IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 5, Ver. II (Sept - Oct. 2016), PP 88-94
www.iosrjournals.org

Using The Pictorial Structures in 3D Human Body Pose Estimation

Zahra Ekhtiyari[1], Havva Alizadeh Noughabi[2]
[1,2]*Department of computer engineering, Faculty of engineering, University of Gonabad, Iran*

Abstract: *The processing of estimating the configuration and location of human body from the image, is called human pose estimation. If this pose determined in image space, the 2D pose estimation is done. If the images of human be available from at least two camera views, then the pose of body can be obtained in 3D. Using the pictorial structures in two dimensional estimation of human body is so popular. However during the recent years using the pictorial structures in three dimensional estimation of human body has been interested by the researchers and various methods have been presented. Due to aforementioned sentences, these novel methods have been surveyed and compared in present study.*

Keywords: *probability estimation, pictorial structure, human body pose estimation.*

I. Introduction

One of the issues in computer vision is human body pose estimation. One of fundamental problem in Human Computer Interaction, is determining the 3D human body pose that provide information for machine about action and intent of a human. Human body pose estimation has many applications like robotics, controlling, security and so on. Human pose estimation is the problem of estimating the position of humans in images, or more exactly the positionand orientation of the individual body parts. This isn't possible to estimate the position and orientation of all bones of the skeleton. So a simplified skeleton model is assumed.

A usual approach in the field of pose estimation is probabilistic estimation that most models fit into this formalism like part based model. Pictorial structures is a part based model that in human pose estimation parts can be represent as limbs or joints. In limb based model, each part corresponded to a limb which could translate and rotate in 3D but in joint based model, the parts could only depend on the translation of the joints but not the rotation. Limb based model and joint based model has been shown in figure 1.

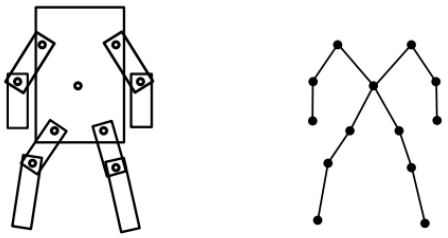

Figure 1: The left hand figure shows a limb-based model and the right hand figure shows a joint-based model

Pictorial structures represent the state of the art for 2D human pose estimation. They work well at dealing with complicated backgrounds. Also pictorial structures are good for general object detection in 2D. After all, pictorial structures haven't been applied as much for 3D pose estimation of humans. In this paper we focus on using pictorial structures in 3D human pose estimation. Pictorial structures will be introduced in section 2. Then, the analysis of the recent methods performed about the 3D human pose estimation using the pictorial structures model would be explained in section 3. Finally, Section 4 presents the conclusion of the paper.

II. Pictorial Structures Model

Fischler and Elschlager introduced the pictorial structures model in 1973 (Fischler and Elschlager, 1973). They discretized the search space and found the globally optimal solution by using dynamic programming. Pictorial structures became more popular when Felzenszwalb and Huttenlocher realized how to make the inference even more efficient using the general distance transform. (Felzenszwalb and Huttenlocher, 2012) (Felzenszwalb and Huttenlocher, 2000)

Felzenszwalb (Felzenszwalb and Huttenlocher, 2005) has stated the concept of energy for a graph in pictorial structures model and introduces total energy of a pictorial structure utilized for recognizing the objects of an image by equation 1.

$$E = \sum_{i=1}^{N} m_i(l_i) + \sum_{i\sim j} d_{ij}\left(l_i, l_j\right). \quad 1$$

Total energy in this model is obtained by summation of unary and pairwise terms. Unary terms are related to energy of each parts singly. Pairwise terms are defined among the adjacent parts in graph.$m_i(l_i)$ represents the unary energy of each *i* part of the object placed in l_i. $i\sim j$ shows parts of the object with places adjacent to each other and$d_{ij}(l_i, l_j)$ shows the amount of energy of this adjacent parts. N is the total number of parts which the object has been divided into them. In order to achieve the most efficient possible model for the object, it is necessary to look for the place of each part by minimizing the energy relation above. In other words, relation 2 is used.

$$L^* = argmin_L \sum_{i=1}^{N} m_i(l_i) + \sum_{i\sim j} d_{ij}(l_i, l_j). \quad \mathbf{2}$$

Where $L = (l_1, \dots, l_N)$ and it is a vector which indicates all parts of the object. L^* demonstrates the most efficient value of L.

