Dirichlet-tree Cascaded Hough forests for Continuous Head Pose Estimation

Yuanyuan Liu[1,2,3]
[1]National Engineering Research Center for E-Learning, Central China Normal University, Wuhan, China
[3] Wenhua College, Wuhan, China

Jingying Chen[*1,2]
[2]Collaborative & Innovative Center for Educational Technology (CICET) Central China Normal University, Wuhan, China

Haiqing Chen[4]
[4]Hankou College, Wuhan,China

Abstract—**We propose a hierarchical regression approach, Dirichlet-tree cascaded Hough forests (DCHF), which is based on deep learning for continuous head pose estimation in unconstrained environment, e.g., poses, illumination, occlusion, low image resolution, expressions and make-up. First, positive facial patches are learned and extracted from facial area to eliminate the influence of noise. Then, in order to estimate continuous head pose efficiently, multiple probability models are learned in four layers of the DCHF, i.e., the patch's classification, the head pose angles, and offset probabilities mapping in the Hough space in a hierarchical way. Moreover, our algorithm takes a weighted and cascaded Hough voting method, where each positive patch extracted from the face can cast the efficient vote for head pose estimation. Experimental results on different public databases demonstrate the robustness and accuracy of the proposed approach to continuous head pose estimation.**

Keywords-component; continuous head pose estimation; DCHF; deep and hierarchical learning; weighted and cascaded Hough voting

I. INTRODUCTION

Head pose estimation is often the first step for many applications like face recognition, facial expression analysis and visual focus of attention recognition [1, 4, 7]. While most of existed methods are detected on images in constrained environment, some recent works in head pose estimation have been extended to deal with more challenging face images collected "in the wild"[5,7]. However, 2D continuous head pose estimation in the unconstrained environment remains challenging due to high variations in facial appearance, poses, illumination, occlusion, expressions and make-up.

The approaches to head pose estimation are divided into two categories: model-based approach and appearance-based approaches. Local features and facial geometric model are combined to obtain precise head pose using model-based approach [2]. But they are hard to deal with images with low-resolution. Appearance-based methods are based on machine learning methods to estimate head pose from the entire facial area, such as multi-detector methods [3], manifold embedding methods [4], and non-linear regression methods [6, 9]. In recent year, non-linear regression Random Forest (RRF) becomes a popular algorithm in computer vision because of their powerful calculation capability, high performance, and easy implementation [7, 9, 14]. Gall and Lempitsky proposed the Hough forest from RF to detect objects [1]. Related work showed that Hough forest can map the features to vote in a generalized Hough space [1, 5]. Hough forest can learn a direct mapping between the appearance of input image patch and its Hough vote. Furthermore, Zhang [5] used Hough forest to estimate head pose in the low-resolution images and proved its more powerful capability of classification and regression than random forest. But the occlusion is not solved in the situation. In order to improve the efficiency and robustness, we propose Dirichlet-tree cascaded Hough forests (DCHF) for continuous head pose estimation in unconstrained environment in a hierarchical way.

The Dirichlet-tree distribution was proposed by Minka [9]. It is the distribution over leaf probabilities that result from the prior on branch probabilities. Minka proved the high accuracy and efficiency of the Dirichlet-tree distribution in [9]. Some researchers use a Dirichlet-tree distribution in multi-objects tracking [15] and head pose classification [17]. In this work, the DCHF is proposed to estimate continuous head poses in vertical and horizontal directions in unconstrained environment. And the framework of DCHF can be shown in Figure 1. The procedure of continuous head pose estimation has four hierarchical layers, such as D-L1, D-L2, D-L3, and D-L4. D-L1 and D-L2 are two layers in the horizontal direction, D-L1 represents coarse regression while D-L2 is refined regression. D-L3 and D-L4 are two layers in the vertical direction, D-L3 represents cascaded regression under the two prior layers, while D-L4 represents final continuous head pose regression in the two directions.

