Automatic Detection of Emotion Valence on Faces Using Consumer Depth Cameras

Arman Savran
University of Pennsylvania
Department of Radiology
arman.savran@uphs.upenn.edu

Ruben Gur
University of Pennsylvania
Department of Psychiatry
gur@mail.med.upenn.edu

Ragini Verma
University of Pennsylvania
Department of Radiology
ragini.verma@uphs.upenn.edu

Abstract

Detection of positive and negative emotions can provide an insight into the person's level of satisfaction, social responsiveness and clues like the need for help. Therefore, automatic perception of affect valence is a key for novel human-computer interaction applications. However, robust recognition with conventional 2D cameras is still not possible in realistic conditions, in the presence of high illumination and pose variations. While the recent progress in 3D data expression recognition has alleviated some of these challenges, however, the high complexity and cost of these 3D systems renders them impractical. In this paper, we present the first practical 3D expression recognition using cheap consumer depth cameras. Despite the low fidelity facial depth data, we show that with appropriate preprocessing and feature extraction recognition is possible. Our method for emotion detection uses novel surface approximation and curvature estimation based descriptors on point cloud data, is robust to noise and computationally efficient. Experiments show that using only low fidelity 3D data of consumer cameras, we get 77.4% *accuracy in emotion valence detection. Fusing mean curvature features with luminance data, boosts the accuracy to* 89.4%.

1. Introduction

Machines that can perceive human affect (emotional expression) will have an important role in future human-computer interfaces. For instance, by detecting emotional valence, a computer system can have a clear strategy based on either positive, negative or neutral affect towards the system or an object. Natural social interaction between a virtual agent and human can be established by emotional perception ability. Another application is to augment human judgement with computerized automatic detection. As an example, clinical diagnosis of psychological disorders may be improved by computationally and objectively extracted affect information.

Facial expression is one of the most direct mediums that conveys affect information, making it the main motivation for the considerable research in automatic facial expression recognition, spanning over two decades. Although many techniques have been developed to recognize expressions, as surveyed in [3], it is still an active research area due to the challenging nature of the problem, especially because of out-of-plane head rotations and highly varying illumination conditions. Some researchers have considered the use of 3D facial data in order to alleviate these difficulties. 3D acquisition provides information about the 3D facial geometry directly. Depending on the 3D technology, illumination variation and shading can have negligible effect on 3D data compared to 2D color images. Moreover, out of plane rotations do not affect the data as in 2D where affect recognition can be hindered greatly depending on the pose.

However, due to the high cost, size and operating conditions of 3D systems used in previous expression recognition studies [15, 21], their application remains very limited. Recently, practical 3D acquisition became a possibility owing to the development of consumer level depth cameras. However the data is not of high quality and resolution, and it is unknown whether the low fidelity capture of faces that it provides, would help in recognition.

In this paper, we propose a novel method to enable fully automatic recognition of facial expressions using low quality 3D data composed of various processing steps including face detection, alignment and feature extraction. To the best of our knowledge, this is the first facial expression recognition study that uses low fidelity depth data of consumer level cameras, and hence the first practical 3D system. Our method benefits from fusion of 3D and luminance data for higher recognition accuracies in good illumination conditions, but also can work using only the depth channel which is crucial if the color/luminance channel is not usable due to high illumination variations.

2. Background

2.1. Affect Valence

In daily life, occurrence of pure prototypical expressions of the basic discrete emotion categories described by Ekman [1], like happiness, surprise and fear, is rather rare. First, categorical emotions frequently blend together, such as pleasant and unpleasant surprise, making discrete emotion classification inappropriate. Second, we exhibit complex affective states, such as embarrassment, affection, depression, boredom and confusion. These non-basic affective states can be expressed via a wide range of facial expressions, many of them can share similarities, and the differences between them can be quite subtle. For these reasons, Ekman's theory [1] of basic emotion categories remains limited in real-life expression recognition application scenarios. Dimensional affect theory, on the other hand, proposes a systematic continuous transition between various emotions [6]. Russell [11] showed that many emotion labels can be mapped to a circular configuration called as Circumflex Model of Affect. This circumflex model has affect valence in one axis and arousal on the other.

VALENCE characterizes if the emotion is positive or negative, i.e., unpleasant feelings vs. pleasant feelings. For example, while happiness, pleasure, contentment and affection are positive emotions, fear, anger, disgust and depression are negative emotions. VALENCE is the most commonly analysed affect dimension among the psychology researchers [11], and a comprehensive study has revealed that it is the most important affect dimension [5].

Nevertheless, so far discrete emotion classification has been the most common approach adopted for automatic affect recognition. Only recently, realization of higher potential of the dimensional affect theory in real world settings has attracted many researchers [17], and it has been observed that facial expressions are quite helpful for VALENCE [14].