The energy minimization problem is equivalent to finding the maximum a posteriori estimate of the object configuration given an observed image. The statistical formulation can be applied to learn the parameters of a model. Actually, all model parameters can be learned from a few training data using maximum likelihood estimation. The statistical framework gives a way of finding several good matches of a model to an image rather than finding just the best one. It can be obtained by sampling object configuration from their posterior probability distribution given an observed image. Equation 3 present the statistical formulation of pictorial structures model.

$$P(L|I,\theta) \propto P(I|L,\theta) \times P(L|\theta). \quad 3$$

Assume that θ is set of parameters that define an object model, I is the image needed as observations and L specifies the configuration of the object. $P(I|L,\theta)$ indicates the likelihood of seeing a certain image given that an object is at some location. $P(L|\theta)$ is the prior probability that an object is at a specific location. According to Bayes' rule, equation 3 is for the posterior probability$P(L|I,\theta)$.

Parameters of this problem in pictorial structures model is defined as $\theta = (u, E, c)$ where $u = (u_1, \dots . u_N)$ is the object parts appearance parameters, E defines the edges among the parts in graph connected to each other, and c indicates these adjacent parts $c = \{c_{ij} | (v_i, v_j) \in E\}$.

Likelihood of seeing an image given that the configuration of object is product of the individual likelihoods that equation 4 showed it.

$$P(I|L,\theta) = P(I|L,u) = \prod_{i=1}^{N} P(I|l_i, u_i). \quad 4$$

Prior probability is furnishedby a tree-structured Markov random field with edge set *E*.

$$P(L|\theta) \propto \prod_{(v_i,v_j)\in E} P(l_i, l_j | c_{ij}). \quad 5$$

Posterior probability is also obtained through equation 6.

$$P(L|I,\theta) \propto \prod_{i=1}^{N} P(I_i|l_i, u_i) \prod_{(v_i,v_j)\in E} P\left(l_i, l_j \middle| c_{ij}\right). \quad 6$$

In relation 2,$m_i(l_i) = -\log P(I|l_i, u_i)$, is a match costmeasuring how part $v_i$matches the image data at location $l_i$and $d_{ij}\left(l_i, l_j\right) = -\log P(l_i, l_j|c_{ij})$ is a deformation cost measuring how the locations $l_i$and $l_j$agree with the prior model.

III. The Analysis And Comparison of 3D Human Body Pose Estimation Methods Using Pictorial Structures

We introduce human body pose estimation and pictorial structures model in previous section. This section explain several recent methods for 3D human body pose estimation that they used pictorial structures model. Magnus Burenius presented one type of pictorial structures generalized with limb-based model and modelled limitations of human body and considering prior probability for the translation and rotation of each part of the body in order to reduce the search space (Burenius et al, 2013). He considered the following model for the human body.

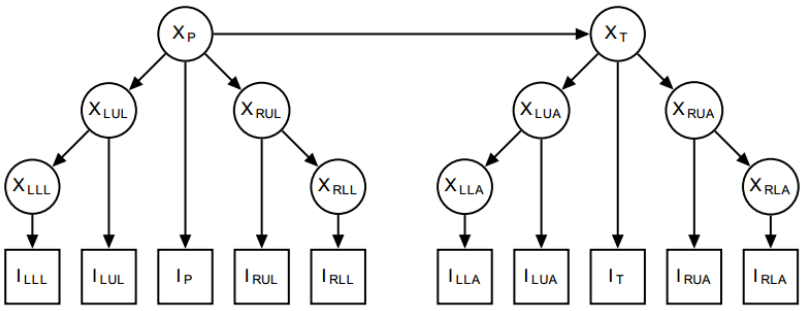

Figure 2: Burenius model for body.The body parts are in topological order: Pelvis, Torso, Left Upper Leg, Right Upper Leg, Left Upper Arm, Right Upper Arm, Left Lower Leg, Right Lower Leg, Left Lower Arm, and Right Lower Arm. The square nodes represent measured variables (Burenius et al, 2013).

He furnished appearance likelihood of each part of the body in each image view using a simple 2D detector which utilizes Histogram of Oriented Gradient (HOG). Then, the 3D likelihood of parts appearance is obtained by assumed that 2D likelihood of each view is independent.

