A Comparison on Different 2D Human Pose Estimation Method

Santhanu P.Mohan[1], Athira Sugathan[2], Sumesh Sekharan[3]

[1] Assistant Professor, IT Department, VJCET, Kerala, India
[2] Student, CVIP, ASE, Tamilnadu, India
[2] Software Engineer, TCS, Kerala, India

ABSTRACT

Abstract—Human Pose Estimation is a fast developing field and lately gone forward with the advanced finding of the Kinect system. For 3D pose estimation, this system performs good but the 2D pose estimation has not solved yet. In Computer Vision, articulated body pose estimation, systems detect the pose of a human body, that consists of joints and flexible parts using. Human body pose estimation models are complex therefore it is one of longest-lasting problems in computer vision.There is a need to develop accurate articulated body pose estimation systems to detect the pose of bodies like hands, legs,head etc. Pose estimation has many applications that can benefit such as robotics, human computer interaction, video surveillance, multimedia, augmented reality, video retrieval and biometrics or intelligent surveillance. Images and videos can have many challenges like background clutters, varying lighting conditions, unconstrained clothing of the person, occlusion etc. A comparative study included in this paper mainly focusing on different 2D human pose estimation methods like pictorial structure, silhouette method, skeletonization, model for shape context matching, segmentation, features extraction and recognition tools, research advantages and drawbacks are provided as well.

Keyword *Articulated pose recognition, model representation, feature extraction, pose estimation*

I. INTRODUCTION

Human pose estimation is the method to determine the pose of the person in an image or image sequences. It is one of the important challenging problems in computer vision that has been researched for many years. The process of human pose estimation is that the parameters of human pose models are determined. Articulated body pose estimation is the process to recover the pose from the complex human body which has articulated configurations. Estimating articulated human poses from images and sequences of images have many difficulties associate with it. Many recent methods are based on the model-based approaches which are resulting on the structural and appearance knowledge of human body that are in various poses. Human body have many joints, therefore, most of the models are represented in terms of a tree of interconnected parts with the vertices that representing parts, the edges that denoting joints and root that denoting torso.

Estimating articulated human pose in still images are challenging because of the factors like image noises, occlusions, background clutter, lighting, loss of depth information, clothing of the person, illuminations etc. In some cases, due to the uncontrolled imaging conditions it is impossible to estimate the articulated human body successfully. Therefore, pre-processing stages such as segmentation, background subtraction, morphological noise removal etc are very important in human pose estimation. These are to obtain a good structural and appearance information of the human body of different poses. From a good structural model of the body pose, features can extract easily, therefore, it is possible to estimate a successful human pose. There are many advantages and disadvantages in these methods while estimating articulated human body parts. Methods are described here to recognize articulated human body parts by using different types of algorithms.

II. COMPARATIVE STUDY

2.1 2D Articulated Human Pose Estimation and Retrieval in (Almost) Unconstrained Still Images

In this paper, a technique is proposed to automatically detect and estimate the upper body parts like head, torso and arms of 2D pose of human in uncontrolled still images. The method tried in images and videos under unconstrained conditions without any knowledge of background, lighting, clothing, location of the person in the image etc. Using an upper body detector, the person in the images and videos are localized in a simplified way using weak constraints on location from the detection. An application called pose search is generated after the successful estimation of human pose. For detecting and estimating the human pose from images and videos, the pre-processing stages such as upper-body detection and foreground highlighting has to be done. Upper body detection finds the position of people in the image and foreground highlighting removes background clutters. Moreover, if we have an assumption about the scale and approximate position of the head and torso, good appearance models can be generated for successful pictorial structures.

First, using a sliding window detector followed by non maximum suppression, the upper body parts in the images are detected based on the part based model. Detection gives the scale and approximate position of people in the image. Each window that is examined is subdivided into small tiles described by Histogram of Oriented Gradients and classified using a linear SVM. Next is foreground highlighting which removes the clutters from background by initializing the Grab-Cut segmentation. This method initialized by utilizing the prior knowledge about the structure of human body and by learning foreground or background color models from the regions where the person is available or unavailable. This reduces the search space by limiting the locations to foreground area. Pictorial structure is a probabilistic mechanism for estimating successful appearance model from images. Body parts are represented by a conditional random field. Pictorial structure model is constructed by taking the rectangular image patches and their (x,y) position, orientation θ, scale s. To get person specific appearance models from still images with unknown appearance, a well known method called image parsing is used. Initially generic features like edges are used and then the process is repeated. The output of foreground highlighting is the area to be parsed. In image parsing, body parts are oriented patches of fixed size, with position parameterized by location (x, y) and orientation θ. They are in a tree structure with edges that carries kinematic constraints. In an iterative image parsing approach, first, only the image edges are considers with part templates. Soft-segmentation is obtained from marginal distribution that is rolled together with a rectangle that represents the body parts. From soft-segmentation, the appearance models that are represented by color histograms are derived.

