International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.8, No.3 (2015), pp.1-10
http://dx.doi.org/10.14257/ijsip.2015.8.3.01

Contour Model and Robust Segmentation based Human Pose Estimation in Images and Videos

Yunheng Liu

School of information technology, Nanjing Forest Police College, Nanjing, Jiangsu, China
789liuyunheng@163.com

Abstract

Pose estimation which is regard as the cross-technology of computer vision and pattern recognition, and an important prerequisite for human behavior understanding. Human pose estimation which use the probability theory, machine learning, pattern recognition, graph theory and other theories to get the position, the deflection angle of the various parts of the body. Then make the detection and estimation parameters for the human body pose. When the image has interference in the background, color and scale changed, human pose complex, self-occlusion and interpersonal interaction occlusion may make the precision and accuracy of pose estimation face great challenge. Thus, according to the above problems, this paper use the advanced model of the human body as contour model to descript the complex pose, in order to make the model more accurate and suitable for various human pose, we pre-clustering the human body pose of the training samples before we trained the model and in order to ensure the accuracy of the pose we use robust segmentation of multi-view with a novel shape prior. The experiment shows that the algorithm performs better than the classic algorithm on the public datasets.

Keywords: *contour model, pose estimation, robust segmentation, shape prior*

1. Introduction

Over the past decade, many researchers pay much attention on the human pose estimation, but because of changes of appearance, illumination, shape of body parts, body parts self-occlusion, complex structure of the human body and the huge state space, although a variety of pose estimation algorithm has proposed, but so far the problem has not been solved. Image segmentation which based on Markov random field model (MRF), through the region structure it can be bring richer local statistical information for the image, but as the overly smooth of the prior region, often lead to the contour blur, even appear marginal zone. Chen et al. who based Pairwise MRF theory proposes higher order proxy neighborhoods model, HOPS [1], by message passing between the local regional nodes, they lead more local feature information in and approximated the high-order MRF neighborhood effective. Because the border among the segmentation regions are highly kurkotic [2], therefore, the algorithm which based on point-to-point interaction MRF model often bring significant approximation error. Establish effective image edge priori model is an important way to improve the quality of image segmentation. Koray, who established a Cauchy distribution which based on the edge priori model [3], this model uses the dependency of pixel space maintain the clarity of the edges effectively, but such Cauchy scale parameter choosing still influence the edge distribution. Zhang et al who established the Laplacian prior model

ISSN: 2005-4254 IJSIP

[4], extracting the edge of the image, but the TV priori cannot sufficiently captured image edge and may cause the piecewise linearity. Katsuki, who use a causal Gaussian MRF model as a priori knowledge [8], although this model considered the edge structure of the image, but it still cannot capture the statistical characteristics of natural images effectively. Zheng et al [3] extended the multi-resolution technology from the pixel level to the regional level, combined with the regional multi-resolution and MRF model; they propose a new segmentation method to improve the multi-resolution segmentation. He *et al*. [4] proposes a fast edge tracing method, keeping the segmentation results and meanwhile greatly reducing the unsupervised MRF segmentation computation time. But traditional methods still have some flaws, the prior model only considers local neighborhood information of the image when the images have serious noise pollution, robustness decline, cannot described the edges well and has poor segmentation, which is due to the assumptions of the impact of each pixel in the neighborhood of the center is the same is unreasonable. Some researchers [5-6] estimate 2D human body pose, this approach reduces the dimensions of the pose estimation, particularly in the occasion that the camera parameters are unknown; estimating the 2D human body poses becomes more feasible. And a lot of work [7-8] estimate the 3D body pose, trying to obtain 3d information of each joint angle. Especially Taylor [8] use a typical method based on geometrical reconstruction and proposed three assumptions: projection model is weak perspective projection model; model of the human limb length is known; the joint points of the model corresponding to the body joint points of the image. Application Taylor polynomial expansion the joints can be estimated for each relative depth. For the ambiguity can be handled by use of relevant constraints. The advantage of this method is that the computational complexity is small, the optimization process is avoided; disadvantage is that in the general case assumption may not be able to meet and the conditions are too harsh, additional the priori constraint is also an important condition for the good estimate results.