Burenius assumed the parts are connected in a tree graph and the pose of part *n* only depends on the pose of its parent$pa(n)$:

$$P_{X_n|X}(x_n|x) = P_{X_n|X_{pa\,(n)}}(x_n|x_{pa\,(n)}). \qquad 7$$

In the above equation, $X_n = (T_n, R_n)$ which $T_n \in \Omega_T \subset \mathbb{R}^3$is translation of the *n-th* part ($\Omega_T$is search space relating to translation), $R_n \in \Omega_R \subset \mathbb{SO}(3)$is rotation of the *n-th* part (Ω_R is search space relating to rotation), and $X = (X_1, \dots, X_N)$.

The joint distribution of all parts then factorizes as:

$$P_X(x) = \prod_n P_{X_n|X_{pa\,(n)}}(x_n|x_{pa\,(n)}). \qquad 8$$

Observation vector of each part (denote by I_n) in C views has been shown by $I_n = (I_n^1, \dots, I_n^C)$ and supposing that various views are independent, equation 9 would be stated.

$$P_{I_n|X_n}(i_n|x_n) = \prod_c P_{I_n^c|X_n}(i_n^c|x_n). \qquad 9$$

Assuming that $I = (I_1, \dots, I_N)$andI_ns are independent, joint distribution on X and I is stated as equation 10.

$$P_{X,I}(x,i) = \prod_n P_{I_n|X_n}(i_n|x_n)\, P_{X_n|X_{pa\,(n)}}\left(x_n\middle|x_{pa\,(n)}\right). \qquad 10$$

Finally, x^* which states the most likely possible pose for the joints is obtained by maximizing the equation 10. Therefore,

$$x^* = arg\, max_x\, P_{X,I}(x,i). \qquad 11$$

Burenius used skeleton model for prior probability$P_{X_n|X_{pa\,(n)}}(x_n|x_{pa\,(n)})$. In this model, translation- and rotation of each part of the body can be recursively furnished using its parent in tree model. R_n and T_n are rotation and translation of the *n-th* part. $R_{pa\,(n)}$ and $T_{pa\,(n)}$ are the parent rotation and translation. ΔR_n and ΔT_n are local rotation and translation of the *n-th* part in terms of its parent. d_n is a constant and it is equal to distance of the *n-th* part from its parent. As a result,

$$R_n = R_{pa\,(n)} \Delta R_n \qquad 12$$

And

$$T_n = T_{pa\,(n)} + R_{pa\,(n)} d_n + \Delta T_n. \qquad 13$$

Assuming that the person is fixed when obtaining joints pose then the local translations during the time is constant. In order to estimate the body state, the translation and rotation of root and the local rotation of the other joints should be furnished. Burenius assumedthat the prior probability of local rotation is uniform. He reduces the complexity of algorithm from$O(|\Omega_X|^2) = O\left(|\Omega_T|^2 \times |\Omega_R|^2\right)toO\left(|\Omega_T| \times |\Omega_R|^2\right)$.

The symmetric appearance of corresponding left and right body parts is difficult to handle for pictorial structure. So after obtaining x^*, by using equation 11, Burenius performed the algorithm two more times for

correspondingparts. He supposed that the right part is identified correctly and the left part should be furnished and he does this supposition another time for the other parts.

Kazemi used a random forest classifier for obtaining the variation in appearance of body parts in2D images. The result of these 2D part detectors provided the 2D part likelihoods and then aggregated across views toobtain the 3D part likelihoods.For resolving the ambiguity in determining symmetric parts introduced a latent variable into his model which represents the correspondenceof joints across theviews and at inference time he optimized for both the best pose and the best values of this latent variable. Figure 3 present a general overview of Kazemi framework. (Kazemi et al, 2013).

