Abstract

Head pose estimation is not only a crucial challenge for many real-world applications, such as driver attention detection analysis, but it represents an interesting strategy to support biometric frameworks as well. This paper aims to estimate head pose from a single image by applying notions of network curvature. In the real world, many complex networks have groups of nodes that are well connected to each other with significant functional roles. Similarly, the interactions of facial landmarks can be represented as complex dynamic systems modeled by weighted graphs. The functionality of such a system is therefore intrinsically linked to the topology and geometry of the underlying graph. In this work, using the geometric notion of Ollivier-Ricci curvature (ORC) on weighted graphs as input to the XGBoost regression model, we show that the intrinsic geometric basis of ORC offers a natural approach to discovering underlying common structure within a pool of poses. Experiments on the BIWI, AFLW2000 and Pointing’04 datasets show that the ORC-XGB method performs well compared to state-of-the-art methods, both landmark-based and image-only.

1. Introduction

Head pose estimation (HPE) is typically described by three angles (pitch, yaw, roll) that express face orientation with respect to a system of Cartesian axes centered on the head. Several applications benefit from understanding the face orientation in 3D space: face alignment, driver-assistance systems, and human-robot interaction just to name a few [3] [22]. The study of HPE is a subset of the broader field of biometrics, in that it can be applied, for example, in the field of behavioral biometrics to study the cooperative behavior of a subject when faced with security cameras [13]. It is possible to divide the different approaches used to solve this problem into two different categories: those that implement methods based on landmarks, and those that process only the raw images. Head pose estimation based on facial landmarks is quite attractive, both because many advanced sensors now integrate landmark detection [20], and because this approach is generally more robust to occlusion since it establishes a correspondence between 2D images and 3D facial models. The main limitation of this strategy is the difficulty of landmark extraction, especially for extreme poses. The idea at the foundation of this paper is to use the Ollivier-Ricci curvature [28] (ORC) as an aid to estimate head pose by evaluating the optimal, minimum movement cost of face landmarks. The regions around landmarks are elements whose geometric relationships can be captured by a suitably chosen metric space or an embedding. ORC is a good example of such a metric space.

We propose a connection graph that takes selected facial landmarks as vertices, on which ORC can be calculated and used as a geometric descriptor. The Ollivier-Ricci curvature distribution is shown to be different from graph to graph and can act as a graph fingerprint or graph kernel. From these premises, the question to be addressed is: Can the ORC curvature help us in solving the challenging problem of head pose estimation? To answer this, we implemented a system whose essential parts are shown in Figure 1. Given an input image, the proposed system constructs a graph based on 468 facial landmarks preliminarily detected. It then regresses the head pose angle using Ollivier-Ricci curvature information for pairs of vertices. There are no other widely known methods for estimating head pose that use this geometric descriptor. The performance of the proposed method is tested on three datasets that are well known in the field. The accuracy is compared with the state of the art by replicating the same conditions and protocols so that the comparison is as meaningful as possible. Whether considering
works that use landmarks or works that do not, the performance obtained compares well; in fact, it even exceeds current accuracies.

1.1. Contributions

The main contributions of the present work are listed below.

• A discretization of the Ricci curvature is used for head pose estimation for the first time.

• ORC is shown to be able to capture and incorporate multiple aspects of the pose, yielding results that are similar, and in some cases better, than the state of the art.

2. Related work

2.1. Application fields of Ricci Curvature

A notion of Ricci curvature (Ollivier-Ricci) based on optimal transportation theory has been proposed by Yann Ollivier in [28].

Recently, this geometric descriptor has become a popular topic and has been applied to various fields: 1) Medicine. In [30], the authors adopt the notion of Ollivier-Ricci curvature to distinguish between cancerous and healthy samples. In particular, they analyze gene co-expression networks derived from large-scale genomic studies of cancer. Cancerous networks show a higher curvature than their normal counterparts. 2) Biology. In the work [36], the goal is the prediction of protein-ligand binding affinity, which is a crucial aspect in drug design. Here, molecular structures and interactions are modeled as graphs on which the Ollivier-Ricci curvature is calculated. This information is then used as a descriptor for a Machine Learning (ML) model. 3) Finance. In [31], the authors use the Ollivier-Ricci curvature as an economic indicator for systemic risk that captures local to global system-level fragility—not only in financial markets, but also in broader economic networks that include the banking ecosystem.