2.2. Facial Expression Recognition with 3D Data

Previous 3D facial expression research has only considered classification of posed expressions of prototypical basic emotion classes, or recognition of facial action units (AUs). Most of these studies have been done on publicly available 3D expression databases (BU-3DFE [22] and Bosphorus [15] databases), and various methods using 3D data were surveyed in [13]. For instance, Wang et al. [21] divide 3D wire-frame faces into seven regions by means of 64 manual landmarks. Then, curvature related rule-based labels were assigned to every vertex, and histograms of the labels over these regions were used to classify six basic expressions. There are also several methods merely based on 3D landmark analysis, like in [18]. However, these methods depend on the detection of high number of fiducial points (83 landmarks), which can be tedious if done manually and perhaps, inaccurate, if done automatically.

Another approach is to perform recognition by fitting face models, for instance, morphable 3D face models as in [10], which can be computationally quite expensive. Some authors used 2D models, like Active Shape and Active Appearance models, to track facial points in the 2D luminance images that are in correspondence with 3D data and then extracted 3D features [20], hence they have the disadvantage of 2D luminance data dependency.

Comprehensive evaluations of 3D versus 2D recognition by analysing 25 AUs are available in [15, 16], where Gabor wavelets are applied on facial surfaces. It has been shown that high quality 3D has significantly higher performance in general, with the exception of some eye related AUs, even under good illumination and under the same frontal pose.

Although there is considerable amount of expression recognition work with 3D data, their applicability is limited since the sensors used are very expensive and not practical to deploy for many application scenarios. Fortunately, recently developed low cost 3D sensors have increased the feasibility of low cost, high throughput acquisition. On the other hand, low quality depth acquisition makes the extraction of facial expression related information highly challenging, compared to high fidelity data used in previous work. Also, many of them perform experiments on manually and carefully segmented 3D face data. Currently, face detection on low fidelity depth data is possible with good accuracy [2], however analysis of facial expressions on highly degraded 3D data is major challenge.

3. SBIA RGB-D Affect Database

There is no any sufficiently large publicly available database to study expression recognition with consumer depth cameras, to the best of our knowledge. A recent 3D dataset acquired by a Kinect sensor is presented in [9], where 451 video segments are labelled according to 12 complex mental states. However, there are only 7 subjects. Also, it consists of many hand-over-face gestures, greatly reducing the number of unoccluded faces. Therefore, we prepared a sufficiently large database for our study [1].

Our database is composed of RGB images and depth maps which were recorded in sync via Kinect sensor, in 640×480 pixels resolution, and were registered in the spatial domain. Distance of subjects to camera was about 100 cm, (depending on their movements, it can range from 80 to 120 cm). Consequently, the eye-to-eye distance of subjects is about 40 pixels on average. All the sessions were recorded in the same good illumination condition in a studio environment. An example acquisition is shown in Figure 2.

[1] SBIA database is currently being prepared for release. Up to 5 subjects can be made available for testing of algorithms, on a collaborative basis.

The dataset is composed of semi-spontaneous facial expressions since it involves spontaneous interactions between actors and professional directors, as well as acting based on scripts which were supervised by the directors. The subjects were free to rotate their heads in any direction or to speak as in a real-world setting. The samples are the short segments of the facial expressions cropped from the original footage which give rise to positive and negative feelings, as well as the segments without emotional expressions. The apex frames of the facial expressions are also annotated.

There are 707 segments from 20 subjects. The dataset is divided into three affect valence classes as positive, negative and neutral valence. The sample size of positive, negative and neutral classes are 317, 337 and 53, respectively. The samples of positive and negative emotions involve various emotions, such as joy, happiness, affection, pleasure and pleasant surprise as positive samples, and anger, disgust, dislike, fear, startled surprise, and unpleasant surprise as negatives. Some instances from our dataset are shown in Figure 1. In each class, there are different type of expressions with various intensity levels and considerable out-of-plane head rotations are involved.

4. Fully Automatic 3D Expression Recognition

The basic stages of our 3D system are depicted in Figure 2. The initial stage is pre-processing, where smooth 3D point clouds are extracted from the facial surfaces (Section 4.1). In the next stage, we approximate facial surfaces by local robust estimation (Section 4.2) in the form of point-based surfaces. Our meshless surface approximation algorithm provides sub-sampling for fast processing, smoothes the point cloud and estimates normals for each point, all in one iteration loop. The third stage is alignment. Having obtained filtered and segmented facial surface point clouds, we can safely apply standard Iterative Closest Point (ICP) algorithm to align the faces to a common coordinate frame. We employ a neutral face as the alignment target as shown in Figure 2. Our descriptors require curvature values, therefore, the fourth stage is curvature estimation (Section 4.3). The curvature estimation method is based on normal section curvature approximation and can efficiently work on point-based surface approximations. Finally, we design histogram-based curvature descriptors to train valence classifiers, using feature selection, as described in Section 4.4. This creates a fully automatic 3D valence recognizer. We now describe each of these steps in broader detail.

