FacialStereo: Facial Depth Estimation from a Stereo Pair

Gagan Kanojia and Shanmuganathan Raman

Electrical Engineering, Indian Institute of Technology Gandhinagar, Ahmedabad, India

Keywords: Sparse Stereo, Active Shape Model, Face Detection.

Abstract: Consider the problem of sparse depth estimation from a given stereo image pair. This classic computer vision problem has been addressed by various algorithms over the past three decades. The traditional solution is to match the feature points in two images to estimate the disparity and therefore the depth. In this work, we consider a special case of scenes which have people with their front-on faces visible to the camera and we want to estimate how far a person is from the camera. This paper proposes a novel method to identify the depth of faces and even the depth of a single facial feature (eyebrows, eyes, nose, and lips) of a person from the camera using a stereo pair. The proposed technique employs active shape models (ASM) and face detection. ASM is a model-based technique consisting of a shape model which contains the data regarding the valid shapes of a face and a profile model which contains the texture of the face to localize the facial features in the stereo pair. We shall demonstrate how depth of faces can be obtained by the estimation of disparities from the landmark points.

1 INTRODUCTION

Human visual system can easily perceive the three dimensional information of the world. We can identify the shapes of different objects and their relative distances with ease. While capturing an image, we project the 3D visual data into the 2D space. During this process, we lose valuable information regarding the distance of an object from the camera. Although by looking at an image, one can identify which object is nearer and which one is farther, but the computers can not do so and estimate their actual positions in the 3D world.

This paper describes a novel technique to estimate the position of a person and his facial features (eyebrows, eyes, nose, and lips) in the 3D world. For this purpose, the concepts of stereo matching and disparity are used. The variability of shapes of faces and facial features leads to the need of a flexible model which allows some degree of variability. It should also be able to deal with the varying complexion through different faces. This need motivated the use of active shape models (ASM) to process the facial images in a scene as they modify themselves according to the structure of face and facial features irrespective of the complexion of the skin (Cootes et al., 1995).

The commercially available cameras available today have a built-in face detection module to achieve proper focusing of the salient people in the scene. The present work targets the utilization of this module to also report the depth of the persons in the scene. Though we assume in the present work that the epipoles of the stereo pair captured are at infinity, we can use this approach even otherwise after rectifying the stereo pair (Hartley and Zisserman, 2004).

The most significant contributions of this paper are listed below.

1. Stereo image pair is used to estimate sparse depth of the faces in images. Concept of stereo disparity is used to estimate depth.
2. ASM is employed to obtain the contours of face and facial features using user specified landmark points.
3. The proposed approach does not require detection of feature points using techniques such as scale invariant feature transform (SIFT) and corner detectors (Tuytelaars and Mikolajczyk, 2008).
4. The proposed approach is fully automatic and can be built into a stereo imaging system for detecting depth of the people in a given scene.

The rest of the paper is organized as below. The section 2 describes the previous works performed related to the proposed approach. We shall discuss the necessary background regarding the ASM for con-

Kanojia G. and Raman S..
FacialStereo: Facial Depth Estimation from a Stereo Pair.
DOI: 10.5220/0004826006860691
In *Proceedings of the 9th International Conference on Computer Vision Theory and Applications* (VISAPP-2014), pages 686-691
ISBN: 978-989-758-009-3

verging on the facial contours in section 3. The proposed approach for determining depth of the people present in the scene is explained in section 4. We shall demonstrate the usefulness of the proposed approach for different scenes in section 5. We shall conclude this paper by giving future directions in section 6.

2 PREVIOUS WORK

Marr and Poggio were the first to propose an approach to perform stereo matching between two views of the same scene (Marr and Poggio, 1971). They showed how one can recover the depth of objects in a scene by estimating stereo disparity (Barnard and Fischler, 1982). The disparity can also be estimated by using a window of adaptive size and thereby establishing correspondence (Kanade and Okutomi, 1994). Stereo vision algorithms establish correspondence between two views of the scene using epipolar constraints (Hartley and Zisserman, 2004). A lot of algorithms use this idea to estimate the depth of objects in a scene (Scharstein and Szeliski, 2002). Recently, Chakraborty *et al.* developed a technique to classify people interactions known as proxemics (Chakraborty et al., 2013). They categorized the scene on the basis of the distance between the people in the images.