2.2. Head pose estimation

HPE has been a popular subject of study over the last twenty years [18]. It is possible to divide the different approaches used for solving this problem into two different categories: those that implement landmark-based methods, and those that do not use landmarks and only process simple images.

Landmark-based methods. In [37], the authors construct a landmark-connection graph and use Graph Convolutional Networks, to which they add two more models to improve performance. A different, more geometric idea on how to exploit landmarks can be found in [5]: the landmarks are placed in a spider’s web that overlaps the face. The same idea is developed in [1], where the feature vector is input into a regression algorithm.

Image-based methods. The leitmotif of these methods is to obtain a pose estimate from a simple image, without extracting any landmarks. In [7], this approach is adopted without even applying ML techniques. The self-similarities present in a facial image are used to determine a fractal coding vector. To determine pitch, yaw, and roll, this vector is compared to a reference gallery of poses. A cascade ML approach is used instead in [35]. The authors define four categories of poses and use a classifier to sort images into categories. A regression model refines the estimate. The same approach, i.e., to work on a single image and then apply regression, can be found also in [2], [38], [11], and [34]. In [2], combining self-similar structures utilized for fractal image compression yields the feature vector. By contrast, the idea in [38] is to apply a regression model to intermediate features deemed significant after learning the fine-grained structure mapping for spatial clustering of features before aggregation. Vo et al. [34] use the global features extracted from a histogram of oriented gradients in a multi-stacked autoencoders neural network. In [11], face
bounding boxes are used to extract high-dimensional feature vectors; the vectors are then fed into a mixture of linear regressions followed by Bayes inversion.

Deep networks have been used, too: in [6], an attention mechanism captures subtle changes in images. The idea in [17] is to create fully convolutional neural networks without fully connected layers, starting from a Bernoulli heat map. The authors of [16] also follow the deep learning paradigm, but rather than working on single images, dynamic facial analysis is performed by working on videos. There are several more works that use deep learning to tackle HPE, such as [24] and [4], but a full discussion of the vast available literature is out of the scope of this paper.

3. Methods and materials

All the graphs mentioned in this paper are undirected. A graph \( G = (V, E) \) is a finite nonempty set \( V \) of objects called vertices or nodes, together with a possibly empty set \( E \) of couples of elements of \( V \) called edges. Once we have a fully connected edge weighted graph \((G, W)\), where \( W : E \to [0, \infty) \) is a nonnegative real function on the edges, called their weight, we can induce a metric \( d(\cdot, \cdot) \) on \( V \times V \) by considering the minimum weight path between any two vertices:

\[
d(v_1, v_2) = \min \sum_{i=0}^{n} W(v_{k_i}, v_{k_{i+1}}),
\]

where the minimum is obtained by considering all edge paths from \( v_1 \) to \( v_2 \)—that is, over all choices of \( n \) and \( k_0, \ldots, k_n \) such that \( v_{k_0} = v_1, v_{k_n} = v_2 \), and \( (v_{k_i}, v_{k_{i+1}}) \in E \) for each \( i \).

In this paper, the graph nodes are the facial landmarks extracted by MediaPipe [14]. The framework detects 468 facial landmarks that are arranged in fixed quads and represented by their Euclidean coordinates. The most natural choice for the weight function is therefore a measure of their geometric distance. In particular, the choice fell on the Manhattan distance \( d_M \).

3.1. Ollivier-Ricci curvature for weighted graph

**Definition 3.1 (Ollivier’s Ricci curvature)** Given an \( n \)-dimensional Riemannian manifold \((M, g)\) with Riemannian volume measure \( \mu \), for each \( x, y \in M \) the Ollivier’s Ricci curvature along the shortest path \( xy \) is:

\[
k(x, y) = 1 - \frac{W(m_x, m_y)}{d(x, y)}
\]

where, fixed a value \( \epsilon > 0 \), \( m_x = \rho(B(x, \epsilon))/(\mu(B(x, \epsilon))) \), \( B(x, \epsilon) \) is the ball of radius \( \epsilon \) at \( x \), and \( W(m_x, m_y) \) is the Wasserstein distance with respect to the cost/distance function \( d(x, y) \).

