FASHE: A FrActal based Strategy for Head pose Estimation

Carmen Bisogni, *Member, IEEE,* Michele Nappi, *Senior Member, IEEE,* Chiara Pero, *Member, IEEE,* and Stefano Ricciardi, *Member, IEEE*

Published in: IEEE Transactions on Image Processing journal.

The Version of Record is available online at: http://dx.doi.org/10.1109/TIP.2021.3059409

***Abstract*—Head pose estimation (HPE) represents a topic central to many relevant research fields and characterized by a wide application range. In particular, HPE performed using a singular RGB frame is particular suitable to be applied at best-frame-selection problems. This explains a growing interest witnessed by a large number of contributions, most of which exploit deep learning architectures and require extensive training sessions to achieve accuracy and robustness in estimating head rotations on three axes. However, methods alternative to machine learning approaches could be capable of similar if not better performance. To this regard, we present FASHE, an approach based on partitioned iterated function systems (PIFS) to represent auto-similarities within face image through a contractive affine function transforming the domain blocks extracted only once by a single frontal reference image, in a good approximation of the range blocks which the target image has been partitioned into. Pose estimation is achieved by finding the closest match between fractal code of target image and a reference array by means of Hamming distance. The results of experiments conducted exceed the state of the art on both Biwi and Ponting'04 datasets as well as approaching those of the best performing methods on the challenging AFLW2000 database. In addition, the applications to GOTCHA Video Dataset demonstrate that FASHE successfully operates in-the-wild.**

***Index Terms*—Head pose estimation, partitioned iterated function systems, fractal encoding, face recognition.**

I. Introduction

With regard to pattern recognition, the term "head pose", which actually refers more specifically to face orientation in 3D space, represents a very popular keyword due to the vast number of related research topics and applications [1] [2]. It is a fact, indeed, that determining how human face is rotated with respect to an imaging sensor may provide crucial information for many tasks such as face recognition and analysis [3], body tracking [4], face frontalization, gaze estimation [5], [6], person re-identification [7], just to name a few. For most of these purposes the ideal head pose estimation (hereafter HPE) method should be able to determine three angular values respectively for yaw, pitch and roll axes, associated to head's degrees of freedom, with the highest possible accuracy. HPE can be determined using different source of information. In this paper we propose a method based on a singular RGB image, that is the only source available when we want to perform best-frame-selection in the wild. In fact, in surveillance contexts, it is not usual to have more than a 2D low resolution image. However, HPE in general, can be also performed using depth information. 3D reconstruction results to be a step that is mainly involved in this second category of input. In fact, as demonstrated by recent literature [8] [9], the depth image or frame, together with the RGB frame is significantly involved in HPE. In particular, in [8] it is used directly in the HPE step together with the RGB images, in [9] it is used in the 3D reconstruction but with an initialization that involves the need of a frontal frame of the same subject. We also noticed, however, that by using depth information, e.g. 3D-input method, is it possible to obtain excellent results even if without the 3D reconstruction [10] [11].

Returning to our focus, HPE using a single RGB image, since HPE can be seen as a classification problem, is unsurprising that the majority of the approaches proposed in the last five years are based on machine learning and convolutional/deep neural networks. Though DL/CNN methods have raised the bar in terms of minimizing the estimation error, they inherently require a training stage involving a vast number of positive and negative examples to fully deliver their potential.

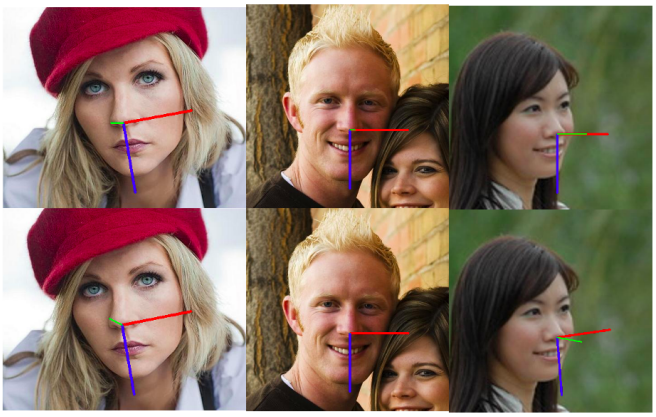

Fig. 1. Visual examples of pose estimation results achieved on images from the AFLW2000 database. Top row: tripods showing ground-truth orientation. Bottom row: tripods showing estimates by PIFS-based method.

However, machine learning could not necessarily be the best performing solution for accurate and reliable HPE, as a few proposals exploiting various image analysis techniques [12] [13] have recently proved. In this context, Partitioned Iterated Function Systems (hereafter PIFS), which are well

C. Bisogni is with the Department of Computer Science, University of Salerno, Italy (e-mail: cbisogni@unisa.it).

M. Nappi is with the Department of Computer Science, University of Salerno, Italy (e-mail: mnappi@unisa.it).

C. Pero is with the Department of Computer Science, University of Salerno, Italy (e-mail: cpero@unisa.it).

S. Ricciardi is with the Department of Biosciences, University of Molise, Italy (e-mail: stefano.ricciardi@unimol.it).

known for compact representation of digital images through lossy compression techniques, can advantageously be used as an effective way to describe face pose variations in terms of their inherent auto-similarities. We explore the possibility to perform pose estimation by finding the best match between the fractal encoding of the input image and a reference pose-dataset. In this paper we present "FASHE", an improved fractal based HPE method that exploits a single frontal face image as the reference to build only once the domain blocks required for the fractal encoding procedure, thus increasing both accuracy and efficiency of pose estimation.

To the aim of objectively assess the performance improvements possibly achieved with this new fractal encoding process, we conducted experiments on three different database, namely the Pointing'04 dataset, the Biwi Kinect Head Pose Database and the AFLW2000 Automated Facial Landmarks in the Wild dataset. The results achieved show, with regard to mean absolute error, an edge over the state-of-the-art for the Biwi database and a performance close to the best approaches for the other two databases considered. As features extractor for HPE, we also tested the use of several lossy coding schemes such as DCT and Wavelet. However, the self-similarity features embedded inside the PIFS technique demonstrated to be more effective as a descriptor for HPE as we will describe in the following sections.

The main contribution of this paper is the adaptation of a well known PIFS Coding Scheme, widely used in the image lossy compression, for HPE. In this context it demonstrated to achieve the state of the art performances and, for certain axes and environment, also to overcome it. To sum it up, the proposals of this work are the following:

- a new fractal encoding approach to head pose estimation under uncontrolled environment;
- high pose estimation accuracy, rivalling the performance of best machine learning based approaches;
- no needs for training - inherent trustworthiness due to the spatial and temporal characteristics of facial dynamics;
- improved efficiency compared to the classical fractal encoding, since the domain blocks are pre-computed only once, instead than for every image;
- a thorough set of experiments conducted on three different datasets.

The rest of this paper is organized as follows. Section II. reports on related works on the HPE topic. Section III. recalls the math background behind PIFS and fractal image encoding; Section IV. describes each step of the proposed approach; Section V. presents the results of the experiments conducted and, finally, Section VI. concludes the paper summarizing the work done and the achievements.

II. Related works

Given the relevance of HPE problem for many application fields, an impressive number of contributions have been proposed on this topic, particularly throughout the last decade. HPE methods have been classified according to different criteria, but, in the end, they all fall into two main types, depending on whether they use either a 2D intensity image or a 3d range/depth image as input. In the following lines we report about related works belonging to these two macro-categories.

A. 2D Methods

This is by far the largest category, also because working on an intensity image considerably extends the application field of HPE while requiring a much simpler acquisition modality. Most of the contributions belonging to this category exploits machine learning and particularly deep neural architectures. This is the case of [14] where a convolutional neural network (CNN) is used to project face images onto a low-dimensional pose-space. Gourier et al. [15] evaluate the HPE from low-resolution images, training an auto-associative memory and computing it for each pose, using a Widrow-Hoff learning rule. A probabilistic framework for continuous regression is proposed by the authors of [16] to address pose estimation in uncontrolled conditions. On a different line of research, an appearance-based approach is proposed in [17], where the input face image is matched against a pre-defined set of exemplars head poses to find out the most similar one according to the highest matching score. In [18], the authors focus on the orientation of the nose as a reliable indicator of whole head orientation, proving that this feature is highly discriminant for pose classification. Instead of relying on a single feature, detector methods such as [19] train a set of classifiers to recognise relevant poses. In [20], head contour results by classifying patches as either head or background, then, a multi-level structured hybrid forest (MSHF) is developed for accurate pose estimation by means of selected patches sub-regions. Drouard et al. [21] propose a mixture of linear regressions with partially-latent output (hGLLiM). Similarly, a regression technique based on Gaussian mixture of locally-linear mapping is used to extract HOG-based descriptors from face-related bounding boxes and mapping them to corresponding head poses in [22]. An evolution of this method is designed to learn both bounding-box-to-face alignments and head-pose parameters, so that the predicted bounding-box-to-face alignments are similar to those used for training, thus minimizing the impact of background variations on pose prediction.