We present a new method to estimate the object pose in images and videos. We use the contour model; it captures the body's natural shape and changes in poses. The model predicted the contour, along with their segmentation into parts forms the training set. Then use the MAP-MRF robust segmentation, which is to solve the pixels assignment before the pose estimation, which to make sure that everyone belongs to their areas of the image. At the same time we introduce a novel shape prior to combine the initial estimate of pose and shape for segmenting the target.

2. Body Model

This Contour model (Figure 1) provides a detailed 2D representation of natural body shape and captures the variation. It represented by part-based of current 2D models with different colors in Figure 1. Importantly, the model also captures the non-rigid deformation of the body. This allows the contour model to accurately represent a wide range of human shapes and poses.

Contour model using the existing detectors and graph structure modeling (PS) to initialize the model and use PCA to capture changes in human body model (and the camera viewpoint changes). Using the MAP-MRF segmentation to segment and refine the targets of the image. The results of pose estimation through the contour model comparison with the segmentation.

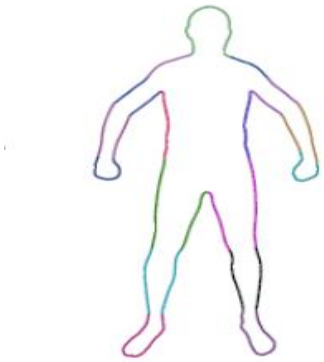

Figure 1. Contour Model

2.1. Pose Clustering

Different from the existing methods, in order to make the model more accurate and suitable for various human pose, we pre-clustering the human body pose of the training samples before we trained the model, then each pose is training alone in order to get the body pose models.

Most of the conventional methods only use a global model, each detector of the body part and Gaussian prior of relative positions of each connecting parts all come from the training samples. However, due to the great differences of the human body pose in parts of the appearance, viewpoint and dress color. It is also serious during the training process which makes the difference of the same class too large and leading to the relationship between the detection model and Gaussian prior is too broad, distinction is lower too.

To overcome this problem, we use contour model proposed by [11] and extended it with pre-clustering the human body pose and divided the pose space is into several different classes each class of the human body pose we will establish the mixed contour model. The benefits of this method: (1) as compared with the global pose space the consistency of the pose with clustering is better, so the priori relationship between the parts after pose of clustering obtain with better loyalty; (2) because the body pose in each class may has the similar structure which make all parts of the body are consistent with this structure so that we can establish the body the appearance model within the scope of pose rather than a global which is more accurately distinguish stronger.

The segmentation of the pose space can be viewed as the maximum likelihood clustering problem, we seek a set with K poses and the true value of each pose sample are similar with a high probability with at least a pose classes:

$$\hat{\Theta}=\arg\max_{\Theta}\prod_{i=1}^{N}\max_{j=1}\left[p(L_i \mid \Theta_j)\right] \quad (1)$$

Where L_i is the pose true value of the N images, Θ_j is the pose parameters of each class that corresponds to the contour model. We should note that the pose of each class is determined by the maximum posterior probability which it belong to.In order to solve Equation (1) which get the model parameters of each class, we use the K-means algorithm [12].

We use the body position coordinates offset between each pair of adjacent parts of the pose as a feature and clustering of all the training pose, the steps of the clustering algorithm is as follows:

(1) We selected K samples as the initial cluster centers arbitrarily;

(2) Calculate the distance between each sample and each cluster center and assign it to the minimum distance class;

(3) Calculate the mean of each sample in the cluster and update of the cluster center;

(4) If the clustering is convergence, that means the sample labels assigned the same, algorithm terminated, otherwise the process returns to step (2).

K-means algorithm is a greedy process, so that you can find the local maximum of formula (1). By experience and experimental verification we take the K as 4, the structure of the human body model with the cluster center of each class are shown in Figure 2.