This definition highlights the correlation between Ricci curvature and the optimal transport problem [19]. Ollivier’s definition of discrete Ricci curvature can be adapted to weighted graphs. In simple terms, given an edge, the Ollivier-Ricci curvature is computed by defining a probability distribution over each of the neighborhoods of the two nodes at its ends and then computing edge weight difference and Wasserstein distance difference between the two distributions. In a weighted graph \((G, W)\), we consider the associated distance function \( d \) as defined in Eq. (1). To calculate the Ollivier-Ricci curvature, for each vertex we must first define the associated probability measures. The probability distribution defined by Lin et al. [23] is extended and generalized in the work of [26] to further account for edge weights. Leveraging concepts from classical differential geometry, they construct the probability measure as follows. Considering a network as a discretization of a smooth manifold, a Riemannian metric and a Riemannian distance should correspond to edge weight and distance. Given an edge weighted graph \((G, W)\) whose associated induced distance in the sense of Eq. (1) is \( d \), the associated normalized probability measure is defined as follows:

\[
m^\alpha_p(y) = \begin{cases} 
\alpha & \text{if } y = x \\
\frac{1}{C} \cdot \exp(-d(x, y)^\alpha) & \text{if } y \in N(x) \\
0 & \text{otherwise},
\end{cases}
\]

where \( N(x) \) is the set of \( x \)'s neighbors, while \( \alpha \in [0, 1] \) indicates the share of mass that should be left on the original node. For example, if \( \alpha = 0.6 \), it means that 0.6 mass units will stay in \( x \), while 0.4 units should evenly spread into \( x \)'s neighbors. The power parameter \( r \) is the measure of the weight that we give to the neighbor \( x_i \) of \( x \) with respect to the distance \( d(x,x_i) \). The probability measure is uniform on all neighbors of \( x \) when \( r = 0 \), as suggested in [23]. Instead, for large values of \( r \), the neighbors that are far away from \( x \) are strongly discounted. It turns out that in most cases, \( r = 1 \) and \( r = 2 \) perform best for community detection task. The quantity \( C = \sum_{x_i \in N(x)} \exp(-d(x, x_i)^\alpha) \) is a normalizing factor. More theoretical details can be found in [9].

4. Implementation details

This section illustrates the steps for implementing the above method.

**Input data.** Given an image as input, we use MediaPipe as the landmark face detector [14]. For each face, 466 landmarks are detected.

**Data normalization.** Given 466 landmark points \( p_i = (x_i, y_i) \), we normalize all points to the same scale by

\[
\hat{p}_i = \frac{p_i - p_0}{p_{\text{max}} - p_{\text{min}}}
\]
where \( p_0 \) is a fixed vertex, while \( p_{\text{max}} \) and \( p_{\text{min}} \) are the maximum and minimum values in the graph.

**Graph construction.** The present work is directed at understanding if and how Ricci curvature can provide an effective descriptor for pose estimation. This is how the graphs were constructed for this experiment; the improvement of the graph structure and its optimization could be the subject of a whole new study.

Given 466 landmark locations, we create a graph by using the landmarks as vertices, then connecting them according to spatial proximity and empirical considerations. Once a connection strategy for the vertices is laid out, the edges are assigned individual weights according to the Manhattan distance between the relevant two vertices. The configuration chosen is the one that performed best in the initial benchmark.

At first, a preliminary study was carried out on several graphs. In these initial tests, pose estimation was obtained through the simple use of different metrics: each test image has been associated to the image coming closest to it according to one specific metric chosen from a large pool of possible distance functions, including the ones made available by the `scipy.spatial.distance` Python module [33]. This preliminary phase also included a test using Delauney Triangulation (DT) [21]. However, despite the supposed rigorous sophistication of this technique, its actual performance was poor. This result can probably be ascribed to the fact that as the points (and thus the graph nodes) get denser, the unique biometric features of the individual subject overtake pose related information. Fig. 1 depicts the configuration chosen after these preliminary empirical tests.

