Facial Landmark Correlation Analysis

Yongzhe Yan\textsuperscript{1}  Stefan Duffner\textsuperscript{2}  Priyanka Phutane\textsuperscript{1}  Anthony Berthelier\textsuperscript{1}  Christophe Blanc\textsuperscript{1}  Christophe Garcia\textsuperscript{2}  Thierry Chateau\textsuperscript{1}

\textsuperscript{1}Université Clermont Auvergne, CNRS, SIGMA, Institut Pascal, Clermont-Ferrand, France
\textsuperscript{2}Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, France
\texttt{yongzhe.yan@etu.uca.fr}

Abstract

We present a facial landmark position correlation analysis as well as its applications. Although numerous facial landmark detection methods have been presented in the literature, few of them concern the intrinsic relationship among the landmarks. In order to reveal and interpret this relationship, we propose to analyze the facial landmark correlation by using Canonical Correlation Analysis (CCA). We experimentally show that dense facial landmark annotations in current benchmarks are strongly correlated, and we propose several applications based on this analysis.

First, we give insights into the predictions from different facial landmark detection models (including cascaded random forests, cascaded Convolutional Neural Networks (CNNs), heatmap regression models) and interpret how CNNs progressively learn to predict facial landmarks. Second, we propose a few-shot learning method that allows to considerably reduce manual effort for dense landmark annotation. To this end, we select a portion of landmarks from the dense annotation format to form a sparse format, which is mostly correlated to the rest of them. Thanks to the strong correlation among the landmarks, the entire set of dense facial landmarks can then be inferred from the annotation in the sparse format by transfer learning. Unlike the previous methods, we mainly focus on how to find the most efficient sparse format to annotate. Overall, our correlation analysis provides new perspectives for the research on facial landmark detection.

1. Introduction

Facial landmark detection is an active research topic of computer vision in recent years, and the aim is to retrieve the coordinates of a given number of fiducial points on a face image. It is an important prerequisite for numerous applications such as face recognition [9], 3D face reconstruction [19], facial expression analysis [29]. Most of the facial landmarks are positioned on the face contour, eyebrows, eyes, nose and lips.

What is facial landmark correlation? Due to the shape and motion of real 3D objects, there exists a natural correlation between landmarks positioned on these objects (e.g. faces, human body, hands or other objects), also in corresponding 2D projections. Especially for faces, the correlation among landmarks is very strong due to the following two reasons: First, the human face is more rigid than the entire human body or hands which have more articulations and may be observed from any point of view and under severe rotations or deformations. Second, recently-released facial landmark datasets are densely annotated with up to 98 landmarks [46], exhibiting an even stronger correlation. Therefore, in this paper, we focus on the correlation of densely annotated facial landmarks in 300W dataset [35] (68 landmarks, see Fig. 1) and WFLW dataset [46] (98 landmarks).

![Facial Landmark Correlation Analysis](image.png)
Motivation of this analysis: The standard evaluation metric for facial landmark detection is the Normalized Mean Error (NME). NME is the averaged Euclidean distance between each predicted landmark and ground truth, normalized by the inter-ocular distance. A smaller NME indicates a more precise prediction and vice-versa. Other commonly used metrics, including Failure Rate (FR), Cumulative Error Distribution (CED), and Area Under Curve (AUC), are all based on NME.

However, we think that the NME can not describe all aspects of the model prediction. A large NME can signify that the prediction is not precise, but it can not reflect how the prediction is mistaken. We will illustrate this in the following example.

Current deep learning-based state-of-the-art methods can be categorized into two types: Coordinate Regression CNNs (CR-CNNs) and Heatmap Regression CNNs (HR-CNNs) [50, 47]. CR-CNNs predict the numeric X and Y coordinate values of each landmark in the last Fully-Connected (FC) layer. HR-CNNs adopt Fully Convolutional Neural Network [26] architectures that estimate a spatial probability map for each landmark. That is, the value of each pixel on the heatmap represents the presence probability of the landmark at this pixel [45].