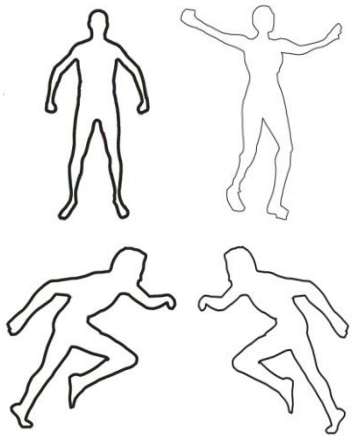

Figure 2. Clustered Pose Model

3. Robust Image Segmentation

3.1. Image Segmentation with MRF

According to Hammersley-Cliford, MRF is equivalent to a Gibbs random field (GRF), the equivalent Gibbs distribution is as follows:

$$P(X) = Z^{-1} \cdot \exp[-E(x)] \quad (2)$$

Where Z is the normalization constant, E (X) is the Gibbs energy function for all the sum of group potential $V_c(X)$of possible group C, $E(X) = \sum_{c \in C} V_c(X)$.

$V_c(X)$ Is the potential function of sub-group c, and its value is only related to the value of the pixel within the sub-group c, C is the set of all the group. We only take the first and the second potential group, so the Gibbs energy function is:

$$E(X) = \sum_{c \in C} \sum_{c \in C} V_i(X_i) + \sum_{i \in S_i} \sum_{\in N_i} V_{ij}(X_i, X_j) \quad (3)$$

The unary potential function $V_i(X_i)$ use to measure possibility that a given data node X is the labeled X_i, the second-order potential function represents the interaction between adjacent nodes i and j it is not only related with the mark of node i, but also related with neighbourhood node j.

3.2. The Energy Function

Segmentation energy function is defined as:

$$E(f) = E_m(f) + E_s(f) \quad (4)$$

The energy function contains two terms; the first term $E_m(f)$ is the Gibbs energy function of formula (1), $E_s(f)$ is the shape priori energy function. Here the shape prior is the binary template which similar to the target.

We need to solve flux vector field t by segmenting the contour S, which is equivalent to divergence of the vector field t from the area in the contour. t refers to the gradient vector field of the contour of the template signed distance map named shape prior vector field. Flux (S) should refer to the intensity of shape prior vector field t which is perpendicular to the segmentation contour surface S of unit area (i.e., the divergence of t) is a scalar.

Flux maximum constraint makes the contour local orthogonal with the shape prior vector field's. If S is a section of arbitrary contour / surface of the target space, then by segmenting the contour S the flux of vector field t is:

$$\text{flux}(S) = \int_S < \hat{n}, t > dS \quad (5)$$

Where <,> represents an inner product, $\hat{n}$ represents the outward unit normal vector of each point on the contour S. We can find that when we follow the outward unit normal vector and the vector field t to align, that means the contour of the contour S is same with the object and contour S has the maximum flux. So the difference between the contour and contour template S are embodiment, if the difference between contour template and the contour smaller, the larger is the flux vector field t by the contour segmenting, otherwise, the flux is smaller.

To show how this constraint joins into the Gibbs energy function, we use divergence theorem for differentiable vector fields:

$$\text{flux}(S) = \int_S < \hat{n}, t > dS = \int_R \text{div}(t) dR \quad (6)$$

Wherein R is the area surrounded by the contour S. Divergence theorem shows that the flux through a section of the contour vector field equal to the divergence of the vector field of the area surrounded by this contour. Note that this theorem is available when the relevant vector field is differentiable. Due to the feature of this theorem, we can find that the maximum flux constraints can join into the Gibbs energy function by the form of first order group of potential function. Based on the above discussion, we define a maximum flux constraint as:

$$E_s(f) = \sum_{i \in p} V_i(f_i) \quad (7)$$

$V_i^F(f_i|D)$ Is defined as follows:

$$V_i(f_i) = -\text{div}(t_i) \quad \text{if } f_i = 1 \quad (8)$$

Where $f_i = 1$ is the region surrounded by contour S, S is the segmentation without shape prior.