The issue with DT-based graphs vs grid-based ones such as ours is that in the former, all edges have positive ORC values. Positive edge curvature means neighbor nodes are well-connected and do not exhibit the bottleneck effect so crucial for pinpointing optimal flow [32]—which our method depends upon. From a qualitative, intuitive point of view, it could be said that DT makes the graph flow problem “too easy” and deprives the algorithm of discriminating power. Negative-curvature edges act as bridges or transitions between neighboring subgraphs (which in turn translate into face areas). This intuitive notion has been confirmed by a series of preliminary experiments where DT-based graphs yielded significantly worse results when used for pose estimation.

**Template creation.** Once a graph has been produced for each image, its Ollivier-Ricci curvature is computed. Every pose is associated with a vector of 446 components, representing the curvature between graph nodes. Referring to Eq. (3), we have \( \alpha = 0.5 \) and \( r = 2 \). These values resemble the process of heat diffusion. For a proper comparison with the latest state-of-the-art work on pose estimation, we evaluated our method on three well known and challenging datasets: AFLW2000 [40], BIWI [12], and Pointing’04 [15]. AFLW2000 consists of 2000 photos of faces, mostly RGB, collected from the social network Flickr. The samples in AFLW include a wide variety of poses, ethnic traits, ages, facial expressions and environmental conditions. The BIWI dataset consists of 24 videos of 20 subjects in a controlled environment. The total number of frames in the dataset exceeds 10000. The Pointing’04 dataset contains faces of 15 subjects acquired in the wild, for a total of 2790 images labeled in steps of 15° and 30° in yaw and pitch between consecutive poses. There is no roll information. Three different experimental protocols were applied depending on the dataset under consideration.

1. **Protocol 1** “One left out” strategy, using only one subject for testing, and the others as a model for carrying out the comparisons. This protocol has only been applied to BIWI, since it is the only dataset with information about the subjects. Overall, 20 experiments were carried out, rotating the test subject. In the same guise as described in other works, one more experiment was performed using a subset of only 10 subjects [7] [2].

2. **Protocol 2** In this case, too, only the dataset BIWI has been used. A part of the videos were used for training, and the others for testing. This protocol has been employed by several pose estimation methods with different modalities, such as RGB, depth, and time. Our method, however, only considers a single RGB frame [38] [6]. In particular, many researchers select 70% of the available videos for training and the remaining 30% (namely, 8 videos) for testing.

3. **Protocol 3** In this case, too, a split is applied on the data: training 70%, testing 30% [1].

Some works excluded extreme poses—that is, those with angles exceeding pitch \([-30°, 30°]\), yaw \([-45°, 45°]\) and roll \([-20°, 20°]\). Therefore, for the sake of fair comparison, an additional experiment has been added to our suite, excluding from the template the poses outside these ranges. Additionally, for Protocol 1, we performed some experiments limited to a metric approach, without the use of ML. As already sketched above while discussing graph construction, for each subject a pose was chosen according to the minimum value of one distance function picked from a pool. The best overall performer, over varying graph configurations and graph weight functions, has always been the Jensen-Shannon distance [27].

**Regression model.** After defining the protocol, a regression model has been applied to obtain an estimate of the pose in terms of the three axes pitch, yaw and roll. We chose the Extreme Gradient Boosting Regressor (XGB), a novel algorithm recently introduced by Chen and Guestrin [10]. This implementation of the gradient boosting algorithm is
both efficient and effective. Its strengths include the ability to deal with overfitting, a high robustness, and a good degree of flexibility. In an earlier preliminary phase, other regressors have also been tested using their default parameter values: Decision Tree, Huber, TheilSen and others. However, their performance on this specific task was not as good. As the focus was on assessing the descriptor for pose estimation, a different choice of regressor with exploration of the parameter space has not been pursued further.

5. Discussion and comparison with the state of the art

This section illustrates an experimental evaluation of the proposed head-pose estimation strategy. To assess our results, we report the mean absolute error (MAE) between the estimate and the ground truth of each angle in the test set. Our method is compared to state-of-the-art methods, both landmark-based and image-based.