Each model has its strengths and weaknesses. HR-CNNs show a strong capability of handling complex pose variations. However they globally lack robustness, and in failure cases, landmarks are predicted at unreasonable positions which are far away from the ground truth (see Fig. 2 (b)). On the other hand, CR-CNNs are generally more efficient in terms of computation and memory usage but also locally less precise. The prediction of single-stage CR-CNNs is usually constrained in a reasonable shape similar to the ground truth, being not extremely precise (see Fig. 2 (a)).

This investigation can be confirmed by the current research trend. Most of the latest HR-CNNs aim at reinforcing the robustness of the detection by introducing global constraints [43, 24, 30] or temporal consistency [40]. However, the recent CR-CNNs enhance local precision using coarse-to-fine frameworks [42, 13, 27, 6, 22, 16, 14].

For instance, the models whose results are illustrated in Fig. 2 (a) and (b) may have similar NME. Nevertheless, these two models have distinct characteristics. In this case, we think that landmark correlation can be the key to explain and, furthermore, to quantify the weaknesses of the two models. We assume that the local imprecision problem of CR-CNNs is due to the fact that the predicted landmark positions are too much correlated (or regularized). In contrast, for HR-CNNs, the “outliers” predicted in unreasonable positions can be considered as a violation of the natural landmark correlation.

We want to make clear that landmark correlation can not be used as a stand-alone evaluation metric, though it provides a new perspective to interpret the model prediction. Similar correlation compared to the ground truth is a necessary condition but not sufficient to precise prediction. Even identical correlation does not ensure precise prediction. However, big correlation difference between prediction and ground truth can conclude that the prediction is not precise. Our contributions can be summarized as follows:

- We present a CCA-based correlation analysis as a novel tool to interpret and quantify the relationship among a set of landmarks (section 3).
- We use this model-agnostic correlation analysis to interpret the three most popular facial landmark detection models in the last decade, including cascaded random forest, CR-CNNs and HR-CNNs (section 4).
- We propose a few-shot learning method to reduce the effort of manual annotation of dense landmarks with the help of the landmark correlation (section 5). By analyzing the landmark correlation in dense format, we are able to form a sparse format by selecting a set of landmarks which are most correlated to the rest of them. We propose to learn dense facial landmark predictions by the images annotated with the sparse format, which requires less annotation cost. Our method shows two advantages: (i) Compared to existing methods which use existing sparse formats [1, 54, 44, 4, 28], the selection of our sparse format is purely data-driven. (ii) The number of sparse landmarks can be arbitr-
ily chosen depending on the minimum correlation required between the selected landmarks and the rest.

2. Related Work

Facial landmark detection in the last decade: In 2010, Dollár et al. proposed Cascaded Pose Regression [10], which laid the foundation for several well-known cascaded regression methods including SDM [49], ESR [5] and ERT [20]. In the deep learning era, cascaded CR-CNNs [39, 53, 42] continue to follow its general coarse-to-fine structure. HR-CNNs [45, 32, 3], originally introduced in 2005 [12], gained much popularity in recent years. In this paper, we propose facial landmark correlation analysis to take a closer look into three of the most important models in the last decade: Cascaded Random Forest model [20], CR-CNN [13] and HR-CNN [3].

Component analysis in facial landmark detection: The use of Principal Component Analysis (PCA), especially the 3DMM model [2], is of great importance in the current research of face analysis. PCA has been used for facial landmark detection since 1995 [7] in the Point Distribution Model. PCA is used to analyze the shape variance with respect to the mean shape, including face rotation, facial expressions and identity variance. The biggest difference between the PCA and our CCA study is that our CCA study analyzes the relationship between individual facial landmarks while PCA focuses on the global face shape.

CNN Interpretation via CCA: Lately, using CCA to interpret CNN representations [33, 31, 21] is an emerging subject. They used CCA to analyze the representations of two different CNNs and gained some insights on the learning process. They mainly focused on attenuating the noise in the CNN representation, which is brought by different initializations. However, as we will show in Sec. 4.3, when we use CCA to analyze the CR-CNNs, we analyse the correlation between different neurons in the same layer. As being trained altogether, no such noise will be involved. Therefore, we do not apply any pre-preprocessing steps such as Singular Value Decomposition (SVD) as in [33].

Few-shot learning for facial landmark detection: Few-shot learning, or weakly supervised learning, is now attracting increasing attention in the community. A recent work [11] proposed a mechanism to enable the training on fewer labeled images. In this paper, we focus on how to learn with fewer landmarks rather than fewer images.