The main drawbacks of this articulated human pose estimation approach is that it can only work on specified datasets. It is specific to upper-bodies and needs improvement because currently this work cannot handle occlusions and impossible to recover from a wrong initial scale estimation.

2.2 Attributed Relational Graph Based Feature Extraction of Body Poses in Indian Classical Dance Bharathanatyam

Articulated body pose estimation is an important problem in computer vision because of convolution of the models. It is useful in real time applications such as surveillance camera, computer games, human computer interaction etc. Feature extraction is the main part in pose estimation which helps for a successful classification. In this paper, we propose a system for extracting the features from the relational graph of articulated upper body poses of basic Bharatanatyam steps, each performed by different persons of different experiences and size. Our method has the ability to extract features from an attributed relational graph from challenging images with background clutters, clothing diversity, illumination etc. The system starts with skeletonization process which determines the human pose and increases the smoothness using B-Spline approach. Attributed relational graph is generated and the geometrical features are extracted for the correct discrimination between shapes that can be useful for classification and annotation of dance poses. We evaluate our approach experimentally on 2D images of basic Bharatanatyam poses.

In this work a method is proposed to increase the quality of skeleton of human object in 2-D images of Indian classical dance- Bharathanatyam. A morphological skeleton of a binary image is created to enhance the skeleton quality in order to find more efficient features for classification. This allows finding a graph model with more attributes. In this method the skeleton characteristic points is represented as an attributed relational graph to model the skeleton. The future research will include finding a more accurate graph model, allowing the determination of more efficient attributes for the nodes and the edges.

2.3 Recognizing Human-Object Interactions in Still Images by Modeling the Mutual Context of Objects and Human Poses

In this paper, proposed a human object interaction method where a mutual context model between objects and human poses are described in which the object recognition can benefit from estimating the poses. The model representation of human pose estimation is taken here. Articulated human body part is a big challenge for human pose estimation where the human body parts can be self-occluded. A conditional random field model that values the spatial and co-occurrence contexts between human poses and objects are developed. In this approach, the atomic poses are the dictionary of human poses that are used for modeling human object interactions. If we know the atomic pose that corresponds to the particular image, human pose estimation can be much easier. Atomic poses can be obtained by grouping the human body part configurations. First, annotation of body parts in an image is defined. The number of body parts and vector of each part are indicating the position and orientation. To know whether the torso in all images have same position and size, align the annotations in a specific manner and then normalize the variations of position and orientation. Then a hierarchical clustering method with maximum linkage measure is used to get a set of clusters which represents an atomic pose. This is a weekly supervised approach for clustering human poses.

The deformable part model is used to train a detector for each object and human body parts. Based on the histogram of gradient feature a mixture of discriminatively trained latent SVM classifiers are used. Spatial pyramid matching method is used to train the activity classifier. Then extract salient invariant features and apply the histogram intersection kernel on a image pyramid which has three layers. The model parameters are estimated by assigning each pose to its similar atomic pose. This is a standard conditional random field model that has no hidden variables so the human part detector and object detector are applied to the given annotations of human body parts and object bounding boxes. At last, a maximum likelihood approach with zero mean Gaussian is used to estimate the model parameters. The main disadvantage of this work is that the method is impossible to apply in the situations like absence of a context between object and human interactions. Another limitation is that this work has to annotate the human body parts and objects in each training images.

2.4 Silhouette Analysis-Based Action Recognition via Exploiting Human Poses

In this paper, proposed a new idea of human action recognition from silhouettes by combining both global and local features advantages. Mutual connections between the sequential human poses in an action are taken for the action recognition from silhouettes. To encode the local features of actions temporally, set of correlated poses called correlogram of human poses which is a modified set of words model is introduced. First, to each frame of the silhouette sequence, the smallest rectangle called bounding box is applied and then it is normalized to a fixed size. Therefore, the original dimension of each frame is reduced and the translation and global scale variations are removed in pre-processing stage. Using morphological transformations like dilation and erosion, the noise during the normalization process can be reduced.