978-1-4799-5835-1/14/$31.00 ©2014 IEEE

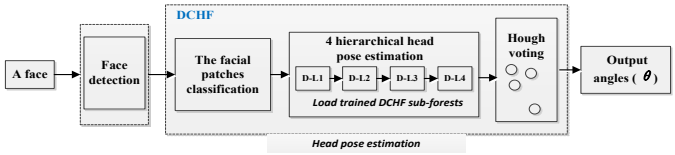

Figure 1 The flowchart of head pose estimation using the DCHF

The main contributions of this paper are as follows. First positive facial patches are learned and extracted to eliminate the influence of noise. Then, in order to estimate continuous head pose efficiently, multiple probability models are learned in four layers of the DCHF, i.e., the patch's classification, the head pose angles, and probabilities mapping in the Hough space in a hierarchical way. Furthermore, our algorithm takes a weighted and cascaded Hough voting method, where each positive patch can cast the efficient vote for the head pose. Details are discussed in the following sections.

The rest of the paper is organized as follows: Sections 2 details about Dirichlet-tree cascaded Hough forests; Section 3 describes continuous head pose estimation in a hierarchical way; Section 4 presents the experimental results and discussions; Section 5 gives the conclusions.

II. DIRICHLET-TREE CASCADED HOUGH FORESTS

Random forests are an ensemble machine learning algorithm where many random decision trees are combined together for powerful classification or regression. Each tree in random forests independently grows with random samples from the dataset. Hough forests [1, 5, 10] combined random forests with Hough voting and improved random forests to learn a direct mapping in generalize Hough space (see Figure 2(b)). They have power capability of mapping input spaces into discrete or continuous output spaces.

From Figure 2(a), we can see that the Dirichlet-tree is the distribution over leaf probabilities $[p_1 \ldots p_i]$ that result from this prior node probabilities $[a_1, a_2, \ldots a_k]$ on branch probabilities b_{ji} [9], where i is the number of a leaf , k is the number of a prior node, and j is the layer of a branch. The Dirichlet-tree distribution algorithm is introduced into the framework of Hough forests as shown in Figure 2(c). Hough forests are reconstructed to a Dirichlet-tree structure and become the DCHF.

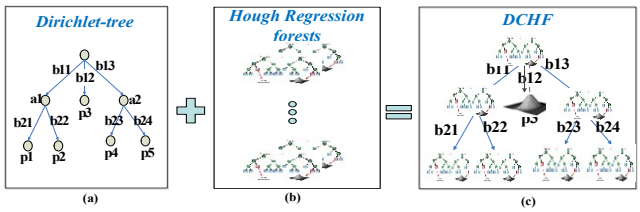

Figure 2: A general DCHF.

A. Training

Each tree T in the forest $\{T_t\}$ is built and selected randomly from a different dataset. From a selected image from the dataset, we extract several facial patches $P_i=\{I_i, c_i, \theta_i, d_i\}$. Where I_i represents the appearance, θ_i represents the set of head pose angles associated to each rotation freedom, $d_i=||N-P_i||_2$, denote the offsets from the centroid of each patch P_i to tip of the nose N.

In the work, the patch appearance I_i is defined as multiple channels $I=\{I_i^1, I_i^2, I_i^3, I_i^4\}$. I_i^1 contains grey values of the raw facial patch with dimension as 30*30. I_i^2 represents the Gabor features based PCA of the facial patch with dimensions as 35*12. I_i^3 represents LBP features of the patch with dimensions 30*30, I_i^4 contains the Sobel edge features in horizontal and vertical directions as 2*30*30. $c_i=\{0,1\}$ contains the patch's information, only the positive patch with $c_i=1$ can be used to estimate head pose. The set of $\theta_i = \{\theta^x_{yaw}, \theta^y_{pitch} \mid \theta^{x\text{-}1}_{yaw}, \theta^{y\text{-}1}_{pitch}\}$ contains head pose parameters in horizontal and vertical directions, where $\theta^x_{yaw}, \theta^y_{pitch}$ are the head rotation angles in the x-th and y-th layer of the DCHF , and $\theta^{x\text{-}1}_{yaw}, \theta^{y\text{-}1}_{pitch}$ are the angles in the prior layer and node of the DCHF.