For the sake of meaningful comparison, the experiments presented here only include works that use the same protocols. Furthermore, when other works limited their experiments to a subset of the poses or adopted a different ratio of training vs. test data, we mimicked the same choice of data and ratios.

Before commenting on our results, a clarification is in order. In recent years, particularly in the Deep Learning field, the 300W-LP dataset [40] has been quite popular as a testbed. A few examples of papers using 300W-LP are [6], [38], and [17]. A salient feature of this dataset is that, along each original image, there are several synthesized poses derived from it. Data of this kind, however, are not suitable for processing with our method, since it relies on the analysis of the relative positions of face landmarks. Synthetic deformations distort the very foundation for methods based on the correct calculation of distances and curvatures. For these reasons, we excluded 300W-LP from our experiments.

Table 1 shows the results on the BIWI dataset using the 3 protocols. In all cases, the biometric component and subject peculiarities influence the performance when considering all subjects (as opposed to one or just a few). This can be noticed comparing Protocol 1 vs. Protocol 3: there is a difference of ≈2-3° between the data in, say, rows 9 and 27. In the case of Protocol 1, the subject whose pose is to be estimated is not in the training set; on the contrary, with Protocol 3 the same subject can be in both the training and the testing sets. When this happens, the method seems to ‘recognize’ the subject and associate the picture with the picture of the same subject in the closest pose.

Comparing Protocol 1 vs Protocol 2, neither of them has images of the same subject in the training set and in the test set. The difference is that in Protocol 2 the training set contains fewer videos. The MAE variation between the two protocols is around 1°. This shows that a larger number of

| Method                  | Protocol 1: one left out technique | Protocol 2: split on video | Protocol 3: split on the datasets images |
|-------------------------|-----------------------------------|---------------------------|------------------------------------------|
| Method                  | Pitch    | Yaw    | Roll   | MAE    | Pitch    | Yaw    | Roll   | MAE    | Pitch    | Yaw    | Roll   | MAE    |
| 1 HP[H]FS-LR [2]        | 5.46     | 6.59   | 3.8    | 5.28   | 2.31     | 2.30   | 1.76   | 2.45   | 6.23     | 4.05   | 3.30   | 4.52   |
| 2 HP[H]FS [2]           | 6.23     | 4.05   | 3.30   | 4.52   | 4.61     | 3.13   | 2.74   | 3.50   | 4.61     | 3.13   | 2.74   | 3.50   |
| 3 FASHE [7]             | 4.61     | 3.13   | 2.74   | 3.50   | 4.61     | 3.13   | 2.74   | 3.50   | 4.61     | 3.13   | 2.74   | 3.50   |
| 4 ORC_XGB               | 3.31     | 2.30   | 1.76   | 2.45   | 4.06     | 3.07   | 2.55   | 3.23   | 5.36     | 3.64   | 3.53   | 4.18   |
| 5 ORC [2]               | 4.73     | 2.79   | 2.82   | 3.44   | 4.06     | 3.07   | 2.55   | 3.23   | 5.36     | 3.64   | 3.53   | 4.18   |
| 6 ORC_XGB               | 4.06     | 3.07   | 2.55   | 3.23   | 5.36     | 3.64   | 3.53   | 4.18   | 5.36     | 3.64   | 3.53   | 4.18   |
| 7 ORC [2]               | 4.06     | 3.07   | 2.55   | 3.23   | 5.36     | 3.64   | 3.53   | 4.18   | 5.36     | 3.64   | 3.53   | 4.18   |
| 8 hGLLiM [11]          | 7.65     | 6.06   | 5.62   | 6.44   | 6.23     | 4.05   | 3.30   | 4.52   | 6.23     | 4.05   | 3.30   | 4.52   |
| 9 ORC_XGB              | 4.61     | 3.71   | 3.61   | 3.98   | 6.31     | 4.83   | 4.78   | 5.31   | 6.31     | 4.83   | 4.78   | 5.31   |

Table 1: Performance comparison with state-of-the-art results on the BIWI dataset for the 3 protocols. ‘*’: only 10 subjects; ‘♣’: without extreme poses; ‘§’: training/testing ratio 80/20; ‘*: temporal information; ‘♪’: depth information; ‘|’: without extreme poses; ‘¥’: without ML.

subjects in the training set makes the system more robust and reliable. As for Protocol 1, there are several papers in the literature that work on only 10 subjects out of the 20 present in the dataset, excluding the most extreme poses.