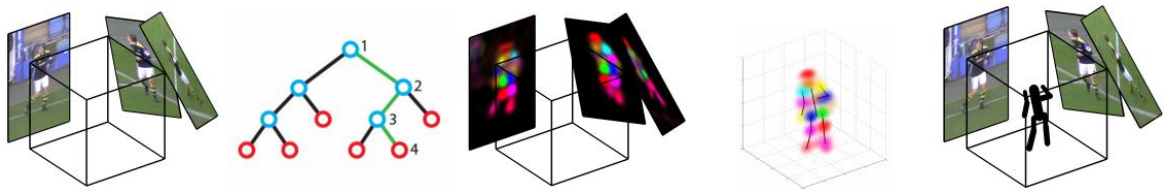

Figure 3: An overview of Kazemi multi-view pose estimation framework. First a random forest is usedto classify each pixel in each image as belonging to a part or the background. Then the results areback-projected to a 3D volume. Corresponding mirror symmetric parts across viewsfounded byintroducing a latent variable. Finally, to estimate the 3D pose used a part-based model. (Kazemi et al, 2013).

In 2013, Sikandar Amin argued that the search complexity can be reduce considerably by formulating the 3D inference problem as a joint inference over 2D projections of the pose in each of the camera views. So he applied the 2D pictorial structure models and estimated the 2D pose then the 3D pose is obtained by triangulation. (Amin et al, 2013).

He introduced pairwise correspondence and appearance terms defined between pair of images. Finally, equation 14 represent the joint posterior over configurations in both views.

$$P(L_1, L_2|I_1, I_2) = \frac{1}{Z} f(L_1; I_1) f(L_2; I_2) \prod_n f_n^{app}(lo_n^1, lo_n^2; I_1, I_2) f_n^{cor}(lo_n^1, lo_n^2). \quad 14$$

In this relation, L_m is the 2D body configuration and the I_m is the image obtained from *m-th* view, $lo_n^m$is the position of *n-th* joint in *m-th* view, $f(L_1; I_1)$ and $f(L_2; I_2)$are the single view factors, and $f_n^{app}(lo_n^1, lo_n^2; I_1, I_2)$is a factor obtained based on the appearance of the both images and $f_n^{cor}(lo_n^1, lo_n^2)$is correspondence terms between their joints.

The factor f_n^{cor} encodes the constraint that part locations in each view should agree on the same 3D position. The factor f_n^{app} encodes the color and shape of the body part seen from multiple viewpoints. Dependency on the camera setup in order to learn pairwise appearance terms is the main weakness of his method.

He extend his approach to a mixture of pictorial structures models by clustering the training data and learning a separate model for each cluster. Figure 4 showed overview of Amin method

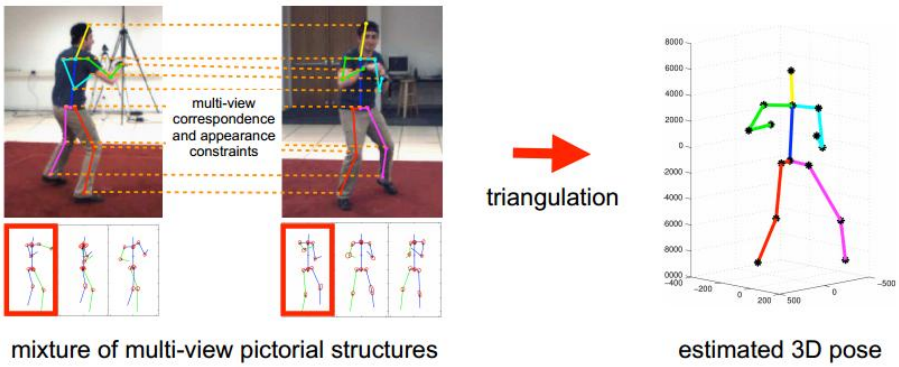

Figure 4: Overview of Amin method, Projections of 3D pose in each view are jointly inferred using a mixture of multi-view pictorial structures models. The body layout priors of each mixture component are visualized below, activated components are highlighted in red. 3D pose is recovered via triangulation (Amin et al, 2013).

Belagiannis (Belagiannis et al, 2014b) presented a 3D pictorial structure for 3D pose estimation of multiple humans from multiple views. His model shown in figure 5. In this figure, the kinematic constrains are shown in green (rotation) and yellow (translation) edges, and the collision constrains are shown in blue edges.

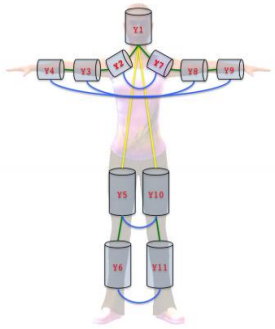

Figure 5: Pictorial structure utilized by Belagiannis (Belagiannis et al, 2014a)

He create a set of 3D body part hypotheses by triangulationof corresponding body joints sampled from the posteriors of 2D body part detectors in all pairs of camera views. Because of unknown personality of humans,triangulation of the corresponding parts of different peoplecreateswrong hypothesesand when different people are in a similar posethat can look correct in the 3D space and can even create a fake skeleton, as shown in Figure 6.