Cootes *et al.* proposed a model-based technique called active shape models (ASM) to deal with the variability of the patterns (Cootes et al., 1995). To achieve this challenging task, they built a model from a training set of annotated images through learning. This model is flexible enough to deal with the probable variation within a class of images. ASM approach was further enhanced in the active appearance models (AAM) (Cootes et al., 2001). ASM and AAM enable one to model the contours of the various features of a face image.

Viola and Jones proposed an approach for a rapid object detection (Viola and Jones, 2004). They developed a machine learning approach which involves efficient classifier by combining the power of a number of weak classifiers. Face recognition is one of the challenging tasks in computer vision research (Zhao et al., 2003). The stereo vision has primarily been used for face and gesture recognition tasks in recent years (Matsumoto and Zelinsky, 2000). A dense depth recovery system from stereo images is proposed by Hoff and Ahuja (Hoff and Ahuja, 1989).

In this work, we concentrate on the sparse recovery of depth in few selected feature points as our objective is to estimate the distance of a person or the feature from the camera. We shall first discuss the application of ASM to estimate the facial contour from the images containing human faces.

3 FACIAL CONTOURS USING ACTIVE SHAPE MODEL

This section provides the 2D formulation of ASM. It is comprised of two models i.e. shape model and profile model.

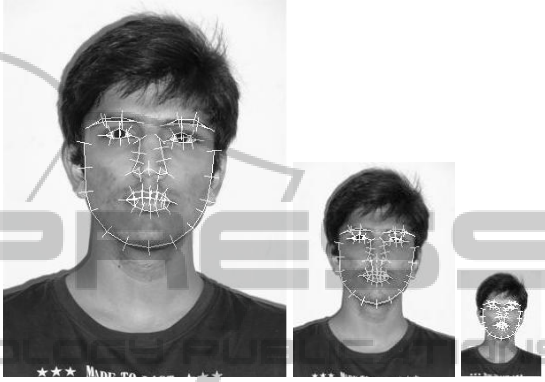

Figure 1: Shape and normals along which gray values are extracted for three different resolutions i.e. 360 × 480, 180 × 240 and 90 × 120 (in pixels) of a facial image.

3.1 Shape Model

A shape is a set of $n \times 2$ ordered points where n is the number of landmark points which signify different locations marked in a face contour. Even after operations like scaling, rotation and translation on a shape, it retains the original shape. For this purpose, the shape is scaled such that $\| x \| = 1$, so that the size of the face does not affect the process. A shape is considered as a $2n$ dimensional vector x

$$x = (x_1, y_1, x_2, y_2, \ldots.., x_n, y_n), \tag{1}$$

where x_n and y_n are the coordinates of the landmark points.

A training set is taken with different shapes corresponding to different faces. To start with, they are aligned by scaling, rotating and translating the shape using a similarity transformation.

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} s\cos\theta & s\sin\theta \\ -s\sin\theta & s\cos\theta \end{bmatrix} + \begin{bmatrix} x_{tr} \\ y_{tr} \end{bmatrix} \tag{2}$$

where, x_t and y_t are the transformed x and y coordinates, s is the scaling factor, θ is the angle of rotation and x_{tr} and y_{tr} are the translation factors.