We define a patch comparison feature as our binary tests φ, similar to [7, 9],

$$\varphi = |R_1|^{-1} \sum_{j \in R_1} I^f(j) - |R_2|^{-1} \sum_{j \in R_2} I^f(j) > \tau \quad (1)$$

Where R_1 and R_2 are two random rectangles within the positive facial patches, $I^f(j)$ is the feature channel $f = \{1,2,3,4\}$ and τ is a predefined threshold.

In order to train a sub-forest in different layers of the DCHF, we divide the set of patches P into two subsets P_L and P_R for each φ,

$$P_L = \{P \mid \varphi < \tau\}, \; P_R = \{P \mid \varphi > \tau\} \quad (2)$$

Where φ is the patch comparison feature (see Eq.(1)) and τ is a predefined threshold.

Selecting the splitting candidate φ when maximizing the evaluation function Information Gain *(IG)*, *IG* is defined as

$$IG = \arg\max_{\varphi} (H(P \mid \theta^{L-1}) - (\omega_L H(P_L \mid \theta^{L-1}) + \omega_R H(P_R \mid \theta^{L-1}))) \quad (3)$$

Where w_R, w_L represent respectively the ratio between the number of samples in the set P_L (arriving to left subset using upper binary tests), set P_R (arriving to right subset using upper binary tests) and set P (total node samples).

We model the vector θ_i at each node in the DCHF as realizations of a random variable with a multivariate Gaussian distribution. Therefore, $H(P|\theta^{L-1})$ can be written as,

$$H(P \mid \theta^{L-1}) = -\sum_{i=1}^{N} \frac{\sum_i p(\theta_i^L \mid \theta_i^{L-1}, c_i, P_n)}{|P|} \log\left(\frac{\sum_i p(\theta_i^L \mid \theta_i^{L-1}, c_i, P_n)}{|P|}\right) \quad (4)$$

where $p(\theta_i^L \mid \theta_i^{L-1}, c_i, P_n)$ indicates the probability that the patch P_n represents the head pose θ_i in the L-th layer of the DCHF under the prior layer L-1 and the patch information c_i.

Maximizing Eq.(3) can obtain the best tests by minimizing the determinant of the covariance matrix $\sum^{\theta}$, which declines the voting impurity for the output angles.

Create leaf l when IG is below a predefined threshold or when a maximum depth is reached. Each leaf stores multiple models, i.e., the head pose angle and offsets from the facial patch to tip of the nose. Otherwise continue iteration for the two subsets P_L and P_R is performed as Eq.(1).

B. Testing

When testing, the patches are evaluated according to the trained tree and passed either to the right or left child node in the t ree until a leaf is reached. By passing all the patches $P_i(c_i=1)$ down all the trees in the DCHF for head pose estimation, the patches $P_i(c_i=1)$ end in a set of leaves of different sub-forests of the DCHF instead of all leaves of random forest. In each leaf l, there are multiple probable models, i.e., the negative/positive patch's class label ($c_i=1$ means that only facial foreground patch reach the leaf), the head pose parameter (θ_i), the list displacements d_i corresponding to the patch, which are used to cast probabilistic Hough votes about the head pose.

The final head pose angles are then obtained by performing weighted Hough voting in the final layer of the DCHF.

III. CONTINUOUS HEAD POSE ESTIMATION

For head pose estimation, face detection under various poses and wide angles is challenging. We firstly use Adaboost with Haar-like features [13] to detect the facial area. Based on this, the DCHF are correspondingly learned to estimate continuous head pose in a hierarchical way. More details will be introduced in the following sub-sections.

