FACE DISTANCE ESTIMATION FROM A MONOCULAR CAMERA

Shashi Kumar M.S, Vimala.K.S. Avinash.N,
Jain University, Bangalore, India
Wittybot Technologies, Bangalore, India
{ shashikumar,vimala,avinash}@wittybot.com }

Abstract – A new era of mobile devices (phones and tablets) has emerged with some of them possessing a front camera facing the person or user and stereo cameras in the rear side. This paper emphasise on a novel method to use such devices with monocular camera to determine the depth between the user and the front camera using a back propagation neural network (BPNN). This depth is successively used to calculate the zooming factor for a legible view, to read a document on the display of the mobile device. In the absence of camera parameters and also considering the fact that the camera and user's face are in constant motion; measuring the distance of the face from the camera becomes a hard problem. We propose the use of frontal facial features acquired from the monocular camera to find the depth information with the use of supervised learning algorithm. Training the BPNN for the facial features for a standard face is one time factory programmed during camera manufacturing. One time new user's face registration for the gadget using rear cameras followed by the simulation of depth from the front camera using the trained BPNN is proposed in this work. The results of distance estimation are likened with stereo camera setup and the approach is validated.

Keywords:
Facial features, Monocular front camera, Stereo rear cameras, Zoom factor, Back propagation neural network.

1. INTRODUCTION

With the advancement in camera technology, some of the mobile phones and tablet PCs are embedded with a monocular front camera facing the user and stereo cameras on the rear (eg. HTC Evo 3D, LG Optimus 3D P920, etc.). User positions himself to the device at a particular distance to read the contents of the display. The work presented in this paper is an approach to use the front facing camera and recover the depth of the user from the monocular camera, which in turn is used to adjust the contents of the display based on the depth. The idea is to give to the user, clarity in reading the contents based on the distance of his eye positioning (as depicted in figure 1).

There are many approaches existing to estimate the depth of an object from camera. Depth from stereo images [1] is the commonly used method to estimate the depth. It generates depth map based on triangulation of correspondence points in stereo images. By using stereo set up the depth information can be obtained with good accuracy. But, stereo setup is computationally expensive than that of monocular setup, because of the fact that it is required to search for correspondence [2,3] points in stereo images.

Structure from motion approach uses the sequence of images from monocular camera to perceive depth information [4]. It needs the relative motion between the object and camera, thus making the correspondence matching throughout the sequence a tough task.

Saxena et al [5] use supervised learning to estimate depth using Markov Random Field (MRF). Beyang Liu et al[6], Ashutosh Saxena et al [7], used similar appearance based approach using semantic segmentation. All these methods are suitable for outdoor environment where segmentation is a difficult task.

We approach the problem on similar grounds, i.e. to solve depth from monocular camera using supervised learning algorithm with the help of an artificial neural network. Most likely than always, the user of the mobile device faces the front camera and hence we consider the use of facial features for the depth estimation as the dominant cues.

In the literature, we find many approaches explained by researchers to detect the face [8] and facial features [9]. We need both pixel distance and 3d distance between these facial components in our approach to extract frontal facial features.

In this paper we propose a novel approach to find depth from monocular camera to a person's face, by supervised training of facial features along with the depth information with the help of a back propagation neural network (BPNN). Using this depth information we suggest a zooming factor for the display device for a legible view of the document.

The rest of the paper is organized as follows: in section 2 we present the overview of the important system. Section 3 explains proposed methodology to solve depth from monocular camera. In section 4 the experimental results and analysis is provided by validating with stereo system and suggesting with scope for improvement, followed by conclusion.

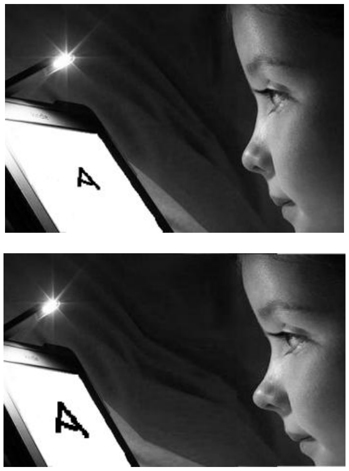


Figure 1: Display contents are zoomed according to the face distance from the device.