In [23] a coarse-to-fine HPE framework which models the uniform geometry representation through a unit circle for the coarse layer and a 3-sphere for the fine layer, is proposed. On a similar line of research, the authors of [24] present a deep CNN based on a supervised two-stages initialization strategy, by either projecting a neutral 3D face shape onto the test image or by searching the closest shapes, pose distance wise, in the training set. A further coarse-to-fine approach is also proposed in [25], where the authors exploit global and local CNN features to achieve both landmark detection and hierarchical head pose estimation. In [26] the authors exploits synthesized heads to build a 3D head centroids metric space and achieve pose estimation by finding the best match between the 2D head landmarks associated to the query image and the combined 3D head centroids/pose hypotheses. To the aim of addressing head pose estimation in multiple-cameras monitored environments,

flexible graph-guided multi-task learning is used to learn multiple region-based classifiers in [27]. On a different take, the authors of [28] propose to use dictionary learning along with a sparse-representation based classifier based to achieve greater pose classification robustness. A regression approach on support vector machines (SVR) is applied to HoG feature is used in [29] for pose estimation on low resolution images. Aiming at HPE in the wild the authors of [30] combine CNN and adaptive gradient to improve accuracy and robustness. Heatmap-CNN trained by 3D-pose, face's visibility is used in [31] to learn regressors and to obtain key-points estimation and pose prediction. The system in [32] adopts a neural network approach to estimate the head orientation from facial images. A synthetic, procedurally annotated, dataset is used to train a CNN and solve the HPE regression problem in [33].

A multi-loss network to estimate head pose Euler angles directly from face image intensities using ResNet50 is described in [34], while deep multi-task learning is used in [35] to learn shared features from low-res intensity images. Multi-task learning is also used in [36] with particular attention to achieving real-time performance in both face detection and pose estimation by means of a cascaded multi-CNN architecture. The work in [37] estimates head poses from color images without depth information via a CNN trained by combining L2 regression loss with ordinal regression loss in a multi-loss scheme. Two joint CNNs specialized for head pose and full-body pose estimation, respectively, are proposed in [38], while the authors of [39] improve estimation accuracy fusing the hidden layers of a first CNN by means of a second CNN along with a multi-task learning algorithm operating on the fused features. A fine-grained structure mapping strategy is proposed in [40] where regression is used to achieve features spatial grouping before aggregation instead of exploiting their spatial relationship or landmark estimation. A recent coarse-to-fine approach featuring two subnetworks trained jointly is presented in [41] where the first stage classifies the query image results into four categories, while the subsequent fine regression stage outputs an accurate pose estimation. Not relying on any learning strategy, the authors of [12] use a quad-tree based representation to model input face pose, with regard to detected landmarks, and then match the resulting binary pose-feature vector to a reference array containing a discrete set of poses to find the best pose estimate. Finally, in [13] instead of learning methods a peculiar Web-Shaped Model (WSM) is applied to a set of previously detected facial landmarks by centering the web on the nose-tip and therefore associating each landmark to a face sector.

B. 3D Methods

Thanks to the growing availability and diffusion of more affordable and portable 3D sensors, the number of HPE approaches exploiting the extra information provided by range/depth images have grown as well, yet it still represents a small fraction of the so called 2D methods. Among the first to work on 3D HPE the authors of [42] aim at finding the nose region in the range image provided as query and therefore exploit a generative algorithm to evaluates a large number of hypotheses by means of parallel computing provided by GPU vector processors. Range data are also used in [19] for pose estimation through a random-forest based regression in which synthetic 3D face renderings are used to train the regressor on labeled data. A related approach is proposed in [43] where a random forest-based framework combined with a voting strategy applied to patches derived from the input range image, is able to cope with large poses and partial occlusions. Also based on random forests, yet trained by SIFT-HOG features, is the method described in [44]. The notion of central profile, a unique characteristic curve defined over the tridimensional facial surface, along with the Hough transform, are exploited thanks to a voting mechanism allowing to determine the symmetry plane and the associated pose in [45]. Realtime HPE by means of affordable Microsoft Kinect depth sensor is proposed in [46], where a viewpoint invariant triangular surface patch descriptor is used to map input face 3d shape into a triangular region and to match it to a previously built reference gallery. The Kinect's RGB and depth cameras are also exploited by [47], where face detection and localization operates on the RGB image aided by depth information and pose estimation is achieved by SVM regressors, through features combining both color and range data. Microsoft Kinect v2 rotational and translational precision are evaluated in the context of real-time HPE by the authors of [48]. Iterative Closest Point algorithm along with particle swarm optimization are used in [49] to address extreme face rotations and partial occlusions by operating on a weighted-vertices morphable face model registered onto captured 3D data. In [50] pose estimation of a single 2D face image is achieved by means of a derived 3D representation achieved by morphing an ethnically coherent prototype face model toward the target query while minimizing the distance between a set of key facial features. The authors of [51] propose an approach which converting the annotated 2D landmark annotation available in many datasets to 3D data, results in a large database containing 230.000 3D facial landmarks to be used for training neural networks for improved HPE. Finally, 3D Dense Face Alignment is proposed in [52], by fitting and aligning a 3D face model to query 2D image through a CNN.

The PIFS based pose estimation approach proposed in this paper belongs to the category of "2D methods without learning" and represents a new application of fractal encoding to head pose estimation problem.

III. Background

Fractals are of great interest in Mathematics since the late 19th century. The first implementation of a Partitioned Iterated Function System (PIFS) was developed by Arnaud Jacquin [53]. The reproduction of the image using the fractal code is more compact than the pictures. Details about the fractal encoding and our version of the algorithm are presented in the following subsections.

A. Introduction to fractal encoding

The main idea under PIFS is that a part of an image can be approximated by a transformed and down-sampled version

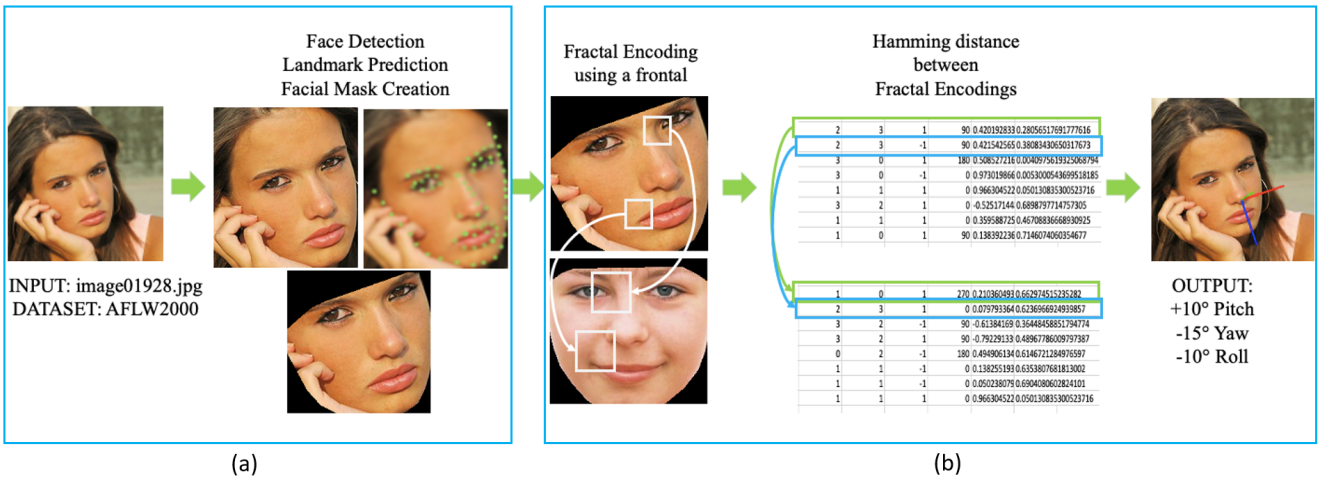

Fig. 2. Workflow of the proposed method. (a) represents the preprocessing, (b) represents the fractal encoding and pose estimation.

of another part of the same image. This property called self-similarity is used in PIFS to compress the image. In order to define PIFS mathematically, we will give some preliminary definitions.