These normalized silhouettes are used as the features for the set of pose models. Treat the silhouette as feature in each frame of the video by extending the set of features model. The 2D silhouette mask is converted to 1D mask from each feature vector by scanning it from top to bottom. So, each frame is represented as the binary element vector and the length comprise the multiplication both row and column. The possibility of dimensionality increment can be reduced by using unsupervised principal component analysis (PCA). Then using k-means clustering method the feature vectors are clustered to obtain a visual vocabulary. In each clusters, the centre represented as the codeword. Therefore, the temporal structural features are explicitly encoded by using correlogram. Soft assignment strategy is also called kernel codebook is adopted to reduce the computational complexity, dimensionality of the model and quantization error. During the quantization process after k-means clustering, it preserves the visual word ambiguity that was ignored. The extension of the motion history image descriptor which is called the motion template is developed to get the holistic structural information that can be loss. Classification can be performed by using the Gaussian kernel SVM classifier. The limitation of this work is that this method is not efficient in still images because silhouettes are inherently ambiguous.

2.5 Estimating Human Body Configurations using Shape Context Matching

In this paper, a method is proposed to estimate the human body configurations by locating the joint points and poses in 3D space. A single 2D image which contains a person is used as the input image in this work. In different configurations of human body and viewpoints of camera, number of exemplar 2D views is stored in this approach.

Positions of joint points are marked or annotate manually on each of these stored views. The shape can be represented by a cluster of samples points from the edges of the person in the image. The edges are getting by using edge detector. Then shape matching context method is used to check the best match of shapes with each of the stored views. The body joint positions are then changed to the test shape from the exemplar view. At last, the 3D body configuration and pose based on a single uncalibrated 2D image from the joint locations are estimated.

In this work, shape representation is by a set of points from both the internal and external contours of the person in the 2D image. Therefore, first obtain a set of sample points on the edge of the person by performing canny edge detection method. Next, a set of exemplars extracted, and for each point in the given shape finds the best matching point on the other shape. The shape matching context algorithm is using for this performance. Next is to estimate the 2D image locations of the keypoints like hands, elbow, shoulders, hip, head, waist etc on the body with the correspondences between sample points and the exemplar. To estimate 3D configurations of the body these keypoints can then be used. The main disadvantage is that only a single 2D image is used in this method for testing.

2.6 Recognition of shapes by morphological attributed relational graphs

In this paper, proposed an efficient method for shape matching and estimate articulated poses of any objects. The main aim of this work is that the strength of the skeleton of objects in 2D images is shown in pattern recognition and computer vision. The object features are extracted for shape matching process and classification, therefore a good model of human figure needed to be obtained. A method of skeletonization process which helps to obtain the best representation of human is proposed in this work. Binary images are used as the input and pre-processing includes the morphological operations to remove the noises present in the images. The morphological thinning algorithm is used next to make the object in a single pixel thickness. But there will be unwanted branches called spurs present in the thinned representation after thinning process. The edges, vertices and joint points present in the skeletal representation are obtained by using structuring elements. To remove these spurs from the skeleton, pruning algorithm is used by fixing a threshold value and removes the spurs from the end point. A piecewise cubic B-spline method is used in the pruned skeleton to improve the smoothness of the skeleton and reduce the unwanted amount of data and to get an approximated morphological skeleton.

A skeletal representation which has edges, vertices and joints can be taken as a graph. From the smoothened skeleton, an attributed relational graph can be generated to use it as a structural model by using adjacency matrix to extract features for shape matching. The features that can be extracted are orientation, strength, length of each branch and the feature vectors of each pixel position. These extracted features are taken for the attributed relational graph matching which is also called graduated assignment. This technique is an optimization method which is for finding a good suboptimal solution and can use as a matrix to denote correspondence between objects. This gives a good result for obtaining the structural model for the articulated human pose estimation and varies from other methods with its simplicity and less computational complexity. Compared to other approaches, this method is simple, less complex computations and works in any type of images.

CONCLUSION

In this survey we have found that there are various techniques for articulated human pose estimation. But algorithm will depend upon the dataset and the environment that are captured. It was found that most of the algorithms described in this paper works in images but there are some limitations in case of its efficiency to reduce memory storage space and noises. In some case, conditional random field framework is used but in some cases, the silhouette, skeletonization and shape context model approaches are used for the estimation.

