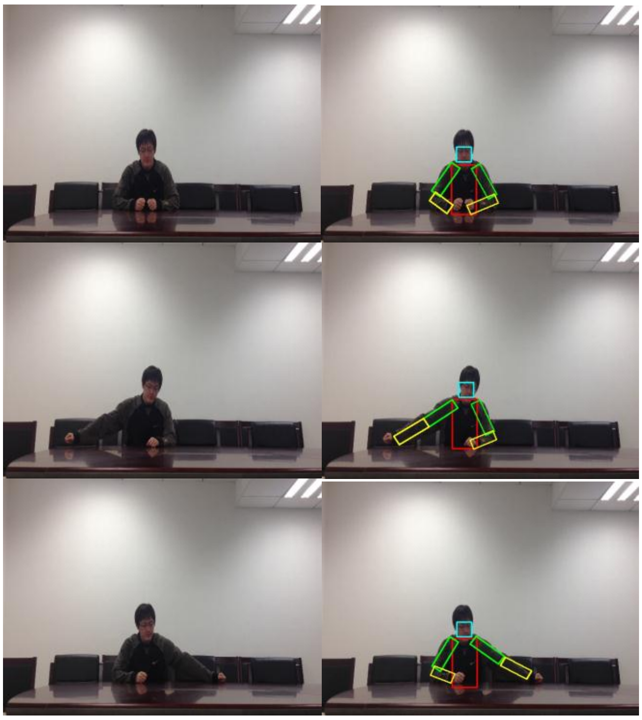


Fig. 14: The experiment results.

6 Conclusions and Future Work

In this paper, we use the pictorial structure model to detect the cheating behaviors. We improve the appearance modeling method proposed in [13] by exploiting skin color information to reduce the configuration search space. A skin color histogram and back projection method are used to detect the locations of head and hands. We infer the image region occupied by torso and arms in the light of required human pose and the location of head and hands. By this mean, we can search the configuration of body parts in a much smaller region and many other foreground objects which have similar appearance with body parts can be excluded. We use edge detection and chamfer matching to detect the individual body parts in the interested image region. Then we obtain more reliable appearance model of human body. Our experiments show the effectiveness of the approach in the application. However, recovering the human body pose from 2D image needs many assumptions. These assumptions limit its application. Now there are many methods to acquire 3D information of human body. 3D information can give more accurate and reliable appearance features. Our approach estimates the pose of human in a single frame. But human motion is continuous. According to the information of previous frames, we can largely reduce the configuration search space.

References

[1] J. K. Aggarwal and Q. Cai, Human motion analysis: A review, *Proceedings of the 1997 IEEE Non-rigid and Articulated Motion Workshop*, 90-102, 1997.

[2] M. Bergtholde, J. Kappes, S. Schmidt, and C. Schnorr, A study of parts-based object class detection using complete graphs, *Int. J. Computer Vision*, 87(1-2): 93-177, 2010.

[3] C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*. Springer-New York, 2006.

[4] J. Canny, A computational approach to edge detection, *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(6): 679-698, 1986.

[5] K. Choo and D. J. Fleet, People tracking using hybrid monte carlo filtering, *Proceedings of the 8th IEEE International Conference on Computer Vision*, 2: 321-328, 2001.

[6] J. Deutscher, A. Blake, and I. Reid, Articulated body motion capture by annealed particle filtering, *Proceedings of the 2000 IEEE Conference on Computer Vision and Pattern Recognition*, 2: 126-133, 2000.

[7] P. F. Felzenszwalb and D. P. Huttenlocher, Efficient matching of pictorial structures, *Proceedings of 2000 IEEE Conference on Computer Vision and Pattern Recognition*, 2: 66-73, 2000.

[8] I. Haritaoglu, D. Harwood, and L. S. Davis, W4: real-time surveillance of people and their activities, *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8): 809-830, 2000.

[9] Y. Hong, J. Yi, and D. Zhao, Improved mean shift segmentation approach for natural images, *Applied Mathematics and Computation*, 185(2): 940-952, 2007.

[10] S. Johnson and M. Everingham, Learning effective human pose estimation from inaccurate annotation, *Proceedings of 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011)*, 2011: 1465-1472.

[11] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[12] R. Poppe, Vision-based human motion analysis: An overview, *Computer Vision and Image Understanding*, 108(1): 4-18, 2007.

[13] D. Ramanan, Learning to parse images of articulated bodies, in *Advances in Neural Information Processing Systems*, 2006: 1129-1136.

[14] D. Ramanan, D. A. Forsyth, and A. Zisserman, Strike a pose: Tracking people by finding stylized poses, *Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, 1: 271-278, 2005.

[15] E. B. Sudderth, A. T. Ihler, M. Isard, W. T. Freeman, and A. S. Willsky, Nonparametric belief propagation, *Communications of the ACM*, 53(10): 95-103, 2010.

[16] A. Thayananthan, B. Stenger, P. H. Torr, and R. Cipolla, Shape context and chamfer matching in cluttered scenes, *Proceedings of 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1: I-127, 2003.

[17] K. Yoyama and A. Blake, Probabilistic tracking with exemplars in a metric space, *Int. J. Computer Vision*, 48(1): 9-19, 2002.

[18] Y. Yang and D. Ramanan, Articulated human detection with flexible mixtures of parts, *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(12): 2878-2890, 2012.