Definition 1: Metric space A Metric Space is an ordered pair (M,d) where M is a set and d is a metric, i.e. a function from MxM to $\mathbb{R}$ such that for any x,y,z in M:

$$\begin{aligned} d(x,y) = 0 \Leftrightarrow x = y \\ d(x,y) = d(y,x) \\ d(x,z) \leq d(x,y) + d(y,z) \end{aligned} \tag{1}$$

Definition 2: Contraction mapping A contraction mapping on a metric space (M,d) is a function f from M to itself, with the property:

$$d(f(x),f(y)) \leq kd(x,y) \tag{2}$$

for all x and y in M where k is a nonnegative real number between 0 included and 1 excluded.
The same definition is true when the map is defined between two different metric spaces.

Theorem 1: Fixed Point Theorem In a complete metric space (M,d) if $f : M \to M$ is a contractive transformation with parameter k, then exist and it is unique, a fixed point $x_i \in M$ such that

$$f(x_i) = x_i \tag{3}$$

and for any point x in M is also true

$$\lim_{n\to\infty} f^n(x) = \lim_{n\to\infty} \underbrace{f(f(f(...(x))))}_{n\,times} = x_i \tag{4}$$

The meaning of the fixed point theorem on an image is the following. We consider images as points in a metric space. Our aim is to find a contractive transformation on this space that has as fixed point the image we want to encode. If we are able to find this transformation, for the fixed point theorem the distance between the point transformed by this contractive function and the fixed point is less than the distance between the initial point and the fixed point. In other words, if we iteratively apply the contractive transformation to an initial point, the transformation result will be closer and closer to the fixed point.

Definition 3: Affine transformation If we consider a grayscale image, an affine transformation is defined as

$$W(X) = AX + B \tag{5}$$

where A is the transformation matrix, X is the array of the image composed by (x,y) coordinates and z gray level, and B is an offset vector. By affine transformation an image can be translated, rotated, scaled or modified in contrast and brightness.

Definition 4: IFS An Iterated Function System is a set F of contractive affine transformations $f_1,...,f_N$, which is itself a contractive transformation. The contraction parameters of F is the maximum of the parameters of its contractive affine transformations. Furthermore, its fixed point X, also called the *attractor*, is such that

$$F(X) = \bigcup_{i=1}^{N} f_i(X) = X \tag{6}$$

Using all this mathematical backgrounds, Jacquin defined the PIFS approximating a part of an image by a transformed and downsampled part of the same image. For this reason, the components of a PIFS system are: a complete metric space M; a collection of sub-domains $D_i \in M$ and a collection of contractive maps f_i.

The steps of the PIFS are the following:

- The image to be encoded is partitioned in R_i, non overlapping range blocks.
- The same image is also partitioned in larger non overlapping blocks D_j called domain blocks.
- Find for every range block R_i, a domain block D_{R_i} such that a contractive affine trasformation f, transform this Domain block in a good approximation of the range block.

It is clear that when a fractal encoding is used, the developer must decide the size of the Domain Blocks and the size of the Range Blocks. Generally domain and range blocks are square and the differences between them means the level of

compression. In FASHE we used an optimization algorithm to speed the search of the blocks in fractal encoding [54].

B. HPE-optimized fractal encoding

If we want to apply fractal encoding to HPE in a classical way, both the domain blocks and the range blocks will come from the same image. This implies that the coded array, representing the positions of the domain blocks after the transformations, is referred to the position of the range blocks of the same image. In addition, the domain blocks should be estimated for every image.

In FASHE, other than apply fractal encoding to this new field, we will use the same reference image to build the domain blocks, regardless of the image we want to estimate the head pose. To this purpose, the best candidate image should have a neutral frontal head pose, 0°Pitch, 0°Yaw, 0°Roll. Consequently, the steps introduced above become the following:

- A frontal reference image is partitioned in large non overlapping blocks D_j called domain blocks.
- The image to be estimated is partitioned in R_i, smaller non overlapping range blocks.
- Find for every range block R_i, a domain block D_{R_i} such that a contractive affine trasformation f, transform this Domain block in a good approximation of the range block.

We also wish to emphasise that the reference image is not required to came from the same subject we want to estimate, rather a generic frontal image of a random subject is used to build the domain block only once in the whole process. Further details are shown in the following sections.

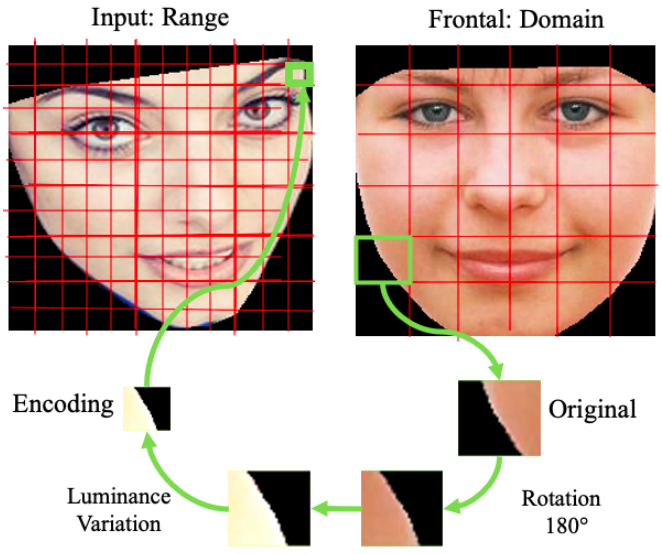

Fig. 3. A detail of the fractal encoding process.

IV. FASHE

In order to apply the fractal theory to estimate the head pose, some preprocessing must be done on the images. The following steps, explained later in details, are the most representative of the overall method.

- Step 1, *Face detection*;
- Step 2, *Landmark prediction*;
- Step 3, *Facial Mask creation*;
- Step 4, *Fractal codec array*;
- Step 5, *Hamming distance*.

First of all, we have to detect the face. In order to do this we used the Viola-Jones (VJ) detector. The VJ face detector [55] is a Real-Time robust and effective algorithm well known in literature. Using the AdaBoost learning algorithm, a classifier is built to select the most relevant features from a very large set of potential features. More classifiers are then used in cascade to fast discard background regions.

Once the squared box with the position of the face is detected, to extract the face more precisely, we use a Landmark predictor [56]. This predictor is very fast, spending only a millisecond per image. It is able to detect 68 relevant point of the face supported by an ensemble of regression trees that performs shape invariant feature. Since the method is a predictor instead of a detector, the landmark will be always 68, even if there are some occlusions on the face. This makes [56] a robust algorithm.

Using the detected boundary of the face, identified between the landmarks 1 and 27, a black mask is created. The obtained image is then resized in order to have the same dimension of the fractal codec array for all the images. As described in section III and in particular, in section III-B, the fractal encoding is applied at the image. The range blocks are built and compared with the domain block of a generic frontal face in order to build the codec array. The codec array dimension will depend on the dimension of the original image and the range-domain block size selected. In our case, the image size is 256x256, and we used a codec of 8x8 pixels as Domain and 4x4 pixels as range, generating a codec matrix of 256 rows and 6 columns. The partitioning made is fixed and the algorithm has been sped up with the selective search of the correspondence between range and domain blocks, as shown in [54]. The computational time required, for this reason, is less than one second per codec, as we further discussed in Table III, where we also show its dependence from the selected range and domain values. In the first and the second columns there are the coordinates of the range block selected, in the third and fourth column the integer representing the inversion and the rotation respectively. In the last two columns there are real value representing the brightness and the contrast applied. This matrix is converted in a 1536 entries array to better perform comparisons.

The comparisons with some models labeled with different poses stored in the dataset is performed using the Hamming distance between the arrays. Given two arrays, one stored and labeled s and one obtained from the image in input to evaluate r, the function used in the Hamming distance [57] is the following

$$\delta(s_i, r_i) = \begin{cases} 1, & \text{if } s_i \neq r_i \\ 0 & \text{if } s_i = r_i \end{cases} \tag{7}$$

where s_i and r_i are the entries of the stored and input arrays, respectively. To evaluate the final distance, the sum is performed

$$d(s,r) = \sum_{i=1}^{n} \delta(s_i, r_i) \tag{8}$$

It is clear that more the encode of the images are different, more the Hamming distance will be higher. And, since different poses produce different encode, the pose of the stored image with minimum Hamming distances compared with the input one, will represent the most similar pose. Consequently, the input image will be labeled with the pose of the image with the minimum Hamming distance.