978-1-4799-2341-0/13/$31.00 ©2013 IEEE

2. SYSTEM OVERVIEW

The overall system (as in figure 2) is divided into two modules: viz., Training module and Simulation module.

Training module is a one-time process for a particular camera performed in the factory during the time of manufacturing the camera. This consists of stereo setup to estimate depth between standard face model and camera (Section 2.2) for neural network training. The appearance of facial features (Section 2.1) of standard face model at different depths is used to train BPNN (Section 2.4).

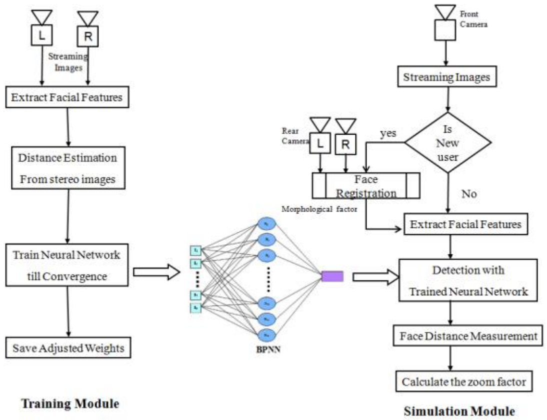


Figure 2: Training and Simulation modules are two main modules of the Proposed System

Simulation module estimates the depth value for the given facial image from monocular camera. It demands the user to register his/her face (Section 2.3) before using document reading application. The ratio of facial features between standard face model and registered face is computed and called as *morphological factor* (Section 2.3) of the face features. The face registration calculates the morphological factor to fit the user face with standard face model. Once the registration is completed the BPNN along with the trained weights and morphological factor, will generate depth value using monocular image stream from the front camera.

The resultant depth information is used to calculate the zoom factor of the document for legible view.

2.1 Facial Feature Extraction

In this paper we use prominent features of face – face boundary, eyes, and nose. Face boundary, eyes and nose are key features of the face as these features are spatially far apart over the entire face region, thus does not requiring any complex mathematical calculation.

We calculate the distance between eyes, and eye-nose tip, also face height and face width all in image co-ordinate system (ICS) in pixel units. This is used both during training and simulation module as the basic features. The distance between same facial features are also calculated in camera co-ordinate system (CCS) in meter units to find morphological factor while face registration. We have used haar features based feature extraction system to extract face, eyes and nose region from the given image. The detected face boundary is fitted with an ellipse [12] to find height and width of the face, corresponding to the major and minor axes. The identified features in our approach are as shown in figure 3.

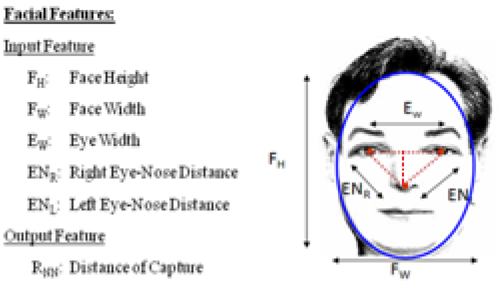


Figure 3: Facial Features

Figure 4, illustrates that with the increase in the distance between the camera and the object the size of the object becomes smaller but are linear perspective in nature[14]. Also illustrated by Filippo Brunelleschi in 1413.The facial features chosen in this paper also change linearly with distance.

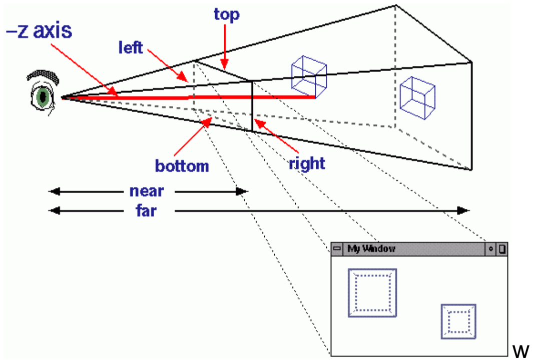


Figure 4: Linear perspective principle example

2.2 Depth Measurement using Stereo setup

Accurate measurement of dimension is important in real-time application. The proposed setup requires measurement of depth between standard face model and camera. Depth is measured using stereo triangulation technique from stereo image pair [1]. Distances between facial features in CCS are required to calculate morphological factors. To find this, a stereo setup is used during training of BPNN. The stereo system calculates depth from the camera using eyes as correspondence points in stereo images.