A. Facial patch classification

The extracted face area may include occlusion or noise. In order to eliminate occlusion and noise, the facial patches are segmented into the positive and negative patches as shown in Figure 3. The positive patches consist of real facial patches that contribute to estimate head pose, while the negative areas may introduce errors to the task. The modeling procedure is the first layer in the DCHF. Positive patches from the sub-region are labeled as $c=1$ and others as $c=0$. The training and testing are similar to Section 2. When a test patch P arrives at a leaf, we use the probability $p(c \mid l_t(P))$ stored at a leaf to judge whether the test patch belongs to the positive facial area. Only the positive patches are used for head pose estimation.

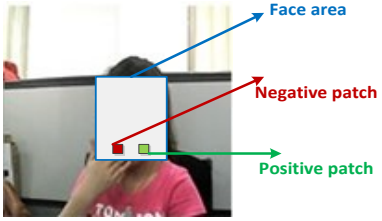

Figure 3: The positive and negative patch

B. Continuous head pose estimation using the DCHF

In order to obtain continuous head pose estimation in the horizontal and vertical directions in unconstrained environment, DCHF is trained as described in Section 2 like a Dirichlet tree-structure regression. The Dirichlet-tree distribution can obtain adaptive input variable value result from the prior node probability instead of uniform values.

$$p(\theta(\theta_{yaw},\theta_{pitch}) \mid \{T_L\}_{t=1}^{T}) = \begin{cases} \frac{1}{T}\sum_{t=1}^{T} p(\theta^{L} \mid c=1,T_L) < \frac{1}{T}\sum_{t=1}^{T} p(\theta^{L-1} \mid c=1,P) & \text{go to the right regression} \\ \frac{1}{T}\sum_{t=1}^{T} p(\theta^{L} \mid c=1,T_L) = \frac{1}{T}\sum_{t=1}^{T} p(\theta^{L-1} \mid c=1,P) & \text{output} \\ \frac{1}{T}\sum_{t=1}^{T} p(\theta^{L} \mid c=1,T_L) < \frac{1}{T}\sum_{t=1}^{T} p(\theta^{L-1} \mid c=1,P) & \text{go to left regression} \end{cases} \tag{5}$$

Where $\{T_L\}_{t=1}^{T}$ represents the sub-forest in the *L-th* layer of DCHF, T is the number of Hough trees in the sub-forest. Figure 4 shows the images for each horizontal regression in D-L1 and D-L2 of the DCHF in a hierarchical way. When the yaw angles have been estimated by the regressions in D-L2， the same deep learning regressions in D-L3 and D-L4 should estimate the pitch angles in the vertical direction. The continuous head pose angles can be obtained by Eq.(5). The final head pose angles are obtained when all positive patches reached the leaf nodes in the cascaded sub-forests from the DCHF instead of all leaves in DCHF.

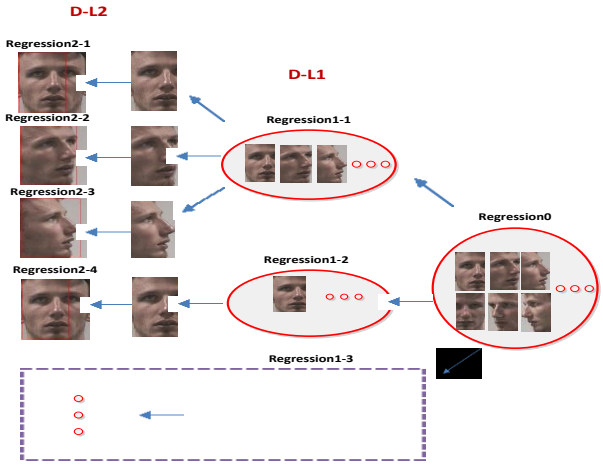

Figure 4: Positive patches for horizontal regression learning using the DCHF.

C. Weighted and cascaded Hough voting

We use a weighted Hough voting method in a cascaded way. Both classification and regression voting are used to the DCHF. In order to eliminate imbalance of samples, we also store the weight $w_s=P_S/P$ that is defined as the ratio of the number of samples in each subset P_S and the total number of samples P in each single tree of the DCHF.