Then the mean shape

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{3}$$

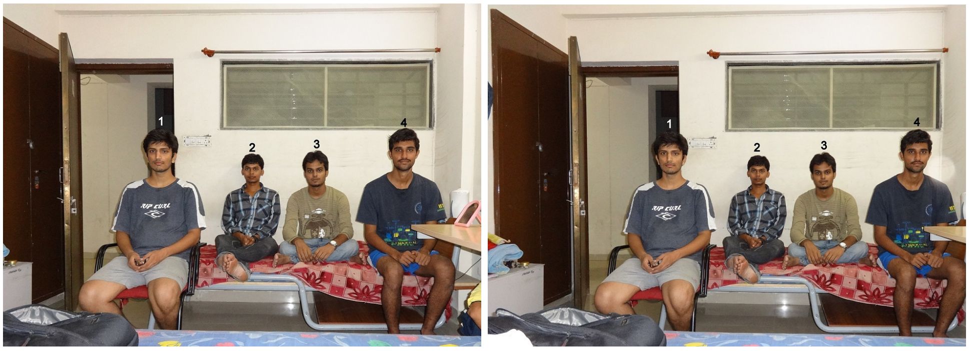

Figure 2: Stereo image pair of a scene which has multiple faces at different depths. The images are of size 4608 × 3456 pixels.

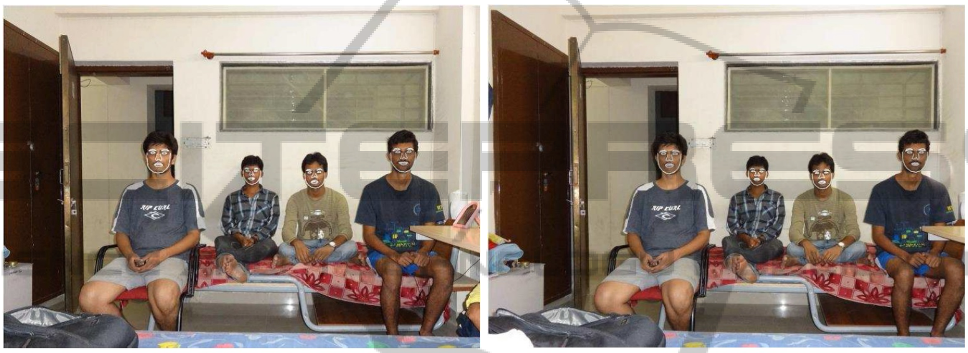

Figure 3: Stereo image pair after the proposed algorithm is applied. On both the images, the obtained active shape contours and the landmark points are displayed.

and covariance is computed

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T \qquad (4)$$

where, n is the number of shapes in the training set.

The shape can be approximated as

$$\hat{x} = \bar{x} + b\Phi \qquad (5)$$

where $\bar{x}$ is the mean shape, Φ is the matrix of eigenvectors of covariance matrix and b is a vector. The value of b is constrained to be between $\pm m\sqrt{\lambda}$, where, m is either 2 or 3 and λ is a vector having eigenvalues as its elements, to generate a face-like structure.

Principal component analysis (PCA) is applied on the ordered eigenvalues and the corresponding eigenvectors so that only significant eigenvectors remain and also for the removal of noise components.

3.2 Profile Model

This model describes the one dimensional pixel profile around a landmark point. Its job is to give the best approximate shape according to the given image when a suggested shape by the shape model is given.

For the purpose, gray scale pixel values are used as the profile data. In this, we sample the image at each landmark point along the normal to the contour and extract k values on both sides of the landmark point as shown in Fig.1. This way we get a profile of $2k+1$ values. Then, for the profile model, mean profile $\bar{g}$ and the covariance matrix S_g is computed for each landmark point across all the images in the training set.

4 SPARSE FACIAL DEPTH ESTIMATION

Let us consider a stereo image pair of a scene which contains facial images as shown in Fig.2. We shall explain why stereo image pair is taken and how depth can be calculated from them soon. On the given images, Viola-Jones face detector is applied to detect faces within the scene (Viola and Jones, 2004). This detector detects all the face-like structures present in the image. They may or may not correspond to an actual face. To increase the probability of getting an actually face, an eye detector is applied on the face-like structure detected by the face detector. If it detects eyes in the detected face-like region, then it is considered as a face else it will be discarded.