2.3 Face Registration

Face Registration is a onetime process for a particular user on the mobile device, unless otherwise the user changes. It is used to calculate the morphological factor of the user's face with that of the standard trained face. This gives the factor by which facial feature need to shrink or stretch to fit the user's face with that of a standard face model.

The Morphological factor can be calculated as follows.

Step1: Read left and right images from stereo setup and locate nose tip and both eyes in them.
Step2: Extract facial features from both cameras of the stereo setup.
Step3: Find the distances in CCS of a new user for all the facial features (F_{Ri}) as mentioned in section 2.2.
Step4: Fetch all the facial features of the standard face (F_{si}) measured in CCS which were computed during the training of standard face for BPNN.
Step 5: Find the morphological factor (M_i)for each of the facial features

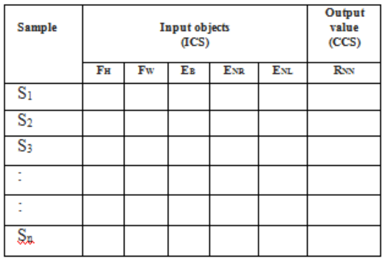
$$M_i = \frac{F_{Si}}{F_{Ri}} \tag{1}$$

where,
F_{Si} is facial feature of standard face
F_{Ri} is facial feature of registering face

These Morphological factors are used to calculate morphed facial features to estimate depth from monocular images.

2.4 Back Propagation Neural Network

Table 1: Sample Training Constraint



Sample	Input objects (ICS)					Output value (CCS)
	F_H	F_W	E_B	E_{NR}	E_{NL}	R_{NN}
S_1						
S_2						
S_3						
:						
:						
S_n						

Neural networks [13] have proven to be effective mapping tools for a wide variety of problems. We use supervised learning to train the BPNN to estimate the depth between user and front camera, based on the facial features (either standard facial features or user's morphed facial features). In BPNN, the error is back-propagated to adjust the weights so that the error is reduced between actual and estimated depth. Here facial features are the input objects, while depth is the estimated value. Here we used a neural network with 5 neurons in the input layer, 1 neuron in output layer, and 2 hidden layers with 38 neurons altogether. The sample training set is as shown in table 1. Samples are collected as explained in section 3.1.

3. PROPOSED APPROACH

The proposed approach can be divided into two modules of development.

- Training module
- Simulation module

3.1 Training Module

In training module we place a standard face model at random position from stereo setup to cover the field of view. At all possible positions in the field of view, the depth between the camera and face model is found using stereo triangulation w.r.t. CCS for the correspondence points, (i.e. eyes and nose tip).

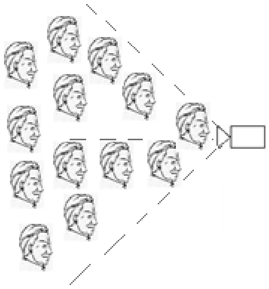


Figure 5: Training of standard face model

Using the image from one of the cameras of the stereo setup, facial features are detected in ICS too. These distances of the facial features in ICS and corresponding depth values in CCS are trained using a BPNN as designed in section 2.4. The algorithm for the training module of the proposed method is given below.

Step1: Read left and right images from stereo setup.
Step2: Extract the coordinates of both eyes and nose tip from the left and right images of stereo setup.
Step3: Using stereo triangulation technique, depth between face and the camera is determined from correspondence points in the image pair.
Step4: Use the left image of the stereo setup; extract facial features in ICS as explained in section 2.1.
Step5: Repeat Step1 to Step4 till enough data is populated in table 1, to cover the entire field of view of the camera from various distances as shown in the figure 4.
Step6: Train BPNN with the populated data until convergence with facial features as input objects, and corresponding depth between user and the camera as the output value.
Step7: Save weights after convergence.