It is important to underline that if the image in input is very rich in information in the background or because of makeup or hair on the face, before to perform the abovementioned steps is required a filter. To understand when the filter is necessary, the overall entropy of the image is estimated

$$E = -\sum_{i} p_i \log_2(p_i) \tag{9}$$

where p_i is the probability of the i-th gray level. When the entropy is higher than the mean entropy of the images stored in the model, it can be supposed that there is a lot of information in the image that can cause excessive difference between two image even if they have the same head pose. For this reason a Gaussian filter will be applied on the original image in that case. An example of the step performed in this case is represented in Figure 4. This step will be added between module (a) and module (b) of the workflow proposed in Figure 2. More information about the filter involved can be found at [58].

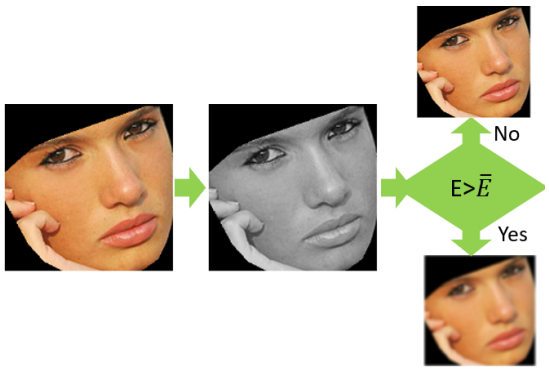

Fig. 4. The additional step for chaotic images.

V. Experiments

A. Datasets

In order to test and evaluate the performance of the proposed scheme, the experiments were conducted on images taken from the *Biwi Kinect Head Pose Database* [59], *Annotated Facial Landmarks in the Wild* (AFLW2000) [60] and *Pointing'04* [61]. Figure 5 shows some image samples of these datasets.

Biwi includes over 15K images of 20 people (6 females and 14 males - 4 people were recorded twice). It contains, for each frame in the captured sequences, both the RGB-image and the depth image. Biwi Kinect Head Pose Database also provides an *.obj* file per subject, which represents the 3D model of the head of the subject, from which the following pose rotations have been obtained:

- Pitch: from -30° to +30°
- Yaw: from -45° to +45°
- Roll: from -20° to +20°

Ground truth is provided in the form of the 3D location of the head and its rotation. A 2D synthetic face image is obtained from 3D model, reproducing the texture characteristic of the original subject; all the figures are annotated with Pitch, Yaw, and Roll parameters, to use them as clear ground truth for the experiments. The total of the poses is 2233 per model/subject, considering all the possible combinations in terms of Pitch, Yaw and Roll angles, for a total of 44660 and the maximum angular support for each of the three degrees of freedom corresponds to 5.

Further experiments were carried out with AFLW2000 dataset and Pointing 04' dataset. Annotated Facial Landmarks in the Wild provides a large-scale collection of annotated face images gathered from Flickr, exhibiting a large variety in appearance (e.g., pose, expression, ethnicity, age, gender) as well as general imaging and environmental conditions. AFLW2000 includes 2000 images from the AFLW dataset, re-annotated for 3D face alignment; the 2D landmarks are skipped in this dataset, since some of the data are not consistent to 21 points, as the original paper mentioned. Pointing'04 dataset consists of 15 subjects and each set includes 2 series of 93 images. The subjects were captured twice, once in each of the two sessions in different poses. The head poses are represented by Yaw and Pitch; Pointing'04 does not contain Roll information.

Fig. 5. Some examples from the datasets. From left to right: Biwi, AFLW2000, Pointing'04.

B. Results with different model configurations

The tests conducted on Biwi adopt a *one left out* strategy, using only 1 subject to test and the others 9 as a model for carrying out the comparisons. Using the test subject in rotation, 10 experiments were carried out. These experiments have a different number of elements in the test set, due to the different percentage of faces detected by the Viola Jones algorithm in the initial stages of the proposed approach. The results refer only to the detected faces. The frontal image to build the domain blocks in the experimental setup is chosen randomly among the subjects included in Biwi. Table I reports the results in terms of MAE for each subset tested. Note the presence of variations in errors between an experimental subset

TABLE I
MEAN OF PITCH, YAW AND ROLL ANGLES ON THE SUBSETS OF BIWI.

Subset	Yaw	Pitch	Roll	overall MAE
1	4.95	8.27	3.02	5.41
2	3.27	4.63	1.42	3.11
3	3.24	3.92	11.6	6.26
4	2.56	3.79	1.46	2.60
5	2.69	3.28	1.38	2.45
6	2.38	3.95	1.51	2.61
7	3.54	5.84	2.03	3.80
8	3.14	3.62	1.53	2.76
9	2.75	4.32	2.15	3.07
10	2.77	4.52	1.32	2.87
mean	**3.13**	**4.61**	**2.74**	**3.50**

TABLE II
AFLW CONFIGURATION GAUSSIAN FILTER.

Dataset	Yaw	Pitch	Roll	MAE
AFLW2000	5.71	7.14	4.5	5.78
AFLW2000_Blur	6.23	7.68	4.06	5.99
AFLW2000_Blur Selective	**4.54**	**6.42**	**3.71**	**4.89**

and the other; this is due to the presence of different geometric conformations in the faces of the subjects.

As regards the AFLW2000 dataset, the choice of Gaussian filter [58] to be applied to the images was made by evaluating their entropy value. The Table II shows the results in terms of MAE for AFLW2000 images tested in different conditions.

- *AFLW2000* is obtained using the proposed fractal encoding algorithm. The Gaussian filter is not applied on the images.
- *AFLW2000_Blur* is obtained using the proposed fractal encoding algorithm and a Gaussian filter with a standard deviation of 7. The application concerns the whole dataset.
- *AFLW2000_Blur Selective* is obtained using the proposed fractal encoding algorithm and a Gaussian filter with a standard deviation of 7. The application only concerns images that have a high entropy value.

As can be seen, the best results are obtained when a selective application of the Gaussian filter is made based on the entropy. This can be interpreted as a normalization in the amount of details in the images.

As claimed in Section IV, our method uses a configuration of 4 as Range and 8 as Domain. To justify this choice we used Biwi to test also other range and domain configurations and highlight that (4,8) represents the best choice. First of all, since we are not interested in compression, we can simply set the domain as double of the range, then we can observe that both range than domain values must exactly divide the dimension of the image, 256 in our case. From those observations we can examine the following couples of ranges and domains, (2,4), (4,8), (8,16), (16,32), (32,64) (64,128) and (128,256). We will exclude from our analysis the first couple, because it needs 2.53 seconds per image to be computed, leading to a non-real time application. We also exclude the last two couples because they leave not enough blocks to compare. Analysing the remaining couples, we will obtain the related errors on Biwi and computational times shown in Table III. From those results we can find the best configuration in (4,8) that we chose in our experiments and comparisons with the state of the art.

TABLE III
DIFFERENT ERRORS AND COMPUTATIONAL TIME IN SECONDS, IN DIFFERENT CONFIGURATIONS OF RANGE AND DOMAIN.

Configuration	**Err_yaw**	**Err_pitch**	**Err_roll**	**Comp_time**
(4,8)	3.13	4.61	2.74	0.1553
(8,16)	3.41	4.81	3.22	0.013
(16,32)	4.13	5.51	4.09	0.0015
(32,64)	9.83	10.08	7.26	0.0002

C. Comparisons with the state of the art

The experiments obtained from the proposed approach are compared with the main proposals in literature and presented in Section II. As mentioned in Section II, these strategies can be divided into two classes: 1) model-based methods and 2) neural network-based methods. Choosing one method over another has its pros and cons. For example, neural network based methods require an adequate training phase, with a significant number of positive and negative examples, but are more reliable for incorrectly identifying of the face landmarks; on the other hand, faster model construction can be achieved through model-based approaches. The performances achieved by means of general methods such as Support Vector Regression (SVR) [62], Gaussian processes - GPR [63] and Partial Least Square Regression - PLS [64] are also reported. Table IV, VI and VII respectively show the comparison of results obtained on Biwi Kinect Head Pose Database [59], AFLW2000 dataset [60] and Pointing'04 [61].