On the detected face region, the mean shape sug-

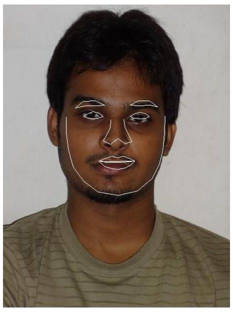

Figure 4: Estimated mean shape from the training set is placed on the face detected using Viola-Jones face detection algorithm. The image size is 3456 × 4608 pixels.

gested by the shape model is placed (i.e. the mean shape calculated from the training shapes) after scaling it up in accordance with the coordinates (approximate width of the eye region) provided by the eye detector (since the shapes are scaled such that $\| x \| = 1$). After placing the shape on the image, the output image will look like the one shown in Fig.4. Then the image is sampled at each landmark point along the normals to the contour and p values are extracted on both sides of the landmark point such that $p > k$. In this case, gray scale pixel values are extracted. This way a search profile gets created for each landmark point. Along this search profile, a profile of k values is found which matches best the model profile. For this, Mahalanobis distance is computed by moving the model profile $\bar{g}$ along the search profile.

$$r = (g - \bar{g})^T S_g^{-1} (g - \bar{g}) \tag{6}$$

where, r is the Mahalanobis distance. The landmark points get moved along the normal to their new location corresponding to the minimum Mahalanobis distance. Then the constraints are applied on the shape which is obtained after moving each landmark point independently to their new location in accordance to the equation (5) to get a face-like shape. By doing this iteratively, a shape surrounding the contours of all face features is obtained. To get better results, multi-resolution approach is used in which the above algorithm is applied on different resolutions of the same image moving from coarse to fine level. In this the result obtained in coarse level is taken as reference for the next level.

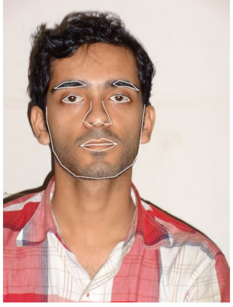

Figure 5: Contours obtained on a facial image.

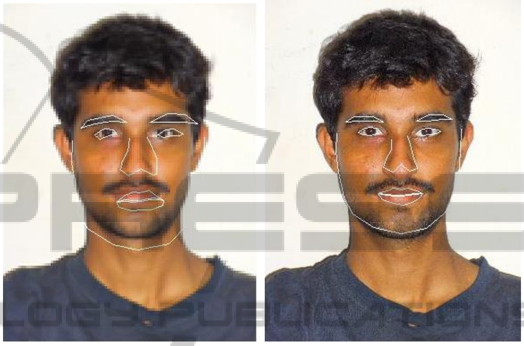

Figure 6: Contours obtained on the same images of different spatial resolutions. The left image is of poor quality with resolution 90 × 120 pixels while the right image has better quality with resolution 360 × 480 pixels.

After obtaining the contours of all face features successfully in each of the stereo images, the mean positions of face and facial features of each face in the given images are calculated. This is done by taking the mean of the coordinates of the points encircling them. After getting the mean positions, disparity among the corresponding faces and facial features in the stereo image pair is computed. Disparity can be calculated by computing the Euclidean distance.

$$d = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{7}$$

where (x_i, y_i) and (x_j, y_j) are the coordinates of a single point in the two stereo images. Then, depth is computed using the following relation.

$$Z = f\frac{B}{d} \tag{8}$$

where, Z is the depth, f is the focal length of the imaging system in pixels, B is the baseline between the stereo cameras, and d is the disparity in pixels between a pair of landmark points. We shall assume that we know the focal length and the baseline distance for a given stereo pair.

5 RESULTS AND DISCUSSION

The 1-D search along the normals make sure that the obtained contours in the both the images (stereo pair) are same. As the initial shape i.e. the mean shape, is same in both the cases, so for the frontal images of the same person the contour obtained is same. So, the difference in the coordinates of the landmark points is