3.2 Simulation Module

Simulation module estimates the depth between user and display using monocular image stream. This module is executed on the smart mobile device used along with the document viewer which can automatically zoom its contents based on the face distance of the user. However considering the fact that the facial features of all individuals are different, the user of the smart mobile device requires to register the face for the first time of its usage (as explained in section 2.3). In face registration, the user's face is registered as against the sample face used for training with the help of a morphological factor for normalizing the facial features. This is depicted in the simulation module of the figure 2.

Algorithm for estimating the depth from monocular camera from BPNN is given below.
Step1: Read an image from image stream of monocular camera.

Step2: Extract distance of facial features in image co-ordinate (F_i)

Step3: Normalize all facial features to fit the standard face model as given below.

$$F_{Ni} = F_i \times M_i \tag{2}$$

Where,

Mi is the morphological factor for the corresponding facial feature.

Step4: Simulate the output depth using morphed facial feature and the trained weights as input to the BPNN

Step5: Find the zoom factor (Z) as a function of face depth from the camera for the particular mobile device.

$$Z = f(estimated\ depth) \tag{3}$$

The calculated zoom factor is updated for a legible view of the document reader.

4. EXPERIMENTAL RESULT AND ANALYSIS

The accuracy of proposed system is calculated for the depth values obtained from BPNN algorithm as against stereo setup results. Experimentally it is verified with a system as shown in figure 5. This system consists of a stereo setup of similar cameras (i.e. same intrinsic parameters). Extract facial features from stereo images (as explained in section 2.1) in ICS. Using left camera, depth value is estimated with BPNN. Using both cameras, distance of the face is measured using stereo triangulation. Considering the stereo measurement as the baseline, the percentage accuracy of proposed depth estimation system is calculated using the following formula.

$$\%Accuracy = \left\{1 - \left(\frac{|R_{NN} - R_{ST}|}{R_{ST}}\right)\right\} \times 100 \tag{4}$$

Where

R_{NN} – Depth measurement using neural network

R_{ST} – Depth measurement using stereo setup

During training of facial features of standard face model, more than 500 input images were captured to train BPNN so that all possible view and distance is covered as shown in figure 4.

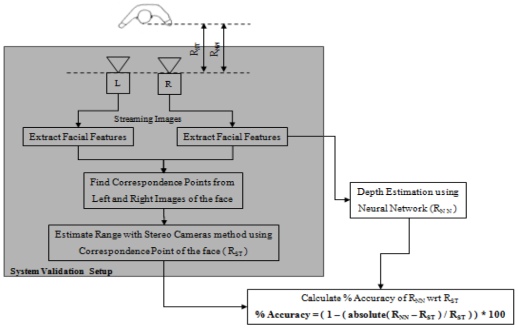

R_{ST}
R_{NN}
L
R
Streaming Images
Extract Facial Features
Extract Facial Features
Find Correspondence Points from Left and Right Images of the face
Estimate Range with Stereo Cameras method using Correspondence Point of the face (R_{ST})
System Validation Setup
Depth Estimation using Neural Network (R_{NN})
Calculate % Accuracy of R_{NN} wrt R_{ST}
% Accuracy = (1 – (absolute(R_{NN} – R_{ST}) / R_{ST})) * 100

Figure 5: Experimental setup for validation model

The images were captured with a minimum distance to cover the full face of the person in the field of view of the camera and a maximum distance of around 1.5 meter from the camera for populating data.

The speed of the proposed validation system was nearly 12 fps on a dual core 2 GHz, 1 GB RAM. Both stereo camera approach and monocular approach of face distance measurement algorithms were run as a single system. This validation system was tested for 30 different users with about 2000 frames per user.

Average accuracy for each person for all frames was computed. The least of the average accuracy in depth estimation using BPNN algorithm was 88% upon face registration and was 72% without face registration (considering morphological factor=1). The maximum of the average accuracy turned out to be 95% and 79% respectively.