TABLE IV
MEAN ABSOLUTE ERROR OF PITCH, YAW, AND ROLL ANGLES ACROSS DIFFERENT METHODS OVER THE BIWI DATASET. ALL METHODS MARKED WITH (*) DO NOT USE NEURAL NETWORKS OR MACHINE LEARNING APPROACHES. THE (+) SIGN IDENTIFIES THE SYSTEMS THAT USE 3D TECHNIQUES.

Method	Yaw	Pitch	Roll	MAE
QT PYR [12] *	5.41	12.80	6.33	8.18
WSM [13] *	6.21	**3.95**	4.16	4.77
hGLLiM [21]	6.06	7.65	5.62	6.44
CNN-Regression [33]	6	6.1	5.7	5.94
Multi-Loss ResNet50 [34]	5.17	6.97	3.39	5.177
QuatNet [37]	4.01	5.49	2.93	4.14
FSA-Net [40]	4.27	4.96	2.76	3.996
Coarse-to-Fine [41]	4.76	5.48	4.29	4.84
SIFT-HOG [44] +	8.8	8.5	7.4	8.23
SVR [62]	6.98	7.77	5.14	6.63
GPR [63]	7.72	9.64	6.01	7.79
PLS [64]	7.35	7.87	6.11	7.11
FASHE *	**3.13**	4.61	**2.74**	**3.50**

To determine the errors and compare the Mean Absolute Error (MAE), is used as performance index. The MAE represents the distance between the predicted and the effective value, as defined by:

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y_j}| \tag{10}$$

where y_j is the ground truth, i.e the true angular value and $\hat{y_j}$ is the prediction, i.e the predicted angular value. MAE is

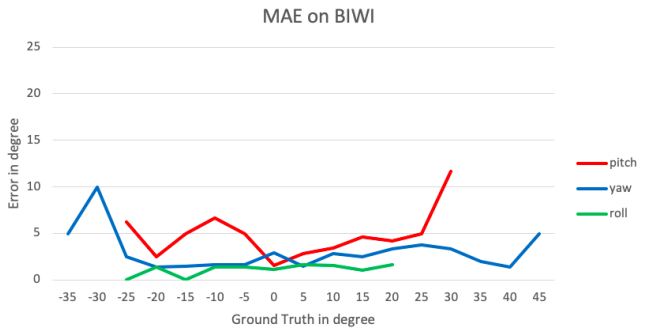

Fig. 6. Errors on BIWI in terms of angular poses.

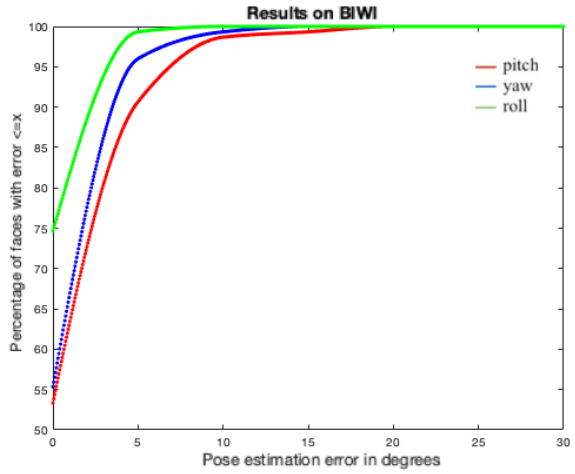

Fig. 7. Errors on BIWI in terms of percentage of tested images.

calculated for each of three degrees of freedom and also an overall MAE of the error along the three axes. Therefore, the search for the perfect model translate into finding the lowest possible MAE value.

The results obtained on Biwi and compared to state-of-the-art methods are reported in Table IV. Almost all the approaches adopted for the comparison are based on neural networks, with the exception of QT PYR and WSM. In this dataset, our approach provides a lower MAE value than all other methods, including Yaw and Roll except for the WSM (only Pitch angular error). In addition, we also compared the features extractor, fractal encoding-based, we propose in FASHE, with another features extractor, to demonstrate the improvements of our descriptor. We used FaceNet [65], a popular and powerful neural network to perform face recognition, to extract a features array of 128 elements from Biwi face images and then proceed as the proposed method (e.g. we substitute, in fact, the FaceNet features array to the Fractal Encoding array, and then we build the model as usual and used Hamming as distance). From the results in table V, it is clear that not every features array can be used to perform head pose estimation and, the particular characteristics of fractal encoding (self-similarity through affinity transformations) highlight the motivations behind FASHE.

The Table VI shows the results on AFLW2000. Considering

TABLE V
COMPARISON OF ANGULAR ERRORS ON BIWI DATASET BETWEEN FASHE AND FACENET

Method	Yaw	Pitch	Roll	MAE
FaceNet [65]	33.82	22.21	16.18	24.07
FASHE	**3.13**	**4.61**	**2.74**	**3.50**

TABLE VI
MEAN ABSOLUTE ERROR OF PITCH, YAW, AND ROLL ANGLES ACROSS DIFFERENT METHODS OVER THE AFLW2000 DATASET. ALL METHODS MARKED WITH (*) DO NOT USE NEURAL NETWORKS OR MACHINE LEARNING APPROACHES. THE (+) SIGN IDENTIFIES THE SYSTEMS THAT USE 3D TECHNIQUES.

Method	Yaw	Pitch	Roll	MAE
QT PYR [12] *	7.6	7.6	7.17	7.45
WSM [13] *	**3.11**	**4.82**	**2.25**	**3.39**
KEPLER [31]	6.45	5.85	8.75	7.01
Multi-Loss ResNet50 [34]	6.470	6.559	5.436	6.155
QuatNet [37]	3.973	5.615	3.92	4.503
Hyperface [39]	7.61	6.13	3.92	5.89
FAN [51] +	6.358	12.277	8.714	9.116
3DDFA [52] +	5.400	8.530	8.250	7.393
FASHE *	4.54	6.42	3.71	4.89

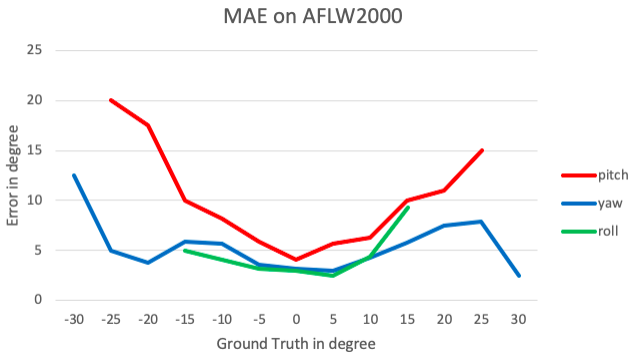

Fig. 8. Errors on AFLW2000 in terms of angular poses.

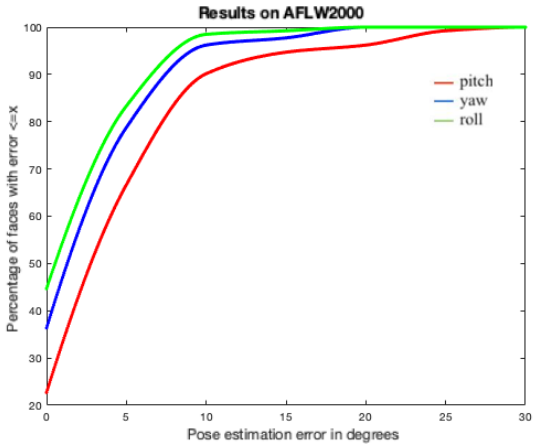

Fig. 9. Errors on AFLW2000 in terms of percentage of tested images.

that the dataset has a not entirely precise annotation of the poses, the results obtained favourably compete with the others. The overall MAE is similar to approaches that use CNNs and, in some cases, it is better. Again, the WSM is the exception.

TABLE VII
MEAN ABSOLUTE ERROR OF PITCH AND YAW ANGLES ACROSS DIFFERENT METHODS OVER THE POINTING'04 DATASET. ALL METHODS MARKED WITH (*) DO NOT USE NEURAL NETWORKS OR MACHINE LEARNING APPROACHES. THE (+) SIGN IDENTIFIES THE SYSTEMS THAT USE 3D TECHNIQUES.