The Eq. (5) defines the probabilistic votes cast by a single sub-forests from DCHF for head pose estimation. To integrate the votes coming from different patches in different sub-forests, we accumulate them into a 3D Hough image $V(y_i,d,,\theta)$, in which for each pixel-position y_i, the votes (see Eq.(6)) add up in the sub-forest $\{T_L\}_{t=1}^{T}$:

$$V(y_j) = \sum p(d_i \mid c=1;\{T_L\}_{t=1}^{T}) \tag{6}$$

The procedure of estimation calculates the Hough image V and deletes unwanted patches with minor probabilities. The $V(y_j)$ values represent the confidence measures for each hypothesis vote. In a Hough image, mean-shift is used to find the maxima in an alternative way as Hough voting-based frameworks [1, 16]. If votes for the head pose θ are $\theta_{yj}(\theta^L| c, \theta^{L-1})$ in the patch location y_j, then we set the weighted Hough voting model as

$$V(\theta) \propto K((w_S V(y_j) - (y_i + \overline{w_S V(y_j)}))/h_j) \quad (7)$$

A Gaussian Kernel K and the bandwidth parameter h_j are given by Gaussian filter. In θ_{yi}, c=1 is the positive facial patch, $\theta^{L-1}=\{\theta^{x-1}_{yaw}, \theta^{y-1}_{pitch} |L\text{-}1\}$ represents the Hough voting result for head pose in a Hough image. Then regression voting provide good results by evaluating sparse patches in the cascaded sub-forests, rather than all patches in the forest.

IV. EXPERIMENTS

In this section, we thoroughly evaluate the proposed DCHF framework for the task of continuous head pose estimation in unconstrained environment. This subsection provides experimental results under various conditions, such as pointing'04 head pose database [13], LFW database [14] and our lab real-time database (see Figure 5). The Pointing'04 database consists of 2940 images with different poses and expressions. The LFW database consists of 5749 individual facial images and 13300 images. The images have been collected 'in the wild' and vary in poses, lighting conditions, resolutions, races, occlusions, and make-up. Our lab database has been collected using 20 different persons with different poses, expressions and occlusions, and the reference angles have been annotated using the method similar to LFW [21].

Figure 5: Examples of images from there databases, Pointing'04 database (the first row), LFW database (the second row), and our lab database (the third row).

For evaluation, we divided the databases into a training set and a testing set. The training set includes 2100 images from Pointing'04 database, 12000 images from LFW databases and 300 images from our lab database. The testing set includes the rest of 840 images from Pointing'04 database, 1500 images from LFW database and 200 images from our lab database.

A. Training

For continuous head pose estimation, we trained trees of the DCHF with the three databases. We fixed some parameters on the empirical values, e.g., the maximum depth of the tree is 15 and the splitting candidate is 2000 and the thresholds is 25. Other parameters include the number of patches extracted from each image (fixed to 200) and the patch's size.

Figure 6 describes performance of the algorithm when we varied the size of the facial patches used for training each tree. In Figure 6, the blue, continuous line shows the percentage of estimation accuracy as the patch size, when 300 train images per trees are used. The red, dashed line shows instead the percentage of false estimation rate as the patch size. The plot shows that a minimum size for the patches is critical since small patches can not capture enough information to reliably predict the head pose. However, there is also a slight performance loss for lager patches. In the case, the trees become more sensitive to occlusions since the patches cover a large region and overlap more. Having a patch size between 20*20 seems to be a good choice where the patches are discriminative enough to estimate the head pose.

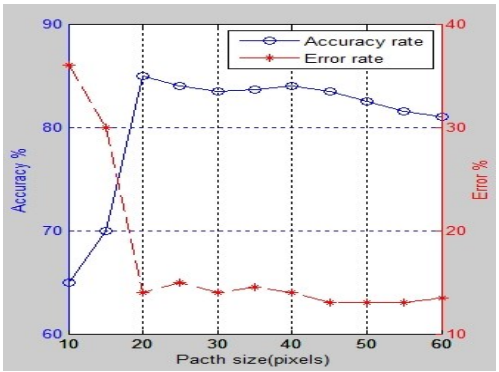

Figure 6: Estimation accuracy depending on the patch size overlaid to the mean error rate.