We assume that a landmark can be easily transferred from another landmark that is highly correlated. This is not new and has already been proved in several work which focus on transferring the data between two annotations with different semantic meanings [38, 55, 52]. We also find similar ideas in some existing coarse-to-fine approaches [27, 6, 36, 37], where the entire set of landmarks is divided into several partitions inside which the authors assume a strong correlation. Specifically, Tan et al. [41] proposed a few-shot learning method to reconstruct the global shape from a sparse landmark format. DeCaFA [8] can be trained with coarsely annotated examples by exploiting landmark-wise attention. However, in the above works, they mainly focused on how to improve the model performance given the pre-defined sparse format. The choice of the sparse format and their partitioning are heuristic. In contrast, in this paper, we focus on how to find the best sparse format that will most benefit the few-shot learning. Our selection of the sparse landmark format is entirely based on the statistics of the underlying data. Our approach is inspired by the work on multi-task learning [51, 23].

3. Facial Landmark Correlation Analysis

3.1. Canonical Correlation Analysis [18]

Given a $p$-dimensional random variable $U \in \mathbb{R}^p$ and a $q$-dimensional variable $V \in \mathbb{R}^q$, CCA aims to find the best linear transformation $a \in \mathbb{R}^p$ and $b \in \mathbb{R}^q$ that maximize the correlation:

$$
\text{Cor}(U, V) = \frac{a^T \sum_{UV} b}{\sqrt{a^T \sum_{UU} a} \cdot \sqrt{b^T \sum_{VV} b}},
$$

where

$$
\sum_{UV} = \text{Cov}(U, V) = E[(U - E[U])(V - E[V])].
$$

The operator $E$ denotes the expected value of its argument. This problem can be solved by SVD after basis change. This gives $\min(p, q)$ correlation coefficients sorted from the most correlated to the least correlated canonical directions. We consider the mean value $\text{Cor}(U, V)$ of the correlation coefficients as an overall measure [34].

3.2. Facial Landmark Correlation

To focus on the variance of the face shape, we apply an important pre-processing step. We crop, center all the faces and then further normalize their sizes. We consider the 2D Cartesian coordinates as a two-dimensional variable. Specifically, we calculate the absolute value of the correlation coefficients (ranged from -1 to 1) as we are interested in the magnitude of the correlation between two landmarks but not their directions.

To be clear, the canonical correlation between the $i$-th and the $j$-th landmark in the annotation format can be found at the $i$-th row and the $j$-th column on the affinity matrix $A$:

$$
A_{i,j} = \left| \text{Cor}(L_i, L_j) \right|,
$$

where $L_i, L_j \in \mathbb{R}^2$ indicate the annotation of the $i$-th and the $j$-th landmark on the entire dataset.
The correlation affinity matrix on the 300W train subset [35] is shown in Fig. 1. We draw several conclusions from the affinity matrices in Fig. 1: (i) The correlation among the landmarks belonging to the same facial component is generally more significant than the others. (ii) Some landmarks from the same component are less correlated (such as upper-lip and lower-lip). This is due to the shape variance e.g. different facial expressions. (iii) Certain facial components from different facial components are strongly correlated, such as eyebrows and eyes, the outer and inner contour of lips, which is plausible.

4. Facial Landmark Model Interpretation

We now use the proposed landmark correlation analysis to interpret three important facial landmark detection models: cascaded random forest [20], cascaded CR-CNN [13] and HR-CNN [3]. We will focus on three aspects. (i) What are the characteristics of the final prediction from each model? (ii) Are there any meaningful differences between cascading and stacking? (iii) Can we interpret the learning dynamics of the CNN models for landmark detection?

4.1. Model Settings

All of the analyzed models are trained on the 300W train subset and analyzed on 300W validation subset.

**Cascaded Random Forest:** ERT [20] consists of 10 cascaded random forest regressors. Each regressor comprises 500 trees and the depth of the trees is 5. We use the implementation from [48]. The initialized shape is the mean shape of the train subset. The NME of this model is 6.18% for the validation set.