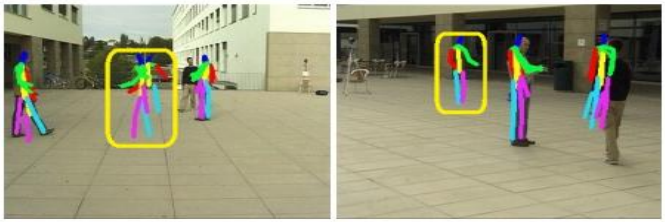

Figure 6: Detection of a fake structure in Belagiannis approach

In his pictorial structure, the unary termsare computed from the confidence of the 2D part-based detectors and reprojection error of the joint of the corresponding body parts. For modelling occlusions and resolving geometrical ambiguities proposed the part length and visibility unary terms. Human body prior represent the relation between the body parts and learned from one camera setup but works it with any other setup. Human body prior got the pairwise terms. So, this way in comparison to Amin waythatthe pairwise termsis independent of the camera setup. He introduced an extra pairwise collision termsto forbid collisions of symmetric body in 3D.

Assume that $Y_i = [X_i^{pr}, X_i^{di}]^T$ is a variable correspondence to *i-th* part of the body, X_i^{pr} is 3-D position of the *i-th* part beginning, and X_i^{di} is ending position of this part the global coordinate system. Each Y_i variable gets value from Λ_i sample space. Furthermore, $Y = (Y_1, \dots, Y_n)$ determines the overall body configuration. Belagiannis computes the posterior probability through equation 15.

$$P(y|x) = \frac{1}{Z(x)} \prod_i^n \phi_i^{conf}(y_i, x) \prod_i^n \phi_i^{repr}(y_i, x) \prod_i^n \phi_i^{vis}(y_i, x) \prod_i^n \phi_i^{len}(y_i, x) \prod_{(i,j)\epsilon E_{kin}} \psi_i^{tran}(y_i, y_j). \quad 15$$

In this relation, x is the observations furnished from the images and y is the body configuration which should be furnished. $Z(x)$ is the partition function, E_{kin} are the edges that representthe kinematic constraints between the body parts, and E_{col} are the edges that represent the collision between symmetric parts. $\phi_i^{conf}(y_i, x)$ is related to confidence of detection, $\phi_i^{repr}(y_i, x)$ isreprojection error and $\phi_i^{vis}(y_i, x)$ is visibility of body part in multi-view and $\phi_i^{len}(y_i, x)$ is body part length. Kinematic constraints on the translation is $\psi_i^{tran}(y_i, y_j)$ and on rotation is $\psi_i^{rot}(y_i, y_j)$. $\psi_i^{col}(y_i, y_j)$ is used to prevent the symmetric parts collision with each other.

In most of the articles in this field, Percentage of Correctly EstimatedPart (PCP) factor has been utilized for evaluating pose estimation accuracy. Suppose that $\widehat{s_n}$ and $\widehat{e_n}$ are the beginning and ending of the *n-th* part available through ground truth data set. s_n is the beginning and e_n is the ending of this part furnished by estimation method.In PCPevaluation, a part is declared correctly estimated if: (Burenius et al, 2013).

$$\frac{\|\widehat{s_n} - s_n\| + \|\widehat{e_n} - e_n\|}{2} \leq \alpha \|\widehat{s_n} - \widehat{e_n}\| \quad 16$$

The PCP score is more informative than one based on the Euclidean distance (Burenius et al, 2013). Recent methods have reported their results on KTH Multi view Football data set. Burenius and Belagiannis has reported the PCP in 2D by projecting the 3D estimation across each view. Burenius expressed the 2D PCPfor $\alpha = 0.5$ and $\alpha = 0.2$ in table 1 and Belagiannis explicated the 2D PCP for $\alpha = 0.5$ in his way with result of Burenius in table 2.