Depending upon the survey, in this paper, the morphological skeletonization process which is a powerful tool for representing human pose are determined as the best method for estimating the articulated human body. Using of B-spline gives a smoothened approximated skeleton and helps to generate an articulated relation graph to extract the features for graph matching and estimating the poses efficiently.

REFERENCES.

1. M. Eichner • M. Marin-Jimenez • A. Zisserman •V. Ferrari, "2D Articulated Human Pose Estimation and Retrieval in (Almost) Unconstrained Still Images," International Journal of Computer Vision September 2012, Volume 99, Issue 2, pp 190-214 Springer 2012.

2. Athira Sugathan and Suganya R , "Attributed Relational Graph Based Feature Extraction of Body Poses in Indian Classical Dance Bharathanatyam ",International Journal of Engineering Research and Applications, ISSN 2248-9622, Vol. 4, Issue 5 (version 1),May 2014,pp 7-11.

3. Bangpeng Yao and Li Fei-Fei,"Recognizing Human-Object Interactions in Still Images by Modelling the Mutual Context of Objects and Human Poses," Proceedings of the IEEE Transaction on Pattern Analysis and machineintelligence,Vol.34,No.9,2012,pp.1691-1703.

4. D. Ramanan, "Learning to Parse Images of Articulated Objects," Proceedings of Advances in Neural Information Processing Systems, 2006.

5. P. Felzenszwalb and D. Huttenlocher, "Pictorial Structures for Object Recognition," Proceedings of International journal of Computer Vision,vol.61, no. 1, pp. 55-79, 2005.

6. M.-H. Y. Gang Hua and Y. Wu. Learning to estimate human pose with data driven belief propagation. In CVPR, 2005.

7. • M. Andriluka, S. Roth, and B. Schiele, "Pictorial Structures Revisited,People Detection and Articulated Pose Estimation," Proceedings of the IEEE Conf.Computer Vision and Pattern Recognition, 2009,Vol.pp. 1014 - 1021.

8. B. Yao and L. Fei-Fei, "Modeling Mutual Context of Object and Human Pose in Human-Object Interaction Activities," Proceedings of IEEE Conference in Computer Vision and Pattern Recognition, 2010.

9. B. Yao, A. Khosla, and L. Fei-Fei, "Classifying Actions and Measuring Action Similarity by Modelling the Mutual Context of Objects and Human Poses," Proceedings of International Conference in Machine Learning, 2011.

10. Di Wu and Ling Shao, "Silhouette Analysis-Based Action Recognition Via Exploiting Human Poses," Proceedings of the IEEE transactions on circuits and systems for video technology, vol. 23, no. 2, February 2013.

11. P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in Proc. 2nd Joint IEEE Int. Workshop Vis. Surveillance Performance Eval. Tracking Surveillance ,Oct. 2005, pp. 65–72.

12. C. Di Ruberto, G. Rodriguez, L. Casta, "Recognition of shapes by morphological attributed relational graphs," Proceedings of the IEEE Conf.Computer Vision and Pattern Recognition, 2004 ,37 ,1 ,Vol.pp. 21-31.

13. C. Di Ruberto and A.G. Dempster, Attributed skeleton graphs using mathematical morphology, Electronics Letters, vol. 37 num. 27 (2001) 1325–1327.

14. S. Gold and A. Rangarajan, Graph matching by graduated assignment, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 1996, IEEE Computer Society (1996), 239–244.

15. Yanshu Zhu, Feng Sun, Yi-King Choi, Bert Juettler, WenpingWang, "Spline Approximation to Medial Axis," arXiv:1307.0118v1 [cs.GR] 29 Jun 2013.

16. K.J Maccallum and J.M Zhang, "Curve-Smoothing Techniques Using B-Splines," The Computer Journal, Vol.29, No.6, 1986.

17. Greg Mori and Jitendra Malik, "Estimating Human Body Configurations using Shape Context Matching," Proceeding of the 7th European Conference on Computer Vision, pages 666-680,ECCV 2002, Springer-Verlag, ISBN:3-540-43746-0.

18. S. Belongie, J. Malik, and J. Puzicha. Shape context: A new descriptor for shape matching and object recognition. In NIPS, November 2000.

19. J. Canny, "A computational approach to edge detection," Proceedings of the IEEE Transaction, PAMI, 8:679–698, 1986.