Each tree grows based on a randomly selected subset of 186 images. Sub-forests in different layers of the DCHF have been trained independently in a hierarchical way from the first layer to the fourth layer. There are 15 trees in the first layer D-L1 of DCHF, 15 trees in the second layer D-L2 and 5 trees in per sub-forest under the prior node. In the third layer D-L3, we trained 15 trees, where 3 trees in each head pose class under the node in D-L2. In the fourth layer D-L4, we trained 30 trees totally, where 2 trees in each head pose under the prior layers.

B. Testing

Testing parameters include the Hough forest parameters, the Dirichlet-tree layer numbers, and the Gaussian model parameters. The Hough forest parameters for testing are similar to training in all of our experiments. Other parameters are adaptively selected during testing. Firstly, the faces detected by our trained Adaboost with Haar-features have been normalized to 125*125 pixels. Then 200 patches were densely extracted from the face area and were extracted multiple feature channels and sent them down through all layers in the DCHF in a hierarchical way. In order to eliminate negative

influence of occlusion and noise, only the positive facial patch can be used to estimate head pose in D-L1 to D-L4.

In order to evaluate the performance of our proposed approach, we performed the procedure in 10 times and evaluated the approach using the averaged results.. In the face detection step, we trained our detector under various conditions and have achieved the average detect rate of 93.7%. In the head pose estimation step, head pose angles are obtained by the trained Hough sub-forests in cascaded layers. Some successful estimation examples using the DCHF are shown in Figure 7.

Figure7： Example results of three databases in unconstrained environment using DCHF.

C. *Comparison with state of the art*

The mean absolute error (MAE) is used to evaluate the accuracy rate in the case. The estimated yaw and pitch angles have been compared with the results estimated by state of the art, i.e., random regression forests (RRF) [6], Hough forests (HF) [5]. Also the head pose estimators are performed with positive patch extraction and without positive patch extraction. The comparison results are shown in Table I. From the Table I, one can see that our proposed DCHF with positive patch extraction preforms better than other typical algorithms in unconstrained environment.

TABLE I. RESULTS COMPARISON WITH TYPICAL ALGORITHMS ON THREE DATABASES

Approaches	**Mean absolute error(°)**		
	Different algorithms	***Yaw***	***Pitch***
Positive patch extraction	DCHF	9.6	11.3
	RRF	14.5	16.8
	HF	12.5	13.6
Without	DCHF	11.7	12.2
positive patch extraction	RRF	16.1	17.8
	HF	13.6	14.9

D. *Computation time*

The experiments have been conducted on a PC with Intel(R) Core(TM) i5-2400 CPU@ 3.10GHz. The comparison of computation time is given in Table II. From the table, one can see that the DCHF is faster than the RRF and HF.

TABLE II. COMPUTATION TIME COMPARISION

Approaches	**Computation time (s)**		
	Yaw	***Pitch***	***Total***
DCHF	0.370242	0.352829	0.723071
RRF	0.496753	0.471794	0.968547
HF	0.502315	0.554590	1.056905

V. CONCLUSIONS

In this paper, we propose a robust and efficient approach for continuous head pose estimation in the vertical and horizontal directions under various conditions. First positive facial patches are learned and extracted to eliminate the influence of noise. Then, in order to estimate continuous head pose efficiently, multiple probability models are learned in four layers of the DCHF, i.e., the patch's classification, the head pose angles, and probabilities mapping in the Hough space in a hierarchical way. Moreover, a weighted and cascaded Hough voting method is proposed to select efficient votes for the head pose. Experimental results show the DCHF performs more accurate and efficient than the RRF and HF with positive facial patch extraction. In future work, more experiments will be conducted to evaluate the method's performance under different noise. Also, this approach could be used to estimate the head pose in a wide scene, e.g. the students' attention in a classroom.