4.1. Pre-processing

It is crucial to perform adequate filtering for both 3D face alignment and feature extraction. Solution of automatic ICP alignment can be trapped in local minima because of high noise, and because of dissimilarities due to head rotations or

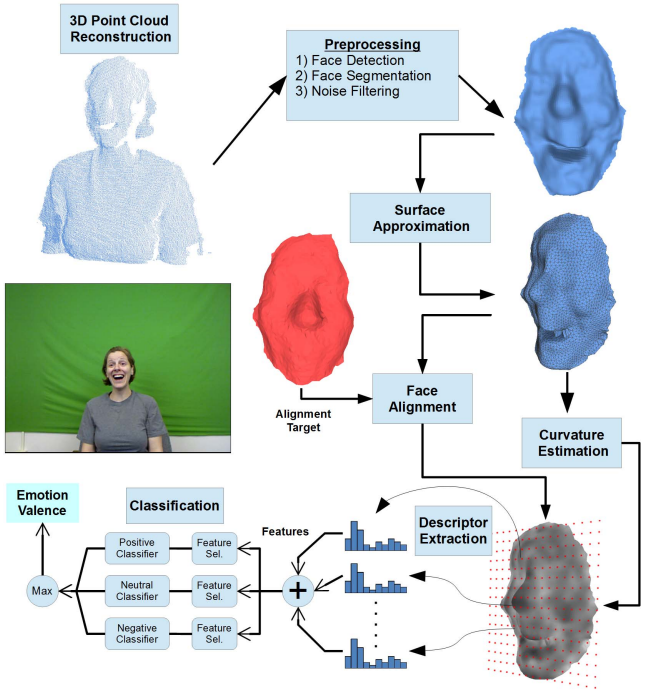

Figure 2. Flowchart of the fully automatic 3D emotion valence recognizer. After 3D reconstruction, the operations in order are: preprocessing, surface approximation, face alignment, curvature estimation, descriptor extraction, feature selection and classification. For visualization, intermediate output faces are rendered as mesh surfaces with artificial lighting, however the system only works on depth maps and point clouds.

surrounding non-face regions like hair, neck, etc. Recognition performance can also be affected from degradation on the features due to excessive noise and holes. To alleviate these issues, we perform several filtering operations both on the noisy depth maps and reconstructed point clouds.

Preprocessing starts with detecting faces using the depth maps. We apply a face detection method which uses random regression forests [2]. In the next step, point clouds are reconstructed from depth maps using sensor calibration parameters, and to remove most of the non-facial points, point clouds are cropped by a polyhedron (height: 200mm, width: 140mm, depth: 170mm) which is positioned where the face detector finds the face. Finally, Euclidean distance clustering [12] is applied and the biggest cluster is selected as the face region. Thus, possible non-facial parts like hair, neck and ears are removed.

In addition, we do hole filling and noise smoothing on the depth map images. Holes are closed by gray-scale morphological closing. Then, Gaussian filtering is applied to obtain smooth facial depth maps. However, before applying the smoothing mask, face boundaries are extrapolated appropriately by morphological operations so as not to introduce abrupt artefacts at the face boundaries.