Method	Yaw	Pitch	MAE
WSM [13] *	10.63	**6.34**	8.4
Gourier et al. [15]	12.1	7.3	9.7
hGLLiM [21]	7.93	8.47	8.2
Probabilistic HDR [22]	8.70	8.85	8.775
Stiefelhagen [32]	9.7	9.5	9.6
Kong et al. [50] +	10.98	9.71	10.345
SVR [62]	12.82	11.25	12.035
FASHE *	**6.6**	9	**7.8**

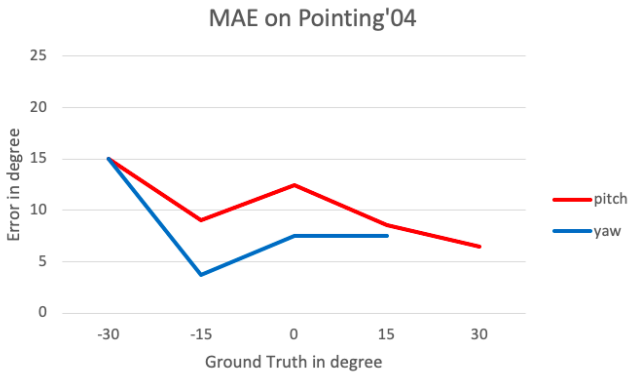

Fig. 10. Errors on Pointing'04 in terms of angular poses.

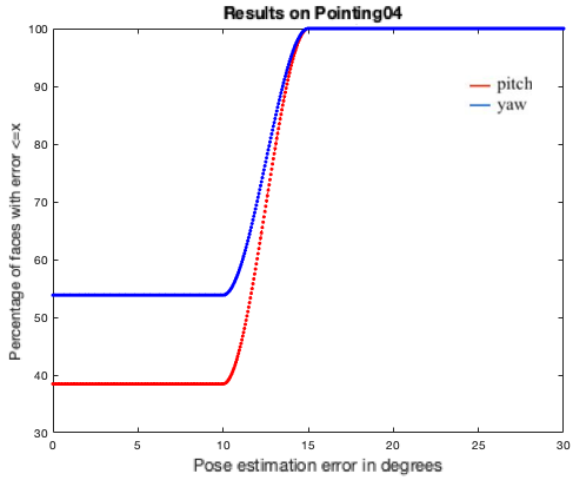

Fig. 11. Errors on Pointing'04 in terms of percentage of tested images.

Table VII shows the results over Pointing'04 dataset. As can be observed on Biwi's experimental results, the proposed approach on fractal encoding provides a lower MAE value than all other methods, including Yaw angles. Roll information are absent from the dataset, therefore none of the state-of-the-art methods include Roll in the experiment results.

Considering the computational time, as can be seen from Table III, if we consider that the faster method at the state of the art is WSM, that has a similar preprocessing to us and spent 0.2 seconds in mean in the core of the method, with our 0.1553 seconds required, we outperform the state of the art. In particular, this computational time is obtained for a fixed partition of range and domain, 4 and 8, respectively, used for each dataset involved. Those values give us a balance between computational time performances and errors performances. It is possible that, paying in computational time, the results over AFLW, that is the only dataset on which we do not outperform the state of the art, could be improved.

Some additional visual results are shown in the lines graphs and histograms graphs. In particular in Figures 6, 8 and 10 can be observed the behaviour of FASHE correlated to head pose angular values on , AFLW2000 and Pointing'04 datasets, respectively. On Biwi, Pitch error increases when angular value is large and positive, whereas Yaw error increases when angular value is large in module, while Roll error is generally low over all angular values. On AFLW2000 both Pitch and Yaw errors increase when angular value is large in module, while Roll error is maximum only for positive large angular values. Finally, on Pointing'04 Pitch error and Yaw error both increase when the pose's angular value is large and negative.

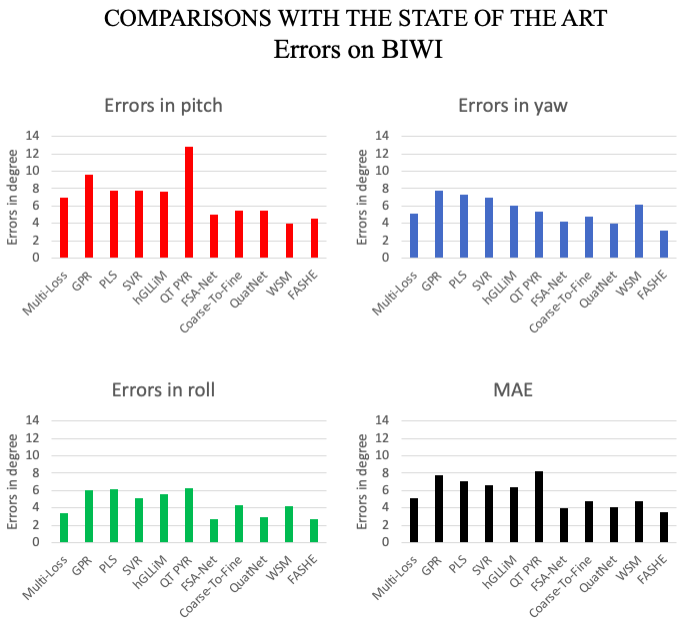

Fig. 12. Comparisons in terms of Errors between state-of-the-art methods on BIWI.

In Figures 7, 9 and 11 can be observed the percentage of images with error below a given angle on Biwi, AFLW2000 and Pointing'04 datasets, respectively. On Biwi, more than a half of images have an error equal to 0°, and less than 5% of images have an error higher than 5°. On AFLW2000, about 30% of images have an error equal to 0°, about 80% of images have an error less than 5°, and less than 5% of images have an error higher than 5°. On Pointing'04, the difference in pitch and yaw is more marked. About 40% of images for pitch and 60% of images for yaw have an error around 0° and there are no images with error higher than 10°.

Figures 12, 13 and 14 graphically represent the comparisons with the state-of-the-art on Biwi, AFLW2000 and Pointing'04 datasets, respectively. Here the errors reported are displayed

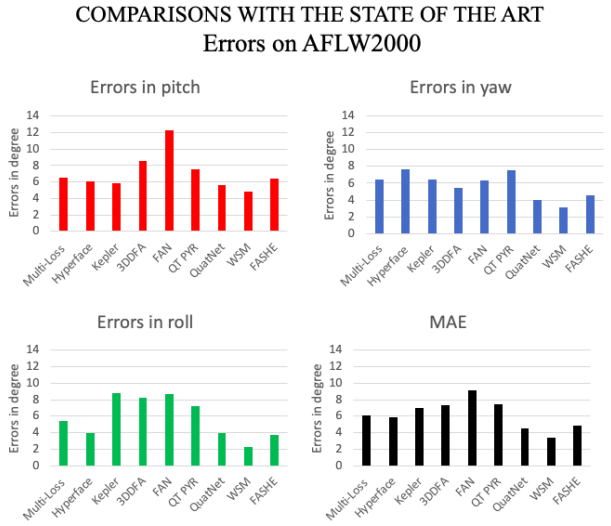

Fig. 13. Comparisons in terms of Errors between state-of-the-art methods on AFLW2000.

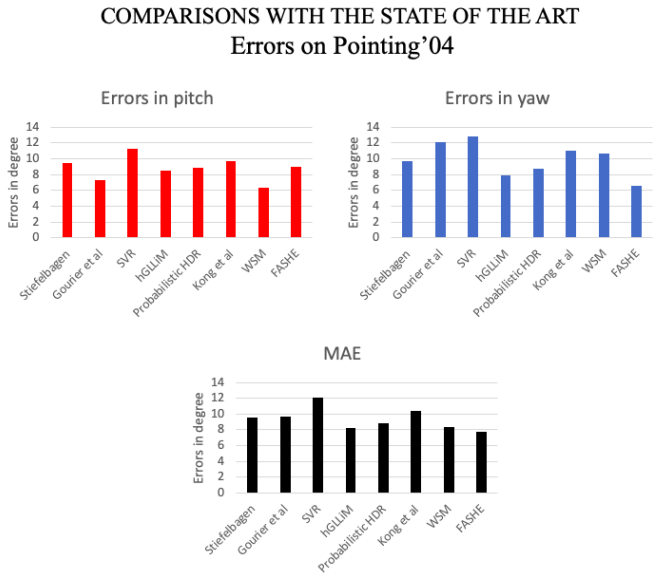

Fig. 14. Comparisons in terms of Errors between state-of-the-art methods on Pointing'04.

on different graphs according to Yaw, Pitch, Roll and MAE, with the exception of Figure 14 where Roll is not present.