The graphs in figure 6 and figure 7 shows distance measurements of two different users. The solid line is the measurement from stereo setup, dashed line is the measurement from BPNN with face registration, and dotted line is the measurement from BPNN without face registration. The average accuracy of user-1 improved to 94.4% with face registration as against 77.43% without face registration. The average accuracy of user-2 improved to 92.6% with face registration as against 72.88% without face registration.

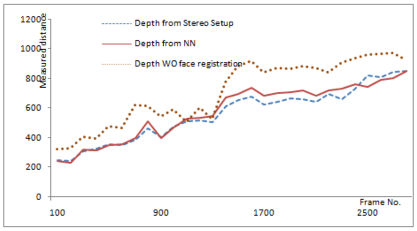


Figure 6: depth comparison for user1

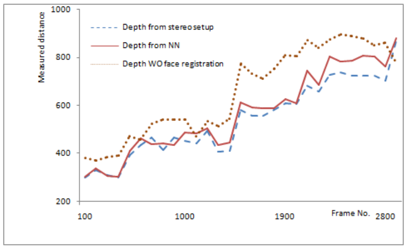


Figure 7: depth comparison for user2

5. CONCLUSION

This paper explains a methodology of estimating the depth of a person's face from a monocular camera using a supervised learning algorithm. We make use of the back propagation neural network for the same. Stereo triangulation method is used to make a baseline for the training of the neural net as well as the validation of the approach. Further it is said that a mobile device on which this system works integrates with a document viewer which zooms the document size as a function of the distance of the reader of the device. The validation results shows the accuracy in our experimentation was above 88% for most of the faces well registered. We understand the process of registering the face is a manual intervention in an automatic system. We consider this a scope of improvement of the system.

6. REFERENCES

[1] R. Hartley and A. Zisserman. Multiple View Geometry in Computer Vision, Chapter 9-12.

[2] S.B Poland, J.E.W. Mayhew, and JPFrisby.PMF:A stereo correspondence algorithm using a disparity gradient constraint. Perception, 14:449-470, 1985.

[3] D. Scharstein and R. Szeliski, “A taxonomy and evaluation of dence two-frame stereo correspondence algorithms,” International Journal of Computer Vision, vol. 47, no. 1/2/3, pp. 7–42, April – June 2002.

[4] Richard Szeliski. A Multi-View Approach to Motion and Stereo, Technical Report MSR-TR-99-19, 1999

[5] A.Saxena, S.H.Chung, and A.Y.Ng. Learning depth from single monocular images. In Proc. NIPS, 2005

[6] Beyang Liu, Stephen Gould, and Daphne Koller. Single image depth estimation from predicted semantic labels.CVPRIEEE ,pp. 1253-126, 2010.

[7] Ashutosh Saxena, Jamie Schulte and Andrew Y.Ng. Stanford University. Depth Estimation using Monocular and Stereo Cues.

[8] Frank Y. Shih, Chao-Fa Chuang. Automatic extraction of head and face boundaries and facial features Information Sciences. Elsevier, pp117–130, 2004.

[9] P.Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In Proc. of IEEE Conference on Computer Vision and Pattern Recognition, Kauai, HI, December 2001.

[10] Paola Campadelli, Raffaella Lanzarotti and Giuseppe Lipori Università degli Studi di Milano Italy. Automatic Facial Feature Extraction for Face Recognition.

[11] Philip Ian Wilson, John Fernandez, Facial feature detection using haar classifiers. Journal of Computing Sciences in Colleges, 21:127–1.

[12] Radim Hal, Jan Flusser. Numerically stable direct least squares fitting of ellipse. Wilson, P. I. and Fernandez, J. (2006).

[13] Jack M.Zuruda. Introduction to Artificial Neural System, chapter 4, pages 165-214.

[14] Raynaud, Dominique. Understanding errors in perspective. R.Boudon, M.Cherkaoui, P. Demeulenaere, eds, The European Tradition in Qualitative Research, chap. 13 1 (2003): 147-165.