ACKNOWLEDGMENT

This work was supported by the National Key Technology Research and Development Program (No.2013BAH72B01) and research funds from Ministry of Education and China Mobile (MCM20130601) and (MCM20121061), Research Funds from the Humanities and Social Sciences Foundation of the Ministry of Education (No. 14YJAZH005)Research Funds of CCNU from the Colleges' Basic Research and Operation of MOE (CCNU13B001), Wuhan Chenguang Project (2013070104010019), Central China Normal University Research Start-up funding (No.: 120005030223) , the Scientific Research Foundation for the Returned Overseas Chinese Scholars (No.:(2013)693), Young foundation of WenHua college (J0200540102), National Natural Science Foundation of China (No.61272206), Hubei province natural science foundation(No. 2013CFB209).

REFERENCES

[1] 'J.,Gall, V.,Lempitsky, "Class-specific hough forests for object detection," Decision Forests for Computer Vision and Medical Image Analysis. Springer London, 2013, pp.143-157.

[2] C. Huang, H.Ai, Y. LI and S.Lao, "High-performance rotation invariant multiview face detection", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.29, No.4, 2007, pp. 671-686.

[3] S. Yan, H. Wang, Y. Fu, J. Yan, X. Tang, and T. S, "Huang, "Synchronized Submanifold Embedding for Person-Independent Pose Estimation and Beyond," IEEE Trans. On Image Processing.Vol.18, No. 1, 2009, pp. 202-210, Jan.

[4] D. Zhu, X. Ramanan, "Face detection, pose estimation and landmark localization in the wild," In Proc. IEEE Conf. CVPR, 2012.

[5] M. Zhang, K. Li, Y. Liu, "Head pose estimation from low-resolution image with Hough forest," Pattern Recognition (CCPR), 2010 Chinese Conference on. IEEE, 2010, pp.1-5.

[6] Y.Li, S.Wang, X.Ding, "Person-independent head pose estimation based on random forest regression," In Image Processing (ICIP), 2010 17th IEEE International Conference on. IEEE, 2010, pp.1521-1524.

[7] M. Dantone, J. Gall, G. Fanelli, L. Van Gool, "Real time facial feature detection using conditional regression forests," In CVPR, 2012.

[8] T. Minka. "The dirichlet-tree distribution," Paper available online at: http://research.microsoft.com/minka/papers/dirichlet/minkadirtree.pdf, 1999.

[9] G.Fanelli, T.Weise, J.Gall, and L.Van Gool, "Real time head pose estimation from consumer depth cameras," in DAGM, 2011.

[10] S. Schulter, C. Leistner, P. M. Roth, et al, "On-line Hough Forests," in BMVC. 2011, pp.1-11.

[11] N.Gourier, D.Hall, and J.Crowley, "Estimating Face Orientation from Robust Detection of Salient Facial Features," in Pointing 2004. ICPR international Workshop on Visual Observation of Deictic Gestures, 2004, pp. 183–191.

[12] G. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Technical report, University of Massachusetts, Amherst, 2007.

[13] M. Jones, P. Viola, "Fast multi-view face detection," Tech. Rep. TR2003-096, Mitsubishi Electric Research Laboratories ,2003.

[14] L. Breiman. Random forests. Machine Learning, 2001: 45(1):5–32,.

[15] X., Yan, C., Han, "Mutiple Target Tracking by Probability Hypothesis Density Based on Dirichlet Distribution," Journal of Xi'An JiaoTong University. Vol.45 No2. 2011.

[16] O. Barinova, V. Lempitsky, and P. Kholi, "On detection of multiple object instances using hough transforms," TPAMI, 2012.

[17] Y. Liu, J. Chen, Y. Liu, Y. Gong, and N. Luo, "Dirichlet-tree Distribution Enhanced Random Forests for Head Pose Estimation," In ICPRAM, 2014, pp.87-95.