**Cascaded CR-CNN:** We reproduced the model of Fan et al. [13]. However, we added two additional stages to further boost its performance. Therefore, the overall structure has four stages. The main network in the first stage is ResNet18 and the sub-networks in the following stages consist of a single ResNet block and a single FC layer. The NME of this model is 3.66%.

**Stacked HR-CNN:** We used the official implementation of [3]. The hourglass models are stacked in 4 stages. The NME of this model is 3.52%.

4.2. Characteristics of the Prediction

In this section, we visualize the affinity matrix error \( A_{pred} - A_{GT} \), where \( A_{pred} \) is the CCA affinity matrix calculated on the output of each model and \( A_{GT} \) is the CCA affinity matrix calculated on the ground truth.

**Cascaded random forest ERT:** In Fig. 3, we show the final prediction of ERT through landmark correlation. We can see that the overall correlation of the prediction is higher than the ground truth. There are more green parts than red parts and the shades of the green parts are higher than the shades of red parts. We can make two important observations from the CCA matrix error.

(i) The landmarks on the face contour are generally more correlated to the other facial components. It means that the predicted face contour from ERT is too regularized.

(ii) Some landmarks on the right are over-correlated with other landmarks on the left (marked in the black rectangles in Fig. 3 (a)). For example, the correlation between the left tip of the left eyebrow (landmark index 17) and the right tip of the lip (landmark index 54) is significantly bigger than the ground truth. It statistically signifies that the prediction of the ERT does not have enough horizontal variance compared to the ground truth, which is due to the failures confronting extreme head poses.

The shown visual examples (Fig. 3 (b)) confirm the above investigations on face contours ((a)(b)) and large poses ((c)(d)). Further, our observations are consistent with the major concern about ERT expressed in the literature [56], which is the poor robustness to pose variations.

**Cascaded CR-CNN & stacked HR-CNN:** In Fig. 4, we show the correlation matrix error of CR-CNN and HR-CNN. Overall, the prediction of stacked HR-CNN has a lower correlation error compared to cascaded CR-CNN. And both of them show a smaller correlation error than ERT (see the scale of colorbar on the right). Also note that the
CNN tends to correlate adjacent landmarks. Both of the CNN based methods share this important characteristic. This is probably due to the convolution operation used in the CNN, which excessively exploits local semantic information. For example, in Fig. 4, the correlation between the left tip of the left eyebrows/eyes and the upper-left face contour (blue rectangles), the correlation between the lip and the bottom face contour (red rectangles) and the correlation among the landmarks on the bottom face contour (cyan rectangles) are significantly higher than the ground truth correlation. Some landmarks that are over-correlated to their adjacent landmarks, show inferior correlation with the more distant landmarks (the correlation between upper-left face contour and lips, black rectangles).

Stacking for HR-CNN and Cascading for CR-CNN behave differently. When comparing Fig. 5 (a) and Fig. 4 (a), we observe that a single-stage CR-CNN suffered from more severe over-correlation problem compared to the cascaded CR-CNN. Therefore, the coarse prediction of the first stage in cascaded CR-CNN is indeed over-regularized. The following stages in cascaded CR-CNN learn the further shape variance, which de-correlates the output shape from the first stage. We note that in CR-CNNs, the output is linearly connected with the previous FC layer. This may explain why there are always excessive correlations present in the output of CR-CNN. When comparing Fig. 5 (b) and Fig. 4 (b), we find that the correlation error of the prediction from the first stage is almost the same as the final prediction on stacked HR-CNN. Therefore, the role of stacking in HR-CNN is different from the cascading in CR-CNN. Further study on cascading and stacking is presented in the supplementary material.

HR-CNN is more likely to violate landmark correlation than CR-CNN under challenging conditions. In Fig. 6 (b), we can see that the correlation between the inner facial components on 300VW Scenario3 is weaker than the ground truth, especially on the right eyes/eyebrows (black rectangle). This is consistent with the weakness of HR-CNN that we mentioned in Fig. 2 (b). If we compare Fig. 6 (b) and Fig. 4 (b), we observed that this problem only happens on 300VW ‘S3, which involves challenging conditions such as occlusions, motion blurs, complex lighting conditions, etc. However, if we compare Fig. 6 (a) and Fig. 6 (b), we find that CR-CNN is still robust under these challenging conditions, especially on inner facial components.