3.3. Shape Prior t

It is come from the gradient vector field of the object symbol distance map. Reference the definition from [13] about the contour of the model; we will define it similar to the signed distance function of level set:

$$\varphi_M(x) = \begin{cases} 0 & x \in M \\ +\min_{x_M \in M} \|x - x_M\| & x \in R_M \\ -\min_{x_M \in M} \|x - x_M\| & X \in \Omega \backslash R_M \end{cases} \quad (9)$$

The contour of the template M divided the image Ω into two regions, *i.e.*, background area $\Omega \backslash R_M$ and area RM surrounded by M, where x = (x, y) are Cartesian coordinates of the image

pixels, $x_M = (x_M, y_M)$ is the coordinates of contour Template M. $\varphi_M(x)$ calculate the closest distance from each pixel in the image to the contour template M at the same time the $\varphi_M(x)$ in the contour template M which is greater than 0, and its outer one is less than 0. You can get multiple contours if we connecting the points have the same distance to the contour template M, (it can be seen as the contour of a different size and contour template). As the gradient direction of the image pixel is the same with the normal vector direction of contour lines, so when we get signed distance map of the contour template, then seek the gradient, equivalent to get the normal vector of contours that is the normal vector of the contour template t.

3.4. Minimize E (f)

To minimize energy function of formula (2), we use the graph cuts. V_{ij} Should be a submodule function and the submodule need to satisfy the following:

$$V_{ij}(0,0) + V_{ij}(1,1) \leq V_{ij}(0,1) + V_{ij}(1,0) \quad (10)$$

To the target / background that we studied, the constraints that we combined are the submodule, therefore energy function can be minimized by the graph cuts.

4. The Pose Estimation

When we get the segmented foreground regions, the maximum a posteriori (MAP) inference result of the mixed contour model for each class of the pose can be obtained by the method of Yang and Ramanant [14]. This makes the segmented foreground regions has a group of pose estimation in each cluster and the pose estimated result corresponding to the confidence probability. So that we need a mechanism to determine the type of the pose belong to which class. However, we note that compare of the highest probability of each class of model directly is unreasonable, because body structure models in each class are trained by their own set of training samples, when under different circumstances compare the probability is unreasonable. To solve this problem, we use the best estimate probability of each class to train the samples and the segmented foreground regions which belongs to true value of the pose class, the training of a weight vector w_c for each class c to weight average of the probability, then use the weighted average result to confirm the segmented foreground regions belongs to which class, as follows:

$$\hat{c} = \arg\max_c w_c^T P(L|I) \quad (11)$$

w_c Is the probability weighting vector for class of c, $P(L|I)$ is a probability vector composed by the best results of various model of class detection probability. To study each class of weight vectorw_c, we use multiple linear regression algorithms [15].

Make the training samples as $(c_i, p_{i1}, p_{i2}, \cdots, p_{im})$, i=1,2,…,n, where $p_{i1}, p_{i2}, \cdots, p_{im}$ is the m variables of sample i, *i.e.* the best detection probability of the model in each class, c_i is the dependent variable of sample i. Because of the independent variable may not able to describe the relationship between the independent variables completely, so we get error term ε_i to describe the other factors influence in addition of variable to c_i. Therefore, the expressions of multiple linear regression models are as follows:

$$c_i = \beta_0 + \beta_1 p_{i1} + \beta_2 p_{i2} + \cdots + \beta_m p_{im} + \varepsilon_i \quad (12)$$

Equation (12) describes the dependent variable is the sum of the mean and error. Linear refers to the mean of c_i is a linear function about the unknown parameters$\beta_0, \beta_1, \cdots, \beta_m$.

As shown in equation (12) above, we have m + 1 parameters $\beta_0, \beta_1, \ldots, \beta_m$ need to determine, for the convenience, we regular equation (12) as a matrix form:

$$C = P\beta + \varepsilon \quad (13)$$

Where C is the column vector which has all the dependent variable of the training samples (*i.e.*, samples pose Class), ε is the error vector, matrix P as independent variables:

$$P = \begin{pmatrix} 1 & p_{11} & \cdots & p_{1m} \\ 1 & p_{21} & \cdots & p_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & p_{n1} & \cdots & p_{nm} \end{pmatrix} \quad (14)$$

Error vector ε needs to satisfy the condition:

$$E(\varepsilon) = 0$$

$$\text{Cov}(\varepsilon) = E(\varepsilon^2) \sigma^2 I \quad (15)$$

So that the unknown parameters can be obtained by the least squares method:

$$\hat{\beta} = (P^T P)^{-1} P^T C \quad (16)$$

5. Results and Analysis

The data collected from the public evaluation of image database Buffy, images in the library were cut from the TV series " Buffy the Vampire Slayer " 3 sets of 376 images, there are complex background, illumination changes, human target of different scales in the images and all kinds of occlusion. The pose changes greatly, and the background and light conditions have dramatic changes in the challenging data set.