Table1: Theresults of Burenius methodfor poseestimation to real images from 20 different frames. PCP scores in % with $\alpha = 0.5$and $\alpha = 0.2$ (in blue) are used to measure performance of pose estimation using 1, 2 or 3 cameras. First only impose view and skeleton constraints, then add intersection constraints for the lower legs (Burenius et al, 2013).

	View & Skeleton Constraints						**View, Skeleton & Intersection Constraints**					
Parts	**C=1**		**C=2**		**C=3**		**C=1**		**C=2**		**C=3**	
Pelvis	97	57	97	35	100	50	97	57	97	35	100	55
Torso	87	40	90	48	100	65	87	38	90	48	100	55
Upper Arms	14	2	55	8	55	15	14	2	53	8	60	20
Lower Arms	6	0	30	6	35	18	6	0	28	7	35	15
Upper Legs	62	8	87	26	90	45	63	9	88	19	100	48
Lower Legs	33	5	68	35	70	57	41	7	82	38	90	60
All Parts	41	13	67	23	70	39	43	13	69	23	77	40

Table2: The PCP scores, for each camera in BelagiannismethodandBurenius method (Belagiannis, 2014a).

	Burenius	Belagiannis	Burenius	Belagiannis
Body Parts	C2	C2	C3	C3
Upper Arms	53	64	60	68
Lower Arms	28	50	35	56
Upper Legs	88	75	100	78
Lower Legs	82	66	90	70
All Parts (average)	62.7	63.8	71.2	68.0

Kazemi reported 3D PCP result in his method while supposing $\alpha = 0.5$ that shown in table 3.
Table 3: The 3D PCP scores in Kazemi method (Kazemi et al, 2013)

Body parts	3D PCP
Upper Arms	0.89
Lower Arms	0.68
Upper Legs	1.00
Lower Legs	0.99
Average	0.89

IV. Conclusion

human body poseestimation is one of the common issues in machine vision and various methods have been presented for that so far which a category of them has used pictorial structures model and reached high accuracy for this problem. The majority of these approaches have been presented for 2D human body pose estimation. Generalization of pictorial structures to 3D is difficult due to freedom degrees' increase. In this paper, several methods which have a distinctive look toward the issue of 3Dhuman body pose estimation by pictorial structures have been studied and pros and cons of each of them have been stated and reported PCP scores for evaluation.

References

[1] M. A. Fischler and R. A. Elschlager, *The representation and matching of pictorial structures*, IEEE Trans. Comput., no. 1, pp. 67–92, 1973.
[2] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, *Object detection with discriminatively trained part-based models*, Pattern Anal. Mach. Intell. IEEE Trans., vol. 32, no. 9, pp. 1627–1645, 2010.
[3] P. F. Felzenszwalb and D. P. Huttenlocher, *Efficient matching of pictorial structures*, in Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on, vol. 2, pp. 66–73, 2000
[4] P. F. Felzenszwalb and D. P. Huttenlocher, *Pictorial structures for object recognition*, Int. J. Comput. Vis., vol. 61, no. 1, pp. 55–79, 2005.

[5] P. F. Felzenszwalb and D. P. Huttenlocher, *Distance Transforms of Sampled Functions.*, Theory Comput., vol. 8, no. 1, pp. 415–428, 2012.
[6] M. Burenius, *Human 3D Pose Estimation in the Wild: using Geometrical Models and Pictorial Structures*, Doctoral thesis, KTH, School of Computer Science and Communication, 2013.
[7] M. Burenius, J. Sullivan, and S. Carlsson, *3d pictorial structures for multiple view articulated pose estimation*, in Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, pp. 3618–3625, 2013.
[8] Kazemi, M. Burenius, H. Azizpour, and J. Sullivan, *Multi-view body part recognition with random forests*, in 2013 24th British Machine Vision Conference, United Kingdom, 2013.
[9] S. Amin, M. Andriluka, M. Rohrbach, and B. Schiele, *Multi-view pictorial structures for 3d human pose estimation*, in British Machine Vision Conference, 2013.
[10] Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic, *3D pictorial structures for multiple human pose estimation*, in Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, pp. 1669–1676, 2014.
[11] V. Belagiannis, X. Wang, B. Schiele, P. Fua, S. Ilic, and N. Navab, *Multiple Human Pose Estimation with Temporally Consistent 3D Pictorial Structures*, in *Computer Vision-ECCV 2014 Workshops*, pp. 742–754, 2014.