With regards to experiments with a reduced, 10-subject dataset, the proposed ORC_XGB method produces results aligned with, or slightly better than, the current alternatives. With the same number of subjects in the dataset, including the more extreme poses, our method equals or exceeds the accuracy obtained by others. This applies even to alternatives that exclude extreme poses in their tests [2]. See for in-
Figure 2 illustrates the error on BIWI, Protocol 1, in terms of angular poses for pitch, yaw and roll. For some ranges, especially the extreme ones, the error grows noticeably. Some investigation shows that the main reason is the detector: when the error is larger, the landmarks themselves are off. This suggests that, if the landmark detection process were performing correctly, the overall performance could be significantly better. Even with the undermining factor of a faulty detection process, the final results are mostly acceptable. This consideration reinforces the conclusion that Ollivier-Ricci curvature works well as a descriptor for pose estimation.

The results in Table 1 show that the method behaves well with Protocol 2, too. The MAE drops well below 3°. The results on the pitch axis are somewhat worse than those shown in [16], although this could be at least partially due to their working with videos: the temporal information that can be extrapolated probably helps to track head movements, improving the accuracy. Focusing on rows 17 and 18, the work in [37] performs better than ORC_XGB, particularly on the pitch axis. Moreover, when the base method (EVA_CGN vanilla) is integrated by an adaptive channel attention module and a densely connected architecture, results improve on the yaw axis as well.

Protocol 3 is perhaps the least interesting in this context, since there is significant interference introduced by individual traits vs. pose differences. The performance of ORC_XGB still ranks at the top, with the best error figures. When excluding poses with extreme roll values, the MAE value is less than 1°. When using the whole dataset, there is an improvement exceeding 3°. The results are particularly good on the pitch axis. Even when the training set is not enlarged (e.g., 80% in [35]), the system holds its own when compared with current methods. It is interesting to note that a smaller training set seems to have a lesser effect on accuracy for the roll axis.

Table 2 reports the results for the Pointing’04 dataset under Protocol 1 and Protocol 3. When running Protocol 1, the proposed system compares well with [4], [11], and [29]. The overall MAE is worse than in [34], but ORC_XGB produces the only value of yaw error below 7°. This experiment shows the effectiveness of adding a regression module. As expected, excluding extreme poses improves the results, even more so than with other methods. Focusing on the comparison with [7] and [5], which do not use ML,
shows that a descriptor based on Ricci curvature can be quite effective even when simply used as a metric, without a regression module. The data about Protocol 3 show good results in different conditions, such as a 50/50 training/testing ratio, exclusion of extreme poses. The results with this dataset are generally better for the pitch axis.

With the AFLW2000 dataset, two sets of experiments were performed with Protocol 3 (Table 3). The first set replicates the conditions described in [1], [2], and [5] by excluding the more extreme poses, while the second set alters the training/testing ratio. With this dataset, too, the overall error is under $3^\circ$. In this case, it is the roll axis that shows the best results, reaching below $2^\circ$. Even including the extreme poses and reducing the training percentage to 50%, our results on pitch and particularly yaw surpass other methods, such as [2].

### 6. Conclusion and Future Work

This paper addressed the problem of the head pose estimation with a Ollivier-Ricci curvature based method. We present a method that uses 468 facial reference points to construct a graph on which the geometric curvature based descriptor is applied. For each pair of edges, a curvature value is computed and inserted into a feature vector. The vector is in turn fed to a regression model—the Extreme Gradient Boosting Regressor. Experiments were performed on three well known datasets, with images collected in both controlled and wild environments. The experimental results show that this approach performs well when compared to state-of-the-art methods under identical protocols conditions. Therefore, Ollivier-Ricci curvature is empirically proved to be a good descriptor for this type of problem. The main limitation of this system is its dependence on the face landmark detector. Landmarks become nodes in the graph, and graph construction is a sensitive phase, which can be studied more deeply. Therefore, we plan on investigating the graph creation phase more systematically. In particular, we would like to find a geometry to improve the spatial relationship between the landmarks. A better designed graph is likely to further improve the discriminating performance. Another area to be investigated is the testing of other regressors besides the ones already tried, hoping to find one that fits our data even better.