The contour model realistically captures a large range of real human poses in the space in which it was trained .This enables it to find segmentations which, while not perfect, are guaranteed to be plausibly human.

The algorithm that we proposed need to cluster the human body pose and appearance and established model and structural model of human body parts in each class, in order to ensure the accurate of the pose clustering and each class has enough training samples parts for the appearance model training and we need a large training set.

Before the experiment, we should determine the parameters in our algorithm at first. To determine the number of clusters when the pose clustering, we make the K=2,4,6 respectively, to cluster the human body pose, the clustering results under different scale values have different result, finally we find that when K =4 the differences of the human pose cluster which make the classes differences bigger and inside the class is become smaller, and the various classes from the human pose clustering are fitting the recognition of the human eyes, when k=4 the cluster human pose estimation result shown in Figure 4.

Tab 1 shows the compare of the proposed algorithm and other classical algorithm use Image Dataset Buffy for pose estimation and the accuracy rate of the detection the various body parts. We use the pre-clustering to preprocess the human pose, and then make the best fitted the maximum probability which from the model detection of the class of pose as an accuracy of the final pose estimation.

The segmentation results from Figure 3, we can find that, the MRF algorithm that we proposed removed the shaded part of the background and the background parts which have the similar color with the object and segment the correct objet with complete contour.

See the results, we can find that the proposed algorithm running well when there are some occlusions in the scene. In addition, within the same time the algorithm that we proposed hase high segmentation accuracy than the classics MRF algorithm.

6. Conclusion

When the image has occlusions and shaded only use the information of the images is not enough. So this paper we fuse the shape priori information into the Markov image segmentation method; and use maximum constraint of the flux to fuse the shape prior information into the Gibbs energy function. Finally we use graph cut minimized to make the Gibbs energy function to achieve the optimal solution and promote segmentation contour close to a given template. The results proved that the algorithm can segment the object accurately when the images contain shadows and occlusion in them.

In the 2D contour model training library as the human body pose of the training sample have large changes that is the differences within the class is large, so the global model trained by this method is not accurate enough and lower discrimination. To solve this problem, we propose the body pose clustering process. After clustering the differences among the poses within the class are smaller meanwhile the poses have the similar appearance and the spatial relationship between the parts therefore the contour model has the higher discrimination after training. And then we can get good results of the pose estimation.

Table 1. The Comparison of our Method about the Parts Estimation Accuracy with State-of-art Methods on the Buffy Dataset %

	head	torso	Upper leg	Lower leg	Upper arm	Lower arm	PCP	detection rate	total probabilities
Yang	94.4	98.2	69.7	60.8	59.8	25.8	60.2	69.8	41.7
ours	97.2	99.3	86.6	77.9	71.6	50.4	76.8	81.9	63.4

(a) (b) (c) (d)

Figure 3. Segmentation and Pose Estimation (a) Input Image (b) Input Image After Background Subtraction; (c) When Occluded Pixels Are Removed Before 2D Diffusion, the Obtained Shape Prior; (d) Pose Match

