21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Face Pose Estimation with Combined 2D and 3D HOG Features

Jiaolong Yang, Wei Liang, Yunde Jia
Beijing Laboratory of Intelligent Information Technology
School of Computer Science, Beijing Institute of Technology, Beijing 100081, P.R. China
{*yangjiaolong, liangwei, jiayunde*}*@bit.edu.cn*

Abstract

This paper describes an approach to location and orientation estimation of a person's face with color image and depth data from a Kinect sensor. The combined 2D and 3D histogram of oriented gradients (HOG) features, called RGBD-HOG features, are extracted and used throughout our approach. We present a coarse-to-fine localization paradigm to obtain localization results efficiently using multiple HOG filters trained in support vector machines (SVMs). A feed-forward multi-layer perception (MLP) network is trained for fine face orientation estimation over a continuous range. The experimental result demonstrates the effectiveness of the RGBD-HOG feature and our face pose estimation approach.

1. Introduction

Many approaches have been proposed to achieve fast and robust face pose estimation over the past two decades. In a technical taxonomy, there exist four main categories of face pose estimation methods. *Multi-detector* methods [5][6] train a serious of detectors related to discrete poses or pose groups and assign a pose sample to the detector with greatest support. *Nonlinear regression* methods [3][10][12] build the mapping between the pose samples and the pose measurement using nonlinear regression. *Motion-based methods* [11][14] or tracking methods track the face in successive frames and recover the motion parameter with registration techniques. *Model fitting* methods [7][13] fit some face models to the samples and estimate the pose from the model parameters. Besides, these methods may also be divided depending on the type of data they require, i.e., 2D images or depth data.

The advantages of nonlinear regression methods such as MLP [10][12] and locally-linear map [9] are that they are very efficient and give some of the most accurate face pose estimates in practice [8]. However, they are very sensitive to localization errors. Multi-detector methods can perform head localization and pose estimation simultaneously, but they can only provide coarse and discrete estimation results. www.baidu.com

In this paper, we propose a three-step face pose estimation approach using both nonlinear regression method and multi-detector method on the 2D and 3D data from a Kinect sensor. Initially, nine detectors corresponding to nine different head orientation groups are used to get the coarse location of the face. These detectors are trained using linear SVMs. In order to achieve more accurate localization results, a refining search on the image coordinate and scale is then implemented based on the coarse localization result. Localization errors are reduced into an acceptable scope after the refining step. Finally, a feed-forward MLP network is applied to achieve orientation estimation results. HOG features are extracted from both 2D and 3D data in face localization and pose estimation. To the best of our knowledge, this is the first trial using HOG features for the task of face pose estimation.

2. Approach

We first introduce the HOG representation of the color image and depth data in our approach. Then the three steps for face pose estimation are described.

2.1. RGBD-HOG Representation

We define a dense HOG representation of an image similar to the construction in [1]. The image is divided into 8×8 non-overlapping pixel cells. Gradient orientations over pixels are accumulated into a histogram in each cell. The gradient of each pixel is discretized into one of nine orientation bins. For color images, the channel with highest gradient magnitude is used at each pixel. In each cell, four different normalization factors are used to normalize the histogram in a cell to build a 9×4 dimensional feature. These factors measure

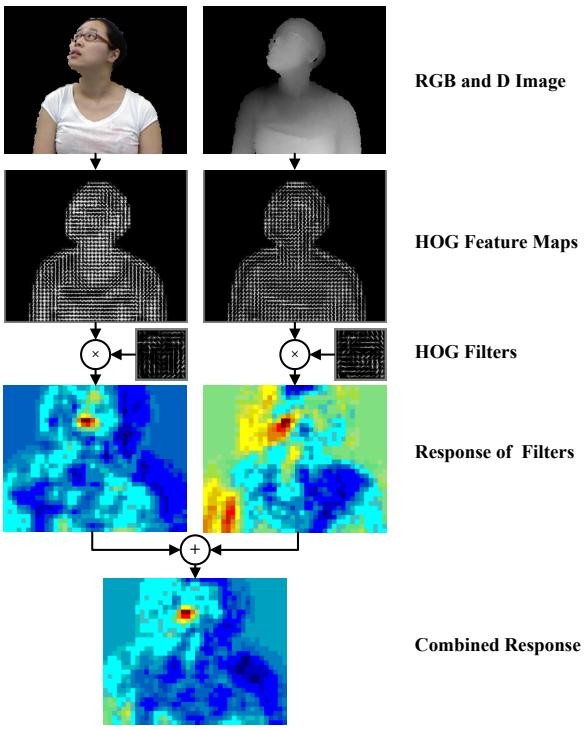

Figure 1. Detection process using filters.[1]