### References

[1] Abate, P., Barra, C., Pero, and M. Tucci. Head pose estimation by regression algorithm. Pattern Recognition Letters, 140:179–185, 2020. 2, 4, 5, 6, 7

[2] Abate, P., Barra, C., Pero, and M. Tucci. Partitioned iterated function systems by regression models for head pose estimation. Machine Vision and Applications, 32(5), 2021. 2, 4, 5, 7

[3] A. F. Abate, C. Bisogni, A. Castiglione, and M. Nappi. Head pose estimation: An extensive survey on recent techniques and applications. Pattern Recognition, 127:108591, 2022. 1

[4] N. Alioua, A. Amine, A. Benshair, and M. Rziza. Estimating driver head pose using steerable pyramid and probabilistic learning. International Journal of Computational Vision and Robotics (IJCVR), 5, 01 2015. 3, 6

[5] 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, PP:1–1, 04 2020. 2, 6, 7

[6] A. Behera, Z. Wharton, P. Hewage, and S. Kumar. Rotation axis focused attention network (rafa-net) for estimating head pose. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12626 LNCS:223–240, 2021. 3, 4, 5

[7] C. Bisogni, M. Nappi, C. Pero, and S. Ricciardi. Fashe: A fractal based strategy for head pose estimation. IEEE Transactions on Image Processing, 30:3192–3203, 2021. 2, 4, 5, 6

[8] Z. Cao, Z. Chu, D. Liu, and Y. Chen. A vector-based representation to enhance head pose estimation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 1188–1197, 2021. 5

[9] L. Cascone, R. Distasi, and M. Nappi. Ollivier-ricci curvature for head pose estimation from a single image, 2022. 3

[10] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pages 785–794, 2016. 4

[11] V. Drouard, R. Horaud, A. Deleforge, S. Ba, and G. Evangelidis. Robust head-pose estimation based on partially-latent mixture of linear regression. IEEE Transactions on Image Processing, PP, 03 2016. 2, 5, 6

[12] G. Fanelli, T. Weise, J. Gall, and L. Van Gool. Real time head pose estimation from consumer depth cameras. In R. Mester and M. Felsberg, editors, Pattern Recognition, pages 101–110, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

---

| Method         | Pitch | Yaw   | Roll | MAE  |
|----------------|-------|-------|------|------|
| 1 WSM          | 4.82  | 3.11  | 2.25 | 3.39 |
| 2 WSM-BRR      | 4.67  | 3.82  | 2.49 | 3.66 |
| 3 HP2IFS-LsR   | 6.90  | 6.70  | 4.48 | 6.02 |
| 4 ORC_XGB      | 3.51  | 3.02  | 1.94 | 2.82 |
| 5 HP2IFS       | 7.46  | 6.28  | 5.53 | 6.42 |
| 6 ORC_XGB      | 3.51  | 2.93  | 1.96 | 2.80 |
| 7 ORC_XGB      | 6.01  | 4.24  | 5.81 | 5.35 |

Table 3: Performance comparison with state-of-the-art results on AFLW2000 dataset, Protocol 3. ‘!’: without extreme poses; ‘§’: training/testing ratio 80/20; ‘♣’: without ML.
[13] D. Freire-Obregón, M. Castrillón-Santana, P. Barra, C. Bisogni, and M. Nappi. An attention recurrent model for human cooperation detection. *Computer Vision and Image Understanding*, 197-198:102991, 2020. 1

[14] Google. Mediapipe on github. [https://google.github.io/mediapipe/getting_started/python.html](https://google.github.io/mediapipe/getting_started/python.html). Accessed: 2021-03-12. 3

[15] 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. 4

[16] J. Gu, X. Yang, S. De Mello, and J. Kautz. Dynamic facial analysis: From bayesian filtering to recurrent neural network. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-January:1531–1540, 2017. 3, 5, 6