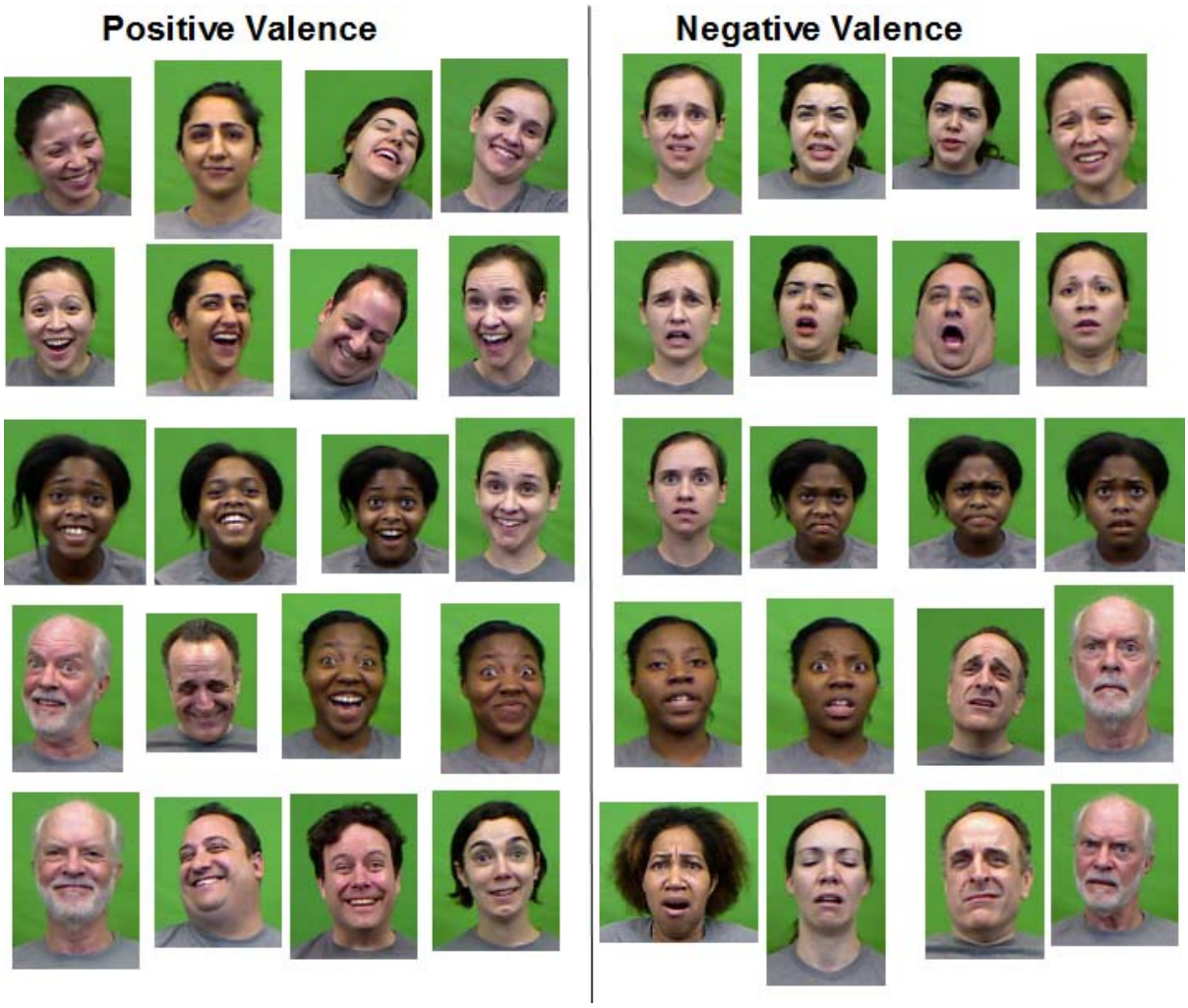

Figure 1. Representative expressions from our experimental dataset. Emotion such as joy, happiness, affection, pleasure and pleasant surprise are categorized in the positive class; while emotions such as anger, disgust, dislike, fear, startled surprise, and unpleasant surprise are categorized in the negative class.

4.2. Robust Surface Approximation

To obtain smooth representation of facial surfaces and for efficient processing, we have developed a meshless surface approximation algorithm. Various techniques have been proposed for point based implicit surface reconstruction. For instance, Hoppe et al. [7] perform principal component analysis to obtain least squares fitting of tangent planes, which serves as local linear surface approximation. Many authors developed techniques based on moving least squares (MLS) approach. Since approximations in least square sense are sensitive to outliers, there is also MLS-based work dealing with the robustness issues [4]. However, existing methods are mostly proposed for computer graphics applications where aesthetics are crucial. On the other hand, our goal here is to develop a fast method which makes it sufficiently suitable to analyse noisy facial surface points for automatic expression recognition.

With this goal in mind, we follow the tangent plane approximation approach due to its low complexity. Our method performs robust estimation to handle noisy range data, and does not require the initial normal estimation step in contrast to MLS methods. Also, data is sub-sampled during surface approximation to reduce computations in the following stages.

In order to approximate the facial surface computationally efficiently, we use a representation based only on a set of points and tangent planes. This implicit surface representation is formally defined as a zero-set $Z(f) = \{\mathbf{x} \mid f(\mathbf{x}) = 0\}$ where $f(\mathbf{x}) : \Re^3 \rightarrow \Re$ is a signed distance function. The distance between an arbitrary point $\mathbf{p} \in \Re^3$ and the surface is calculated by measuring the distance from $\mathbf{p}$ to the closest point on the surface, and the sign of the distance is

determined according to the side of the surface that $\mathbf{p}$ lies. The signed distance function $\mathbf{f_i}(\mathbf{p})$ is evaluated by finding the tangent plane $\mathbf{Tp_i}$ with normal $\mathbf{n_i}$ whose center $\mathbf{q_i}$ is the closest point on the point-based surface S as

$$f_i(\mathbf{p}) = \mathbf{n_i} \cdot (\mathbf{p} - \mathbf{q_i}). \quad (1)$$

By estimating these tangent planes, we obtain local linear approximation of the surface. Hence, our goal is to find a set of points and normals $S = \{\mathbf{q_i}; \mathbf{n_i} \mid f_i(\mathbf{q_i}) = 0 \mid i = 1, \ldots, m\}$.