the total energy in a square block of four cells containing the specific cell. An analytic dimension reduction based on principal component analysis (PCA), which is performed in a similar way to that of Felzenszwalb *et al.* [4], is applied to get a 9-dimensional feature for one cell with no noticeable loss of information.

The Kinect sensor captures color images and depth images in pairs. HOG features are extracted from these pairs. Features of a cell pair holding the same location in the color and depth image are combined as one 2D+3D feature, or *RGBD-HOG feature*, with 9×2 dimensions. All of our processes are conducted on RGBD-HOG features. Experiments show that the use of RGBD-HOG features achieves better face pose estimation results than using color features and depth features separately.

2.2. Coarse Localization

Let H be an RGBD-HOG feature map, and (x, y) denote the position of a particular cell. $H(x, y, w, h)$ is the vector with $w \times h \times 9 \times 2$ dimensions obtained by concatenating the RGBD-HOG features in a $w \times h$ sub-window with top-left corner at (x, y). An RGBD-HOG filter is a rectangular template specifying weights for such a sub-window. A w by h filter F is also a vector of $w \times h \times 9 \times 2$ weights. The response of an RGBD-HOG filter on a feature sub-window is defined as

$$R(x, y) = F \cdot H(x, y, w, h) + b, \quad (1)$$

where R denotes the response map and b denotes the bias. Figure 1 illustrates the detection process using an RGBD-HOG filter (the RGB and Depth images are from [3]).

In order to obtain accurate localization results with HOG features, Dalal et al. [2] run detection on multiple scales to get multiple candidates and chose the one with strongest response. However, the process of multi-scale detection on the whole image is very time-consuming. In our approach, we choose the scale based on the depth data to get coarse localization results. Images are scaled such that people are at roughly the same distance before the camera.

In the training stage, images are scaled as described above. Yaw and pitch angles are discretized into three intervals of <-20°, [-20°, 20°] and >20°, dividing the images into nine groups. We train nine RGBD-HOG filters corresponding to the nine groups using one-against-others strategy. Face patches are cropped out as positive samples for each pose group. Negative samples for group i involves positive samples in group 1, ..., i-1, i+1, ..., 9 and randomly selected patches outside the face region. RGBD-HOG features are extracted from these samples and filters are trained in linear SVMs.

In the detection stage, the sample with a left-top corner at position (x, y) is classified as class k,

$$k = \begin{cases} 0, & \max_i R_i(x, y) \leq 0 \\ \arg\max_i R_i(x, y), & \max_i R_i(x, y) > 0 \end{cases}, \quad (2)$$

where $R_i(x, y)$ denotes the response of the RGBD-HOG filter for group i, and k = 0 indicates a non-face region. A simple non-maximum suppression process is then performed for fusion of overlapping detections.

2.3. Location Refinement

The coarse localization solution is based on an important hypothesis: the detector should give a strong positive response even if the detection window is slightly off-center or off-scale on the object. The primaries of the feature maps are 8×8 cells. During detection, the scanning window of the filter moves cell by cell, which means that 7 pixels will be skipped. Without multi-scale solution, expectation of the localization error is large even if the detector is ideally robust. Besides, the scale we choose based on the depth information is not absolutely reliable since size of the face differs from person to person. So the location should be refined.

[1] It's identical to filter on RGB-HOG features and Depth-HOG features with two filters after which two response maps are added, or to filter on the RGBD-HOG features with an RGBD-HOG filter directly. For convenience we illustrate with the former process. But in practice the latter one is used and we train one RGBD-HOG filter instead of training two filters separately.