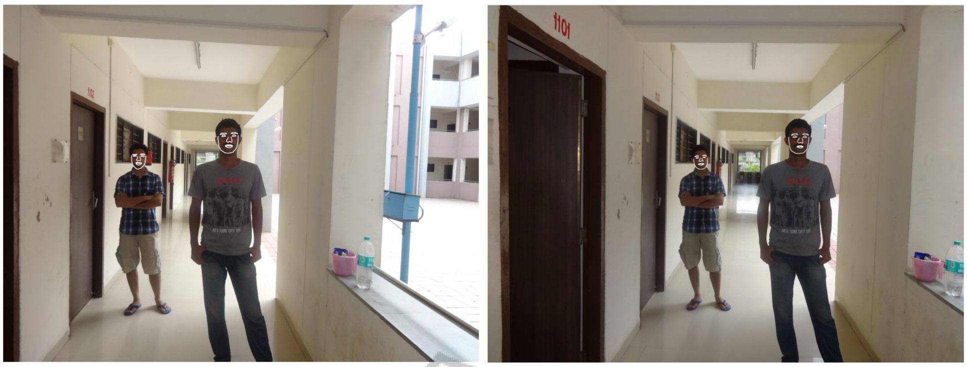

Figure 7: The obtained active shape contours and the landmark points on the stereo image pair of size 4608 × 3456 each.

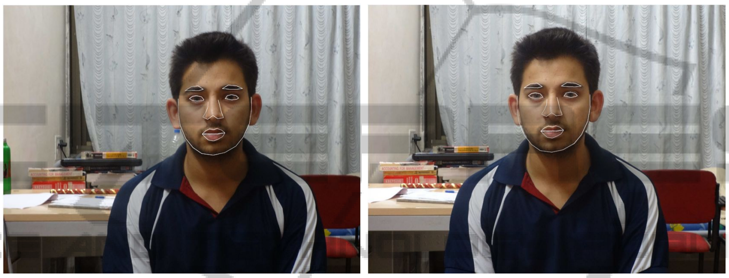

Figure 8: The obtained active shape contours and the landmark points on the stereo image pair of size 4608 × 3456 each.

mainly due to inter-camera distance. The difference due to the change in the shape of contours obtained can be minimized by keeping the inter-camera distance low (approx. 5cm). If the faces are quite close to the camera as shown in Fig. 8 (or faces are zoomed in) then to get better results the face should be at the axis which will perpendicularly bisect the line joining the two cameras but if the faces are at an appreciable distance as shown in Fig. 2 then there is no need of it.

The proposed approach is applied on the stereo image pair shown in Fig. 2. After the algorithm is applied, image pair shown in Fig. 3 is obtained. From the stereo pair in Fig. 3, it can be observed that all the four faces are detected and the contours of facial features are successfully obtained in all the faces. From the Fig. 3, we can easily observe that the proposed algorithm is independent of facial color complexion. It also works fine on the faces with beard and moustache.

The training set contains 28 images that has been manually landmarked with 76 points. As by just looking at the Fig. 2 , it can be perceived that the face marked as 1 is closest to camera and the face marked as 2 is farthest. The images shown in Fig. 2 are of size 4608 × 3456 pixels and the associated focal length and baseline are 4.5mm (2668.7 pixels) and 10cm respectively. The disparity obtained for each face are 232, 177.2, 191.8 and 221.2 (in pixels) respective to the numbering. The depth of each face computed by the algorithm are 115.03, 150.64, 139.12 and 120.65 (in cm) respective to the numbering. The algorithm was implemented using MATLAB R2103a on a laptop with i5 processor and 4GB RAM. The camera used for the experiment is of 16.2 megapixels with 8x optical zoom and 7.77 mm sensor size. The runtime of the algorithm decreases with the decrease in size of the images and number of faces present in the image. Fig.7 and Fig. 8 are another examples of successful application of the proposed algorithm.

The images shown in Fig. 8 are of size 4608 × 3456 pixels and the associated focal length and baseline are 13mm (7709.65 pixels) and 5cm respectively. The depth of face, eyebrows (left and right), eyes (left and right), tip of nose and lips computed by the algorithm are 113cm, 113.29cm, 114cm, 112cm and 112.83cm respectively. From the obtained results it can be seen that even the depth of the single feature can give a close estimate of distance of face from the camera.