Finally, in Figure 15 some experiments on the Gotcha Video Dataset are shown. The Gotcha Dataset [66] is a multiview video dataset containing video from cooperative and non-cooperative subjects. This kind of data, particularly the non-cooperative, are very useful to evaluate the effectiveness of a head pose estimation method to find the (most) frontal face in a video-sequence collected in the wild. According to the results of the experiments, FASHE successfully detected the best approximation to a frontal face pose among the frames available for each subject in the database.

Fig. 15. Experiments on a video sequence on Gotcha. In yellow the most frontal pose detected by our algorithm.

VI. Conclusions

In this paper, FASHE, an improved PIFS-based approach for head pose estimation, is presented. The proposed method relies on contractive transformation of face image auto-similarities to determine a fractal-encoding vector, whose comparison to a reference gallery provide a pose estimate. The method exploits a single reference face image to pre-compute only once the domain blocks required by fractal encoding process. This novel characteristic improves fractal encoding effectiveness as well as its efficiency, since it is not necessary to build the domain blocks for every input image and this is particularly true for video related applications. The fractal code representing the rotational-features of the input face image is therefore matched against a reference array to find the closest approximation according to Hamming distance. The whole process does not involve any training or neural network. Nevertheless, according to the results of the set of experiments conducted on three reference datasets, the reported pose estimation error on AFLW2000 is comparable to that of the best performing methods regardless they are based on machine learning or not. Moreover, on the Biwi and Pointing'04 datasets our method sets a new unprecedented MAE performance, with value as low as 3.50° and 7.80° respectively. These findings support the intuition that not always and not necessarily a learning strategy based on deep CNN architecture represents the most performing approach for complex pattern recognition tasks. Further experiments on

the GOTCHA dataset prove the practical applicability of this approach to the task of face normalization by most-frontal-face frame selection. We are currently considering the use of non-linear transformations to the aim of improving the sensitivity and accuracy in the detection of self-similarity, as well as designing an ad-hoc metric for the distance between IFSs. Another strategy applicable to FASHE to improve its result on heterogeneous dataset as AFLW, is to perform different preprocessing steps in terms of filtering. Since the use of a Gaussian filter on AFLW improved the performances of the method, we can imagine that a move in this direction will make FASHE able to outperform the state of the art also on AFLW. Another possible path to follow could be to evaluate different metrics or classifiers to further improve the results that can be obtained by a fractal encoding method as features extractor for HPE. We are currently investigating the application of FASHE to 3D information, through depth images. This, in fact, completely changes our initial data, and can demonstrate the effectiveness and generalization of the fractal features.

REFERENCES

[1] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 607–626, Jan 2009.

[2] B. Czupryński and A. Strupczewski, "High accuracy head pose tracking survey," in *International Conference on Active Media Technology*, Aug 2014.

[3] C. Ding, C. Xu, and D. Tao, "Multi-task pose-invariant face recognition," *IEEE Transactions on Image Processing*, vol. 24, no. 3, pp. 980–993, 2015.

[4] J. Chen, S. Nie, and Q. Ji, "Data-free prior model for upper body pose estimation and tracking," *IEEE Transactions on Image Processing*, vol. 22, no. 12, p. 4627–4639, 2018.

[5] R. Valenti, N. Sebe, and T. Gevers, "Combining head pose and eye location information for gaze estimation," *IEEE Transactions on Image Processing*, vol. 22, no. 2, p. 802–815, 2012.

[6] Y. Wang, H. Yu, J. Dong, M. Jian, and H. Liu, "Head pose estimation via probabilistic high-dimensional regression," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 2831–2835.

[7] Y. Cho, K. Yoon, and D. Tao, "Pamm: Pose-aware multi-shot matching for improving person re-identification," *IEEE Transactions on Image Processing*, vol. 27, no. 8, p. 3739–3752, 2018.

[8] Y. Yu, K. A. F. Mora, and J.-M. Odobez, "Robust and accurate 3d head pose estimation through 3dmm and online head model reconstruction," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. Ieee, 2017, pp. 711–718.

[9] ——, "Headfusion: 360° head pose tracking combining 3d morphable model and 3d reconstruction," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 11, pp. 2653–2667, 2018.

[10] G. Borghi, M. Venturelli, R. Vezzani, and R. Cucchiara, "Poseidon: Face-from-depth for driver pose estimation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5494–5503.

[11] G. Borghi, M. Fabbri, R. Vezzani, S. Calderara, and R. Cucchiara, "Face-from-depth for head pose estimation on depth images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 3, pp. 596–609, 2020.

[12] A. F. Abate, P. Barra, C. Bisogni, M. Nappi, and S. Ricciardi, "Near real-time three axis head pose estimation without training," *IEEE Access*, vol. 7, pp. 64 256–64 265, 2019.

[13] P. Barra, S. Barra, C. Bisogni, M. De Marsico, and M. Nappi, "Web-shaped model for head pose estimation: an approach for best exemplar selection," *IEEE Transactions on Image Processing*, 2020.

[14] M. Osadchy, Y. L. Cun, and M. Miller, "Synergistic face detection and pose estimation with energy-based models," *Journal of Machine Learning Research*, pp. 1197–1215, 2007.

[15] N. Gourier, J. Maisonnasse, D. Hall, and J. L. Crowley, "Head pose estimation on low resolution images," in *International Evaluation Workshop on Classification of Events, Activities and Relationships*. Springer, 2006, pp. 270–280.

[16] J. Aghajanian and S. Prince, "Face pose estimation in uncontrolled environments," in *British Machine Vision Conference (BMVC)*, 2009, pp. 1568–1572.

[17] B. M. Smith, J. Brandt, J. Lin, and L. Zhang, "Nonparametric context modeling of local appearance for pose- and expression-robust facial landmark localization," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014, pp. 1741–1748.

[18] K. Pawelczyk and M. Kawulok, "Head pose estimation relying on appearance-based nose region analysis," *Computer Vision and Graphics*, pp. 510–517, 2014.

[19] D. Lee, M. Yang, S. Oh, M. Jian, and H. Liu, "Fast and accurate head pose estimation via random projection forests," in *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2015, pp. 1958–1966.

[20] Y. Liu, Z. Xie, X. Yuan, and J. Chen, "Wusong: Multi-level structured hybrid forest for joint head detection and pose estimation," *Neurocomputing*, vol. 272, pp. 206–215, 2017.

[21] V. Drouard, R. Horaud, A. Deleforge, S. Ba, and G. Evangelidis, "Robust head-pose estimation based on partially-latent mixture of linear regressions," *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1428–1440, 2017.

[22] V. Drouard, S. Ba, G. Evangelidis, A. Deleforge, and R. Horaud, "Head pose estimation via probabilistic high-dimensional regression," in *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 4624–4628.

[23] X. Peng, J. Huang, Q. Hu, S. Zhang, A. Elgammal, and D. Metaxas, "From circle to 3-sphere: Head pose estimation by instance parameterization," *Computer Vision and Image Understanding*, vol. 136, pp. 92–102, 2015.

[24] H. Yang, W. Mou, Y. Zhang, I. Patras, H. Gunes, and P. Robinson, "Face alignment assisted by head pose estimation," in *British Machine Vision Conference (BMVC)*, 2015.

[25] X. Xu and I. A. Kakadiaris, "Joint head pose estimation and face alignment framework using global and local cnn features," in *IEEE International Conference on Face Gesture Recognition (FG 2017)*. IEEE, 2017, pp. 642–649.

[26] H. Proena, J. C. Neves, S. Barra, T. Marques, and J. C. Moreno, "Joint head pose/soft label estimation for human recognitionin-the-wild," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 12, pp. 2444–2456, 2016.

[27] Y. Yan, E. Ricci, R. Subramanian, G. Liu, O. Lanz, and N. Sebe, "A multi-task learning framework for head pose estimation under target motion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 6, pp. 1070–1083, 2016.

[28] H. Liao, S. Lu, and D. Wang, "Tied factor analysis for unconstrained face pose classification," *Optik - International Journal for Light and Electron Optics*, vol. 127, no. 23, pp. 11 553–11 566, 2016.

[29] J. Chen, J. Wu, K. Richter, J. Konrad, and P. Ishwar, "Estimating head pose orientation using extremely low-resolution images," in *IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI)*. IEEE, 2016, pp. 65–68.