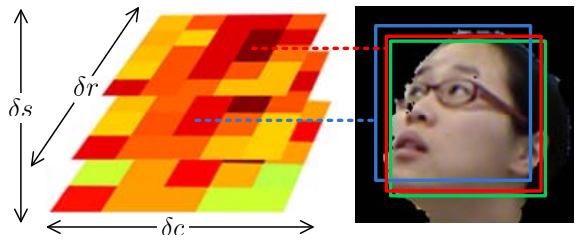

Figure 2. Location refinement with the response pyramid. The blue and red rectangle indicates the face region before and after refinement and the green rectangle indicates the ground truth.

In the refining step, we examine the responses of the selected filter in the coarse localization step pixel by pixel on multiple scales. Suppose we get a detection result at position (x, y) in the feature map, the left-top corner of the face rectangle in the image will be at (r, c) where $r = 8\times x$ and $c = 8\times y$. We move the corner of the rectangle to $(r + \delta r, c + \delta c)$ where δr, $\delta c \in [-d, d]$ and $d \leq 4$. New RGBD-HOG features and the response will be computed after each movement. Besides, the image will be scaled with the factor $1+\delta s$. The search of the highest response in the response pyramid is performed. Influence of different values of d is discussed in Section 3. Let $(\delta r^*, \delta c^*, \delta s^*)$ denotes the optimal displacement for the face window, the final location after refinement is given by

$$\begin{cases} r^* = (r + \delta r^*)/(1 + \delta s^*) \\ c^* = (c + \delta c^*)/(1 + \delta s^*) \end{cases} . \quad (3)$$

Figure 2 illustrates that the localization error can be reduced by choosing the displacement with the highest response.

Note that thought exhaustive search in position and scale is very time-consuming, the computational cost is low since the process is conducted on a small sub-image which is slightly larger than the face region.

2.4. Orientation Estimation

A neural network solution for the task of pose estimation is chosen based on the fact that our system can locate the face with low error so advantages of this non-linear regression method can be used to obtain fine estimation results.

We train a MLP network for face orientation estimation over a continuous orientation range. The network follows a three-layered, feed-forward topology. The input layer has $w\times h\times 9\times 2$ neurons which is the same as the number of weights in an RGBD-HOG filter, so the RGBD-HOG features of a localization result can be fed into the network directly without extra computation of features. The hidden layer includes 10 hidden neurons and the output layer has three neurons corresponding to the yaw, pitch and roll angles of the head.

3. Experiments

The dataset we use is Biwi Kinect Head Pose Database [3], which contains color and depth images in pair from a Kinect sensor. The color image and the depth image are matched based on the camera calibration information. There are 24 sequences of 20 different persons in the dataset. For evaluation, the dataset is divided into a training set and a testing set of 18 and 6 sequences. We select the mean width and height of all the face bounding boxes of 7459 frames in the training dataset, and create new bounding boxes with the original center and new size. Sub-images of the color and depth images are cropped out as the training samples. The size of these sub-images is 98×102. RGBD-HOG feature of 10×10 cells are extracted from the training images and concatenated to be an RGBD-HOG feature vector of $10\times 10\times 9\times 2$=1800 dimensions for one sample. So both the number of weights in an RGBD-HOG filter and the number of neurons in the input layer of the neural network are 1800.

Our first experiment examines the performance using RGBD-HOG features, RGB-HOG features and Depth-HOG features. Coarse localization errors in one test sequence are shown in Figure 3. The RGBD-HOG performs best and Depth-HOG ranks second in this experiment. Table 1 shows the mean and standard deviation of the errors in the face pose estimation experiments on 640×480 color and depth images. Both false positive rate and false negative rate of the detection are rather low ($<1\%$). The more displacement we search (the larger value of d is chosen) in the refining step, the lower the errors would be. Since the scale is rather well after the scaling based on the depth information and small changes are enough to obtain fine estimation results, δs is fixed among three values {-0.05, 0, 0.05}. The pose estimation method runs at about $10 fps$ with d = 4 on a 2.93GHz 6-core Intel Xeon CPU. Figure 4 shows some examples of the face pose estimation results.