If we are given an input point cloud $P = \{\mathbf{p_i} \in R^3 \mid i = 1, \ldots, n\}$, then we obtain the tangent plane at any point $\mathbf{p_i} \in P$, via robust estimation over the support set of point $\mathbf{p_i}$ determined according to radius r_S,

$$N_{Pi} = \{\mathbf{p_j} \in P \mid i \neq j \mid |\mathbf{p_i} - \mathbf{p_j}| \leq r_S\}. \quad (2)$$

For robust estimation of the tangent planes, we use M-estimator Sample and Consensus (MSAC) [19], which is a variant of the robust RANSAC algorithm by utilizing M-estimator. It employs quadratic loss at small error and constant loss at large error to prevent huge loss due to the outliers. Therefore, at each iteration of MSAC, we construct a plane using three points, $\{\mathbf{p_{s0}}, \mathbf{p_{s1}}, \mathbf{p_{s2}}\}$, which are sampled from the support set N_{Pi} (Equation 2) by rejection. The plane is defined by the normal vector $\mathbf{n_i}$ and the constant offset $d_i = -\mathbf{n_i}.\mathbf{p_{s0}}$. The error at each support point is the signed distance given by Equation 1.

However, the direction of the tangent planes are inevitably ambiguous; because, there are two planes passing through three non-collinear points in 3D space, which are located at the same position but have opposite directions. Knowing the viewpoint $\mathbf{v}$ (in our case $\mathbf{v} = \mathbf{0}$), this actually can be resolved since the normals should be toward the viewpoint. This correction is realized by orienting the tangent planes so that they satisfy $\mathbf{n_i} \cdot (\mathbf{v} - \mathbf{p_i}) > 0$. Once the best fitting tangent plane is found by MSAC, the approximated surface point $\mathbf{q}$ satisfying $f_i(\mathbf{q}) = 0$ is obtained by projecting the evaluation point $\mathbf{p_i}$ onto the tangent plane.

$$\mathbf{q} \leftarrow \mathbf{p_i} - \mathbf{n_i}(\mathbf{p_i} \cdot \mathbf{n_i} + d_i)/(\mathbf{n_i} \cdot \mathbf{n_i}) \quad (3)$$

Finally, we integrate sub-sampling functionality into our point-based approximation method. This is simply achieved by keeping track of the neighbourhood of the evaluated points in P, during the iterations, so that those neighbouring points are skipped. Neighbourhood set is defined as $D_i = \{\mathbf{p_j} \in P \mid \mathbf{p_i} \neq \mathbf{p_j} \mid |\mathbf{p_i} - \mathbf{p_j}| \leq r_{sub}\}$ and point density is controlled by the neighbourhood radius r_{sub}. Pseudo-code of our surface approximation algorithm is as follows (Note that the algorithm is implemented by using k-d tree structure for fast processing).

P: input point cloud
S: point based surface approximation

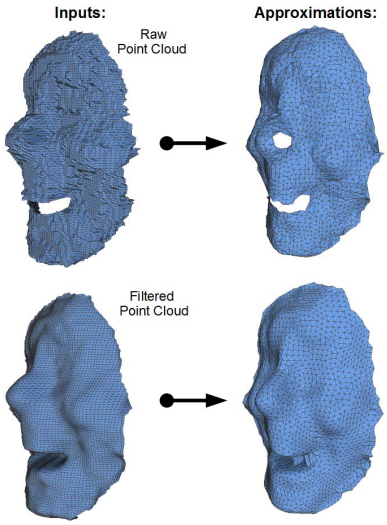

Figure 3. Point-based surface approximations shown for raw and depth filtered inputs by mesh-based surface rendering.

I: set of indices to keep track of occupied points
Initializations: $S \leftarrow \emptyset$, $I \leftarrow \emptyset$
for all $\mathbf{p_i} \in P$ **do**
  if $i \in I$ **then**
    skip to next iteration
  end if
  Construct support set of point $\mathbf{p_i}$, N_{Pi}, with radius r_S
  $\mathbf{Tp_i} \leftarrow MSAC_{Plane}(\mathbf{p_i}, N_{Pi})$
  $\mathbf{n} \leftarrow$ normal of tangent plane $\mathbf{Tp_i}$
  if $\mathbf{n} \cdot (\mathbf{v} - \mathbf{p_i}) < 0$ **then**
    $\mathbf{n} \leftarrow$ flip $\mathbf{n}$ towards viewpoint $\mathbf{v}$
  end if
  $\mathbf{q} \leftarrow$ Projection of $\mathbf{p_i}$ on tangent plane $\mathbf{Tp_i}$
  $S \leftarrow S \cup (\mathbf{q}; \mathbf{n})$
  Construct r_{sub} neighbourhood of points, D_i
  $I \leftarrow I \cup D_i$
end for

We set $r_S = 4mm$ in order to retain as much detail as possible while keeping sufficient number of points for estimation. The sub-sampling radius parameter is set as $r_{sub} = 2.5mm$ so that substantial reduction is achieved on the number of points that approximates the underlying surface. We determined these parameters by visual observation, as well as by recognition tests. For instance, surface approximation outputs of raw and depth filtered inputs are shown in Figure 3. For the filtered inputs, average number of points in the dataset is reduced from 7358 points to 2911.