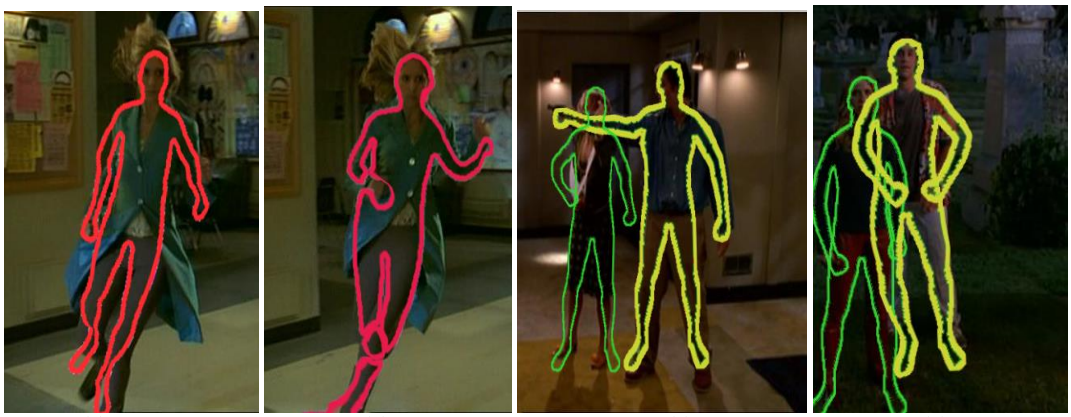

Figure 3. The Result of Pose Estimation

References

[1] A Y C. Chen, J J. Corso and L. Wang, "HOPS: efficient region labeling using higher order proxy neighborhoods", IEEE Computer Society: Proceedings of International Conference on Pattern Recognition, (**2008**), LA, CA.
[2] Q. Gao and S. Roth, "How well do filter-based MRFs model natural images?", Proceedings of DAGM Symposium. Berlin Germany, (**2012**), pp. 62-72.
[3] K. Koray, E. K. Ercan and S. Bülent, "Bayesian separation of images modeled with MRF using MCMC [J]", IEEE Transactions Image Process, (**2009**), pp. 982-994.
[4] H. Zhang, Y. Zhang and H. Li, "Generative Bayesian image super resolution with natural image prior", IEEE Transactions on image Processing, (**2012**), pp. 4054-4066.
[5] C. Zheng, L.-g. Wang, R.-y. Chen, "Image segmentation using multiregion-resolution MRF model [J]", IEEE Geoscience and Remote Sensing Letters, (**2013**), pp. 816-820.
[6] F.-y. He, Z. Tian and X.-z. Liu, "A fast edge tracking algorithm for image segmentation using a simple Markov random field model", International Conference on Computer Science and Electronics Engineering, Hangzhou, (**2012**), pp. 633-636.
[7] X. Lan and D. P. Huttenlocher, "Beyond trees: Common-factor models for 2d human pose recovery", In: ICCV, IEEE conference, (**2005**), pp. 470–477
[8] P. Felzenszwalb and D. Huttenlocher, "Pictorial structures for object recognition", International Journal of Computer Vision, (**2005**).
[9] J. Deutscher, A. Blake and I. Reid, "Articulated body motion capture by annealed particle filtering", IEEE Proceedings CVPR, (**2000**).
[10] C. J. Taylor, "Reconstruction of articulated objects from point correspondences in a single uncalibrated image", CVIU, (**2000**), pp. 349-363.
[11] O. Freifeld, A. Weiss and S. Zuffi, "Contour People: A Parameterized Model of 2D Articulated Human Shape", IEEE Proceedings CVPR, (**2010**).
[12] J A. Hartigan and M A. Wong, "Algorithm AS: A k-means clustering algorithm", Journal of the Royal Statistical Society, Series C (Applied Statistics), (**1979**), pp. 100-108.
[13] G. Tsechpenakis and D. Metaxas, "CoCRF Deformable Model: A Gemetric Model", Driven by Collaborative Conditional Random Fields, IEEE Transactions on image processing, (**2009**).
[14] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-part", IEEE Conference on Computer Vision and Pattern Recognition, (**2011**), pp. 1385-1392.
[15] R A. Johnson and D W. Wichern, "Applied multivariate statistical analysis", Upper SaddleRiver, Prentice hall, (**2002**).

Author

Yunheng Liu, 1975.04, Nanjing, Jiangsu, P.R. China. She is a lecturer of school of information technology, Nanjing Forest Police College, Jiangsu, China. Her scientific interests are computer image identification and computer education.