4. Conclusion

We have proposed an RGBD-HOG feature based approach for estimating face location and orientation with color image and depth data from a Kinect sensor. The combination of color and depth information achieved more robust estimation than using color and depth information separately. A coarse-to-fine face localization method using multiple HOG filters is adopted to locate

Table 1. Mean and standard deviation of the errors in face pose estimation experiments

	Localization error (pixel)	Localization error (mm)	Yaw error (°)	Pitch error (°)	Roll error (°)
Coarse	5.84 / 3.16	8.85 / 5.24	10.63 / 9.83	12.41 / 9.09	10.43 / 7.66
Fine ($d = 1$)	5.19 / 2.86	7.72 / 4.42	10.19 / 9.84	11.16 / 8.42	9.37 / 6.74
Fine ($d = 2$)	4.62 / 2.39	6.83 / 4.33	9.42 / 8.70	10.23 / 8.14	8.23 / 5.92
Fine ($d = 4$)	3.97 / 2.18	6.24 / 3.71	8.92 / 8.27	9.12 / 7.40	7.42 / 4.90

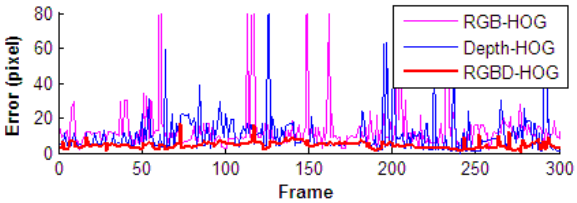

Figure 3. Localization errors with different features. High values indicate false detections.

the face. A feed-forward MLP network is applied for fine and continuous face orientation estimation. Experiments show that our approach achieves fast face pose estimation with low errors.

5. Acknowledgment

This work was partially supported by the Natural Science Foundation of China (60675021), Beijing Natural Science Foundation (4102052), and Beijing Key Lab of Digital Media & Human-Computer Interaction.

References

[1] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893, 2005.

[2] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. pages 428–441. Springer, 2006.

[3] G. Fanelli, T. Weise, J. Gall, and L. V. Gool. Real time head pose estimation from consumer depth cameras. In *Proc. 33rd Annual Symposium of the German Association for Pattern Recognition*, 2011.

[4] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[5] J. Huang, X. Shao, and H. Wechsler. Face pose discrimination using support vector machines (svm). In *Proc. 14th International Conference on Pattern Recognition*, volume 1, pages 154–156, 1998.

[6] M. Jones and P. Viola. Fast multi-view face detection. *Mitsubishi Electric Research Lab TR-2003-96*, 2003.

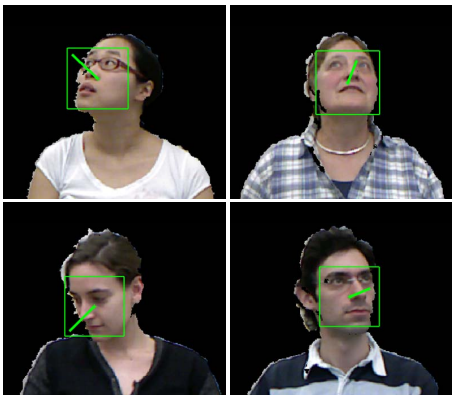

Figure 4. Examples of estimation results.

[7] L. P. Morency, J. Whitehill, and J. Movellan. Monocular head pose estimation using generalized adaptive view-based appearance model. *Image and Vision Computing*, 28(5):754–761, 2010.

[8] E. Murphy-Chutorian and M. M. Trivedi. Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):607–626, 2009.

[9] R. Rae and H. Ritter. Recognition of human head orientation based on artificial neural networks. *IEEE Transactions on Neural Networks*, 9(2):257–265, 1998.

[10] E. Seemann, K. Nickel, and R. Stiefelhagen. Head pose estimation using stereo vision for human-robot interaction. In *Proc. 6th IEEE International Conference on Automatic Face and Gesture Recognition, 2004*, pages 626–631, 2004.

[11] J. Sung, T. Kanade, and D. Kim. Pose robust face tracking by combining active appearance models and cylinder head models. *International Journal of Computer Vision*, 80(2):260–274, 2008.

[12] M. Voit, K. Nickel, and R. Stiefelhagen. Neural network-based head pose estimation and multi-view fusion. *Multimodal Technologies for Perception of Humans*, pages 291–298, 2007.

[13] J. Xiao, S. Baker, I. Matthews, and T. Kanade. Real-time combined 2d+3d active appearance models. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, 2004.

[14] R. Yang and Z. Zhang. Model-based head pose tracking with stereovision. In *Proc. 5th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 255–260, 2002.