4.3. Curvature Estimation on Point Clouds

A plethora of curvature estimation algorithms exist in the literature, mostly because curvature is quite sensitive to noise. The estimation problem becomes even more crucial

with low cost range acquisition devices since underlying surfaces are severely noisy. A comparison of several methods for range data is provided in [8], though they consider relatively higher quality data. Most of the methods require triangular mesh representation of the surfaces, or fit polynomial surface patches for estimation, often quadric or cubic. To handle the noise some authors use methods like robust statistical estimation techniques, while others prefer to apply pre-processing on normal and curvature tensor fields, or post-processing. However, these techniques introduce additional complexity and increased computation times.

In our study, we estimate curvatures by fitting to normal section curvatures [23]. Estimation runs directly on a point cloud with normal vectors, i.e., there is no mesh reconstruction or polynomial surface patch fitting requirement, and thus computational burden is lower. We apply curvature estimation on the point-based surface approximation (see Section 4.2), S, where the normal vector at each point is already available.

Estimation of curvature at the surface point $\mathbf{p_i} \in S$ is carried out over a neighbourhood within radius r_C, defined by the set $N_{Ci} = \{\mathbf{q_j} \in S \mid i \neq j \mid |\mathbf{q_i} - \mathbf{q_j}| \leq r_C\}$. The neighbouring point obtained by fast k-d tree search. The procedure has two steps. First, for every neighbour point $\mathbf{q_i} \in N_{Ci}$, a normal section circle is constructed (osculating circle) that connects $\mathbf{q_i}$ to $\mathbf{q_j}$ according to the normal vectors $\mathbf{n_i}$ and $\mathbf{n_j}$ at these points. The osculating circles are used to estimate the normal section curvatures k_n^j.

In the second step, least squares fitting on the normal section curvatures is done through the Euler formula of the curvature.

$$k_n^j = k_1 \cos^2(\beta_j) + k_2 \sin^2(\beta_j) \tag{4}$$

Here, (k_1, k_2) are the unknown principal curvatures. β_j is the angle between the normal section direction and the first principal direction. Though β_j is unknown, it can be expanded as a sum of a known angle and the unknown angle which depends on the first principal coordinate. Then, it is re-written in the form suitable for least squares solution of the unknowns. The details of the procedure, to obtain principal curvatures and directions, are given in [23].

We empirically determined the best working curvature estimation radius as $r_C = 8mm$. On average 18 points fall within this radius when applied on the simplified surface point clouds via the method described in Section 4.2.

4.4. Recognition with Curvature Descriptor

We design simple local histogram-based descriptors using estimated curvatures. Expressions deform the facial surface, and surface curvature is a good indicator of these deformations since it is the measure of local surface bending. Being the second-order local surface feature, curvature also has the advantage of rotation invariance. It has been shown previously [15] that, in facial expression recognition, mean curvature performs better than other curvature features, such as Gaussian curvature, Shape Index and curvedness. Therefore, we employ mean curvatures.

We first create local histograms $\mathbf{H}_i$ of mean curvatures, $mc = (k_1 + k_2)/2$, over a neighbourhood of the surface point $\mathbf{q_i} \in S$ within a radius r_H defined by $N_{Hi} = \{\mathbf{q_j} \in S \mid i \neq j \mid |\mathbf{q_i} - \mathbf{q_j}| \leq r_H\}$. We found the histogram parameters empirically, as $r_H = 16mm$, 32 bins, and quantization range of $[-0.2, 0.2]$. Values outside of this range are clipped.

The local histograms are extracted according to a uniform rectangular grid laid in front of a facial surface after alignment (see Figure 2). The size of the rectangular grid is determined by the mean of the face bounding boxes, which is 137 mm in width and 198 mm in height. A ray is cast at each grid node which is then intersected with the point cloud. We employ a 16×16 grid. In order to find the intersection points rapidly, we transform point clouds to octree-voxel representation which partitions the space covered by point cloud into voxels. To obtain octree partitioning, we use the implementation in Point Cloud Library [12]. Histograms are evaluated at the intersection points, are normalized and the surface point cloud descriptor is constructed by concatenating the normalized bin values of every histogram into a single vector.

We apply state-of-the-art AdaBoost in combination with Support Vector Machine (SVM) method [15] for recognition of positive, negative and neutral valence states. As described in Section 3, positive and negative classes are comprised of different kind of facial expressions while the neutral class only contains neutral faces. Training is performed on the apex frames of which annotations are given with the database. We first train binary classifiers and then apply one-vs-all strategy for three-class classification. Decision is made by choosing the output of the classifier with the maximum score. AdaBoost is applied for the purpose of feature selection. It runs with the nearest mean classifier on the training set until either there is no performance gain observed or a maximum 200 features are selected. After selecting the discriminative features, i.e., histogram bins, via Adaboost, linear SVM classifiers are trained. Training involves hyper-parameter optimization via cross validation on the training sets.

5. Experimental Results and Discussions

We do experimental evaluations in two lanes. First, we compare our point cloud based 3D-only method with state-of-the art 3D recognition. The importance of 3D-only recognition is that, it can robustly work in wide range of illumination conditions, even in pitch-black dark rooms, whereas color channel degrades severely. Secondly, we evaluate use of luminance data, to see the gain via fusion

of 3D+2D as well as to compare with 2D-only recognition in good illumination conditions.