The results obtained in Fig.2 and Fig.3 clearly states that the proposed algorithm can work on images with any number of the facial images present in the scene. The successful procurement of the depth of facial images also depends on the successful detection of the facial images. Any occluded face will not be detected by the face detector and hence their depths can not be estimated.

On poor quality images, ASM does not work effectively. This is shown in Fig.6. It also goes for the facial images which are at the large distance in the image as the pixel information will not be sufficient because of its small size for the model to work upon. Hence, contours cannot be obtained successfully in such images. Therefore, in such cases depth cannot be estimated by this technique.

The relation given in equation 8 does not apply for objects at large distances. This explains why the depth calculated for the facial image in Fig.9 was incorrect.

6 CONCLUSIONS AND FUTURE WORK

We have developed a novel method to recover the sparse depth information of the persons whose faces are present in a given scene. The approach relies on the ASM features learnt for a given face and therefore does not require explicit computation of the feature detection for extracting the feature points. The advantage with the proposed approach is that we can even calculate the depth of the individual facial features such as eyes and mouth when the images are captured with sufficient zoom.

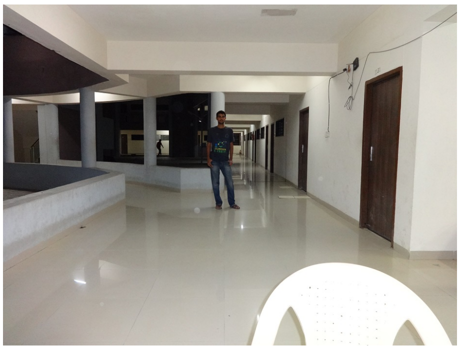

Figure 9: Image of a person standing at a large distance (around 5 metres) from the camera. The image size is 4608 × 3456 pixels. In such cases, the pixel information present in the facial region is not significant enough for the proposed algorithm and also for the Viola-Jones face detection algorithm.

The comparison of our results with state-of-the-art feature detection based sparse depth recovery techniques needs to be performed for validation. We plan to extend this approach to handle scenes which are captured using low resolution cameras and also persons who are located at much larger distance from the camera. These challenging situations can be addressed by using various low level image processing tools as a pre-processing step before using the proposed algorithm. As the stereo cameras have made their way into digital camera market, the proposed approach has the potential to provide information to the user about the proximity of a person from the camera.

REFERENCES

Barnard, S. T. and Fischler, M. A. (1982). Computational stereo. *ACM Computing Surveys (CSUR)*, 14(4):553–572.

Chakraborty, I., Cheng, H., and Javed, O. (2013). 3d visual proxemics: Recognizing human interactions in 3d from a single image. In *IEEE CVPR*, CVPR '13, pages 3406–3413.

Cootes, T. F., Edwards, G. J., and Taylor, C. J. (2001). Active appearance models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(6):681–685.

Cootes, T. F., Taylor, C. J., Cooper, D. H., and Graham, J. (1995). Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59.

Hartley, R. and Zisserman, A. (2004). *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2 edition.

Hoff, W. and Ahuja, N. (1989). Surfaces from stereo: Integrating feature matching, disparity estimation, and contour detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 11(2):121–136.

Kanade, T. and Okutomi, M. (1994). A stereo matching algorithm with an adaptive window: Theory and experiment. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 16(9):920–932.

Marr, D. and Poggio, T. (1971). Cooperative computation of stereo disparity. *Appl. Phys*, 42:3451.

Matsumoto, Y. and Zelinsky, A. (2000). An algorithm for real-time stereo vision implementation of head pose and gaze direction measurement. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 499–504. IEEE.

Scharstein, D. and Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1-3):7–42.

Tuytelaars, T. and Mikolajczyk, K. (2008). Local invariant feature detectors: a survey. *Foundations and Trends® in Computer Graphics and Vision*, 3(3):177–280.

Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *International journal of computer vision*, 57(2):137–154.

Zhao, W., Chellappa, R., Phillips, P. J., and Rosenfeld, A. (2003). Face recognition: A literature survey. *Acm Computing Surveys (CSUR)*, 35(4):399–458.