[17] Z. Hu, Y. Xing, C. Lv, P. Hang, and J. Liu. Deep convolutional neural network-based bernoulli heatmap for head pose estimation. *Neurocomputing*, 436:198–209, 2021. 3, 5

[18] H. Ji, F. Su, and Y. Zhu. Robust head pose estimation via semi-supervised manifold learning with 11-graph regularization. In *2011 International Joint Conference on Biometrics (IJCB)*, pages 1–6, 2011. 2

[19] L. Kantorovich. On the translocation of masses. *Dokl. AN SSSR*, 37:199–201, 1942. 3

[20] L. Keselman, J. Iselin Woodfill, A. Grunnet-Jepsen, and D. Freire-Obregón, M. Castrillón-Santana, P. Barra, C. Bisogni, and M. Nappi. An attention recurrent model for human cooperation detection. *Computer Vision and Image Understanding*, 197-198:102991, 2020. 1

[21] D.-T. Lee and B. J. Schachter. Two algorithms for constructing a delaunay triangulation. *International Journal of Computer & Information Sciences*, 9(3):219–242, 1980. 4

[22] Q. Li, Z. Sun, and R. He. Fast multi-view face alignment via multi-task auto-encoders. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 538–545, 2017. 1

[23] Y. Lin, L. Lu, and S.-T. Yau. Ricci curvature of graphs. *Tohoku Mathematical Journal*, 63(4):605 – 627, 2011. 3

[24] H. Liu, S. Fang, Z. Zhang, D. Li, K. Lin, and J. Wang. Mifnet: Collaborative poses perception and matrix fisher distribution for head pose estimation. *IEEE Transactions on Multimedia*, pages 1–1, 2021. 3, 5

[25] S. S. Mukherjee and N. M. Robertson. Deep head pose: Gaze-direction estimation in multimodal video. *IEEE Transactions on Multimedia*, 17(11):2094–2107, 2015. 5

[26] C.-C. Ni, Y.-Y. Lin, F. Luo, and J. Gao. Community detection on networks with ricci flow. *Scientific reports*, 9(1):1–12, 2019. 3

[27] F. Nielsen. On the jensenshannon symmetrization of distances relying on abstract means. *Entropy*, 21(5):485, 2019. 4

[28] Y. Ollivier. Ricci curvature of markov chains on metric spaces. *Journal of Functional Analysis*, 256(3):810–864, 2009. 1, 2

[29] M. Patacchiola and A. Cangelosi. Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods. *Pattern Recognition*, 71:132–143, 2017. 6

[30] R. Sandhu, T. Georgiou, E. Reznik, L. Zhu, I. Kolesov, Y. Senbabaoglu, and A. Tannenbaum. Graph curvature for differentiating cancer networks. *Scientific reports*, 5(1):1–13, 2015. 2

[31] R. Sandhu, T. Georgiou, and A. Tannenbaum. Ricci curvature: An economic indicator for market fragility and systemic risk. *Science Advances*, 2:e1501495–e1501495, 05 2016. 2

[32] J. Sia, E. Jonckheere, and P. Bogdan. Ollivier-ricci curvature-based method to community detection in complex networks. *Scientific Reports*, 9, 07 2019. 4

[33] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020. 4

[34] M. T. Vo, T. Nguyen, and T. Le. Robust head pose estimation using extreme gradient boosting machine on stacked autoencoders neural network. *IEEE Access*, 8:3687–3694, 2020. 2, 6

[35] 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*, 94:196–206, 2019. 2, 5, 6

[36] J. Wee and K. Xia. Ollivier persistent ricci curvature-based machine learning for the protein–ligand binding affinity prediction. *Journal of Chemical Information and Modeling*, 61(4):1617–1626, 2021. 2

[37] M. Xin, S. Mo, and Y. Lin. Eva-gen: Head pose estimation based on graph convolutional networks. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1462–1471, 2021. 2, 5, 6

[38] 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. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:1087–1096, 2019. 2, 4, 5

[39] T.-Y. Yang, Y.-H. Huang, Y.-Y. Lin, P.-C. Hsiu, and Y.-Y. Chuang. Ssr-net: A compact soft stagewise regression network for age estimation. In *IJCAI*, volume 5, page 7, 2018. 5

[40] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155, 2016. 4, 5