We perform the comparisons by 10-fold subject-independent cross validation, i.e., test subjects are not seen in the training sets. Training and testing are performed on the apex frames of facial expressions. Accuracy of each method with its standard error is calculated for evaluation.

5.1. 3D Point Cloud Based Recognition

We compare three types of 3D-only feature extraction. The pre-processing operations, i.e., face detection, alignment and filtering, are all the same for these three methods. The first row in Table 1 is the surface curvature image based recognition described in [15]. In that method, first, a triangular mesh is generated, and mesh-based discrete curvature estimation is applied. Next, the surface mesh is mapped onto 2D domain in 96×96 pixel resolution image by orthographical projection. Then mean curvature value at each image pixel is obtained by interpolation using barycentric coordinates of mesh triangles. Finally, for each pixel, magnitudes of various Gabor wavelet filter responses are calculated. 20 wavelets, corresponding to four orientations and five scales from four to 16 pixels in half-octave intervals, are applied at each pixel. This results in total 184320 features. We refer [15] for further details of this method.

In the second row of Table 1 we again apply the same Gabor wavelet features, however curvature values are estimated using normal section curvature approximation as explained in Section 4.3 instead of mesh-based discrete curvature estimation. On the other hand, the method in the third row is our completely point cloud based method. Feature vector size of point cloud descriptors is only $(16 \times 16)(gridsize) \times 32bins = 8192$, far less than of the Gabor wavelets.

We see in Table 1 that our point cloud descriptor based method (third row) obtains 77.4% accuracy, which is larger than the 74.2% accuracy of the surface image Gabor method (first row). However, the Gabor method attains 77.2% accuracy if we estimate curvatures over the point clouds. This result points out the importance of curvature estimation component of our method for recognition using low fidelity data of consumer level depth cameras. On the other hand, while the accuracy of our point cloud descriptor is at par with the Gabor method if the same curvature estimation is employed, it provides substantial reductions in computations. This is due to far less number of features, surface point cloud simplification, and no need for surface image feature extraction operations (i.e., mesh formation, surface image generation and Gabor filtering).

5.2. Fusion with Luminance Data

We fuse our 3D-only features with luminance-based features in order to benefit from information provided by facial texture. For luminance-based feature extraction, we use Gabor wavelets since it is a well proven technique as shown in various studies [15]. To apply Gabor wavelets, surface texture images are created by mapping luminance values on the point clouds using orthographical projection and triangular meshes. Exactly the same procedure and parameters as we do in Section 5.1 while extracting mean curvature image Gabor wavelets, is applied for this purpose. The luminance feature vector is concatenated with the 3D feature vector to obtain a hybrid feature vector. Then we run AdaBoost which treats each element in the hybrid feature vector as a weak classifier. Thus, the most discriminative and complementary feature set is selected. AdaBoost is terminated at 200 features, hence the number of features are the same for all the methods. Finally, we apply standard normalization on those selected features and train SVM classifiers.

Curvature Est.	Features	Accuracy
Mesh	Surface Image Gabor	74.2 ± 2.91
Point Cloud	Surface Image Gabor	77.2 ± 2.91
Point Cloud	Point Cl. Descriptor	77.4 ± 3.10

Table 1. Recognition accuracies with standard errors for 3D-only methods. For all the methods, features are based on estimated mean curvature values.

In Table 2, we see that 3D+2D feature fusion achieves 89.4% accuracy, which is a big improvement compared to the 3D-only performance of 77.4%. Improvements by 3D+2D fusion are frequent in previous work, however they are small compared to our comparison (2% in [15]). We attribute this difference to the use of low fidelity depth data. In contrast to high quality 3D data of previous work, amount of facial expression related information is less and also more difficult to extract with low quality data. Therefore, missing information must be compensated from the texture channel.

We also compare with the 2D-only performance of consumer level cameras. The same Gabor wavelets are used on luminance images for the 2D-only method. Facial images are registered by aligning 2D eye coordinates which are found by OpenCV eye detector, and then by scaling to 96×96 pixels resolution. In Table 2, we see that the 2D-only method obtains 84.9% accuracy, which is below than the 3D+2D fusion accuracy while higher than the 3D-only accuracy. The fusion gain with respect to the 2D-only shows the benefit of 3D provided by consumer depth cameras.

6. Conclusion

We have established the use of 3D data from consumer depth cameras for automated recognition of facial expressions, by demonstrating their applicability on the affect valence detection problem. We have developed the first fully automatic and practical 3D expression recognition system.

Method	Accuracy
3D+2D (feature fusion)	89.4 ± 2.21
3D-only (mean curvature)	77.4 ± 3.10
2D-only (luminance)	84.9 ± 2.72

Table 2. Recognition accuracy comparisons (with standard errors) of 3D+2D fusion with 3D-only and 2D-only methods for consumer level depth cameras.