[30] M. Patacchiola and A. Cangelosi, "Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods," *Pattern Recognition*, vol. 71, pp. 132–143, 2017.

[31] A. Kumar, A. Alavi, and R. Chellappa, "Kepler: Keypoint and pose estimation of unconstrained faces by learning efficient h-cnn regressors," in *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, May 2017, pp. 258–265.

[32] R. Stiefelhagen, "Estimating head pose with neural networks-results on the pointing04 icpr workshop evaluation data," in *Proc. Pointing 2004 Workshop: Visual Observation of Deictic Gestures*, vol. 1, no. 5, 2004, pp. 21–24.

[33] X. Liu, W. Liang, Y. Wang, S. Li, and M. Pei, "3d head pose estimation with convolutional neural network trained on synthetic images," in *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 1289–1293.

[34] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[35] B. Ahn, D. G. Choi, J. Park, and S. Kweon, "Real-time head pose estimation using multi-task deep neural network," *Robotics and Autonomous Systems*, vol. 103, pp. 1–12, 2018.

[36] H. Wu, K. Zhang, and G. PaTianrk, "Simultaneous face detection and pose estimation using convolutional neural network cascade," *IEEE Access*, vol. 6, pp. 49 563–49 575, 2018.

[37] H.-W. Hsu, T.-Y. Wu, S. Wan, W. H. Wong, and C.-Y. Lee, "Quatnet: Quaternion-based head pose estimation with multiregression loss," *IEEE Transactions on Multimedia*, vol. 21, no. 4, pp. 1035–1046, 2018.

[38] M. Raza, Z. Chen, S. U. Rehman, P. Wang, and P. Bao, "Appearance based pedestrians: head pose and body orientation estimation using deep learning," *Neurocomputing*, vol. 272, pp. 647–659, 2018.
[39] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 121–135, Jan 2019.
[40] T.-Y. Yang, Y.-T. Chen, Y.-Y. Lin, and Y.-Y. Chuang, "Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1087–1096.
[41] Y. Wang, W. Liang, J. Shen, Y. Jia, and L.-F. Yu, "A deep coarse-to-fine network for head pose estimation from synthetic data," *Pattern Recognition*, vol. 94, pp. 196–206, 2019.
[42] M. D. Breitenstein, D. Kuettel, T. Weise, L. Van Gool, and H. Pfister, "Real-time face pose estimation from single range images," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2008, pp. 1–8.
[43] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool, "Random forests for real time 3d face analysis," *International Journal of Computer Vision*, 2013.
[44] B. Wang, W. Liang, Y. Wang, and Y. Liang, "Head pose estimation with combined 2d sift and 3d hog features," in *IEEE International Conference on Image and Graphics (ICIG)*. IEEE, 2013, pp. 650–655.
[45] M. Lia and W. Pedrycz, "A central profile-based 3d face pose estimation," *Pattern Recognition*, vol. 47, no. 2, pp. 525–534, 2014.
[46] C. Papazov, T. K. Marks, and M. Jones, "Realtime 3d head pose and facial landmark estimation from depth images using triangular surface patch features," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015, pp. 4722–4730.
[47] A. Saeed and A. Al-Hamadi, "Boosted human head pose estimation using kinect camera," in *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 1752–1756.
[48] J. Darby, M. B. Sánchez, P. B. Butler, and I. D. Loram, "An evaluation of 3d head pose estimation using the microsoft kinect v2," *Pattern Recognition*, vol. 48, pp. 3649–3657, 2016.
[49] G. P. Meyer, S. Gupta, I. Frosio, D. Reddy, and J. Kautz, "Robust model-based 3d head pose estimation," in *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2015, pp. 3649–3657.
[50] S. G. Kong and R. O. Mbouna, "Head pose estimation from a 2d face image using 3d face morphing with depth parameters," *IEEE Transactions on Image Processing*, vol. 24, no. 6, pp. 1801–1808, 2015.
[51] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks)," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
[52] X. Zhu, X. Liu, Z. Lei, and S. Z. Li, "Face alignment in full pose range: A 3d total solution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 78–92, Jan 2017.
[53] A. Jacquin, "A fractal of iterated markov operators with applications to digital image coding," 1989.
[54] R. Distasi, M. Nappi, and D. Riccio, "A range/domain approximation error-based approach for fractal image compression," *IEEE Transactions on Image Processing*, vol. 15, no. 1, pp. 89–97, 2005.
[55] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.
[56] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1867–1874.
[57] R. W. Hamming, "Error detecting and error correcting codes," *The Bell system technical journal*, vol. 29, no. 2, pp. 147–160, 1950.
[58] L. G. Shapiro and G. C. Stockman, "Computer vision," 2001.
[59] G. Fanelli, T. Weise, J. Gall, and L. Van Gool, "Real time head pose estimation from consumer depth cameras," in *Joint Pattern Recognition Symposium*, 2011.
[60] X. Zhu, Z. Lei, H. Shi, X. Liu, and S. Z. Li, "Face alignment across large poses: A 3d solution," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
[61] N. Gourier, D. Hall, and J. L. Crowley, "Estimating face orientation from robust detection of salient facial features," in *ICPR International Workshop on Visual Observation of Deictic Gestures*. Citeseer, 2004.
[62] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, Aug 2004. [Online]. Available: https://doi.org/10.1023/B:STCO.0000035301.49549.88
[63] C. E. Rasmussen, *Gaussian Processes in Machine Learning*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 63–71. [Online]. Available: https://doi.org/10.1007/978-3-540-28650-9_4
[64] H. Abdi, "Partial least square regression (pls regression)," *Encyclopedia for research methods for the social sciences*, vol. 6, no. 4, pp. 792–795, 2003.
[65] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815–823.
[66] P. Barra, C. Bisogni, M. Nappi, D. Freire-Obregón, and M. Castrillón-Santana, "Gotcha-i: A multiview human videos dataset," in *Security in Computing and Communications*, S. M. Thampi, G. Martinez Perez, R. Ko, and D. B. Rawat, Eds. Singapore: Springer Singapore, 2020, pp. 213–224.

Carmen Bisogni (M'18) received the B.S. degree and M.S. degree (cum Laude) in Mathematics from University of Salerno in 2015 and 2017, respectively. She is currently pursuing the Ph.D. degree in Computer Science at Biometric and Image Processing Laboratory (BIPLAB) at University of Salerno, Italy. Her research interests include applied mathematics for Machine Learning, Biometrics, Image Processing and Statistical Analysis. Dr. Bisogni is member of IEEE and GIRPR/IAPR.

Michele Nappi (M'11 SM'17) received the Laurea degree (cum laude) in computer science from the University of Salerno, Italy, in 1991, the M.Sc. degree in information and communication technology from I.I.A.S.S. "E.R. Caianiello", Vietri sul Mare, Salerno, and the Ph.D. degree in applied mathematics and computer science from the University of Padova, Italy. He is currently a Full Professor of computer science with the University of Salerno. He is also a Team Leader of the Biometric and Image Processing Lab (BIPLAB). His research interests include multibiometric systems, pattern recognition, image processing, compression and indexing, multimedia databases, human-computer interaction, and VR/AR. He has coauthored more than 190 papers in international conference, peer review journals and book chapters in these fields. He was a member of IAPR. He received several international awards for scientific and research activities. He was the President of the Italian Chapter of the IEEE Biometrics Council, from 2015 to 2017.

Chiara Pero (M'19) received the B.S. and M.S. (cum Laude) degrees in Computer Science from the University of Salerno, Italy, in 2016 and 2018, respectively. She is currently pursuing the Ph.D. degree in Computer Science with the Biometric and Image Processing Laboratory (BIPLAB), University of Salerno. Her research interests include machine learning technics in facial recognition, image processing, behavioral profiling and activity recognition. Dr. Pero is member of IEEE.

Stefano Ricciardi (M'12) was born in Naples, Italy. He received the BSc in Computer Science, the MSc degree in Informatics and the PhD degree in Sciences and Technologies of Information, Complex Systems and Environment from the University of Salerno. Main research interests include biometry, virtual and augmented/mixed reality, haptics systems and human-computer interaction. He has been co-founder/owner of a videogame development company, is currently an Assistant Professor at the Department of Biosciences of the University of Molise - Italy and served as external expert of the of the European Commission's Research Executive Agency for the Horizon-2020 research and innovation programme. He served as reviewers for several international journals, is member of IEEE and GIRPR/IAPR and co-authored more than eighty research papers including international journals, book chapters and conference proceedings.