Although these new sensors are very affordable as well as practical, they have the handicap of capturing low quality 3D facial data. Our 3D processing alleviates this difficulty.

We first showed that our 3D-only method obtains significantly higher accuracy than the state-of-the-art Gabor wavelets based 3D recognition. Our experiments revealed that this improvement is due to robust curvature estimation. Another contribution of our method is its computational efficiency since it works directly on point clouds, and employs surface approximation and less number of features.

Second, we showed that under good illumination conditions, fusion of 3D with luminance channel improves both 3D-only and 2D-only performances. While the 2D-only performance is better than the low quality 3D-only performance, this is realistic in good illumination conditions, as we are using only surface features for 3D. However, because of that, the 3D-only features will perform well in bad illumination conditions, where color/luminance images will fail. In addition, 3D helps alleviate the challenge of head pose. In future, we propose to bridge this performance gap by more suitable 3D surface features and recognition methods, or using temporal information. The improvement provided by the fusion underlines the importance of the information provided by 3D.

7. Acknowledgement

This work was supported by grant NIH R01MH073174.

References

[1] P. Ekman. *Emotion in the human face*. Cambridge University Press, Cambridge, UK, 1982.

[2] G. Fanelli, T. Weise, J. Gall, and L. V. Gool. Real time head pose estimation from consumer depth cameras. In *33rd Annual Symposium of the German Association for Pattern Recognition (DAGM'11)*, September 2011.

[3] B. Fasel and J. Luettin. Automatic Facial Expression Analysis: A Survey. *Pattern Recog.*, 36(1):259–275, 2003.

[4] S. Fleishman, D. Cohen-Or, and C. T. Silva. Robust moving least-squares fitting with sharp features. SIGGRAPH '05.

[5] J. R. Fontaine, K. R. Scherer, E. B. Roesch, and P. Ellsworth. The world of emotion is not two-dimensional. *Psychological Science*, 18:1050–1057, 2007.

[6] H. Gunes and M. Pantic. Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions*, 1(1):68–99, January 2010.

[7] H. Hoppe, T. DeRose, T. Duchamp, J. McDonald, and W. Stuetzle. Surface reconstruction from unorganized points. *SIGGRAPH Comput. Graph.*, 26(2):71–78, July 1992.

[8] E. Magid, O. Soldea, and E. Rivlin. A comparison of gaussian and mean curvature estimation methods on triangular meshes of range image data. *Computer Vision and Image Understanding*, 107(3):139 – 159, 2007.

[9] M. Mahmoud, T. Baltru Saitis, P. Robinson, and L. D. Riek. 3d corpus of spontaneous complex mental states. In *Affective computing and intelligent interaction*, 2011.

[10] S. Ramanathan, A. A. Kassim, Y. V. Venkatesh, and W. S. Wah. Human facial expression recognition using a 3d morphable model. In *IEEE ICIP*, 2006.

[11] J. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161–1178, 1980.

[12] R. Rusu and S. Cousins. 3d is here: Point cloud library (pcl). In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 1 –4, may 2011.

[13] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin. Static and dynamic 3d facial expression recognition: A comprehensive survey. *Image and Vision Computing*, 30(10):683 – 697, 2012.

[14] A. Savran, H. Cao, M. Shah, A. Nenkova, and R. Verma. Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering. In *ACM ICMI'2012, AVEC 2012 Grand Challenge*.

[15] A. Savran, B. Sankur, and M. T. Bilge. Comparative evaluation of 3d vs. 2d modality for automatic detection of facial action units. *Pattern Recognition*, 45(2):767–782, 2012.

[16] A. Savran, B. Sankur, and M. Taha Bilge. Regression-based intensity estimation of facial action units. *Image Vision Computing*, 30(10):774–784, Oct. 2012.

[17] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic. Avec 2011, the audio/visual emotion challenge. In *AVEC 2011 Grand Challenge*.

[18] H. Tang and T. Huang. 3d facial expression recognition based on automatically selected features. In *IEEE CVPR Workshop on 3D Face Processing*, USA, 2008.

[19] P. H. S. Torr and D. W. Murray. The development and comparison of robust methods for estimating the fundamental matrix. *International Journal of Computer Vision*, 24:271–300, 1997.

[20] F. Tsalakanidou and S. Malassiotis. Real-time 2d+3d facial action and expression recognition. *Pattern Recognition*, 43(5):1763 – 1775, 2010.

[21] J. Wang, L. Yin, X. Wei, and Y. Sun. 3d facial expression recognition based on primitive surface feature distribution. In *IEEE CVPR*, Washington, DC, USA, 2006.

[22] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3d facial expression database for facial behavior research. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, Washington, DC, USA, 2006.

[23] X. Zhang, H. Li, and C. Zhanglin. Curvature estimation of 3d point cloud surfaces. In *Proceedings of AsiaGraph*, Tokyo, Japan, October 23-26 2008.