4.3. CR-CNN Learning Dynamics

In this section, we study how the CR-CNN progressively learns from the beginning. To this end, we plot the evolution of the CR-CNN output correlations during training. We do not analyze the learning dynamic of HR-CNN due to the operation of taking the maximum value on the final heatmap. We think that it is necessary to develop a 2D CCA method in the future to analyze the output heatmap directly.

We trained a one-staged ResNet-18 on the 300W dataset. Both the convolutional layers and the FC layers are initialized from a normal distribution [15]. The CR-CNN is trained for 350 epochs with the learning rate decayed by 0.3 for each 70 epochs. We observe the following phases during the first 70 epochs:

**Phase 1 Group Inner Facial Components:** More rigid parts learn first. The first thing that CNN starts to learn is to group the inner facial components. We can observe in Fig. 7 (b) that CNN firstly learns a relatively strong correlation among the landmarks on the inner facial components (eyebrows, eyes, noses) and separate them from the other landmarks on face contours.

**Phase 2 Recognize Each Facial Component:** Next, the CNN starts to gradually identify the facial components (eyebrows, nose, mouths, etc.). In this phase, the correlation among the landmarks which belong to the same facial component grow stronger (see Fig. 7 (c)). The CNN recognizes the eyes, nose and lips almost simultaneously.

**Phase 3 Refine the Prediction:** The CNN learns to refine the prediction in two aspects: (i) enforce the correlation inside each facial component, especially the neighbouring landmarks; (ii) reduce some excessive correlations (e.g. the
Figure 7: CCA affinity matrices on the prediction of CR-CNN in different training epochs. The percentage shown in each figure caption refers to NME.

Figure 8: The workflow of our few-shot learning method. The value of M and N indicates respectively the total number of the landmarks in the dense annotation format and the total number of the images collected. The value of m and n can be arbitrarily chosen depending on the requirements. m can be considered as annotation budget. We save the time to annotate M-m landmarks on N-n images.

Figure 9: Standard deviation (Std) of CCA affinity matrices on the prediction of CR-CNN in different training epochs. The percentage shown in each caption refers to NME at 140th/280th epoch respectively.

5. Few-shot learning

**Motivation:** As the size of datasets grows larger and the landmark format becomes denser [46], it is time-consuming to densely annotate each landmark on all of the images. Few-shot learning has attracted increasing attention in the community. Due to the presence of strong landmark correlation in the dense format, we believe that it is not cost-effective to annotate every landmark, especially when the budget for manual annotation is limited.

Our few-shot learning method is also useful in the following situation: we want to extend an existing format (e.g. 300W format) for a specific use (e.g. detect the landmarks on the wing of the nose or the face contour around the forehead) with limited budget. With this correlation analysis, we are able to find out how the landmarks we want to extend are correlated to the landmarks already annotated and find an efficient strategy for manual annotation.

**Workflow:** We propose a few-shot regression method to find the most cost-effective landmarks to annotate (see Fig. 8). We assume that a landmark can be easily transferred from another landmark that is highly correlated. As a result, to find the most “important” landmarks facilitating the learning of the others, we search for a set of landmarks that

The evolution of the affinity matrices after 70 epochs is difficult to visualize as the evolution of the correlation value is small. To this end, we calculate the standard deviation (Std) of the affinity matrices in different stages (see Fig. 9). We observe that after 70 epoch, the variation of the correlation value related to the landmarks on the face contour is significantly higher than the others, indicating that the correlation between lower face contour and lips is reduced, see Fig. 7 (d)).
has maximum correlation with the rest of the landmarks. The problem to solve in Fig. 8 step (4) can be described as:

Find a set of landmarks indexed by \( m \), which maximize the minimum correlation \( c \) with rest of the landmarks indexed by \( \mathcal{M} - m \):

\[
m = \arg \max_{m \in \mathcal{M}} (c_m),
\]

\[
c_m = \min_{j \in (\mathcal{M} - m)} (\max_{i \in m} (A_{i,j})).
\]

\( \mathcal{M} \) denotes the complete set of landmark index in the dense format, \( A_{i,j} \) denotes the \( i \)-th row and \( j \)-th column of the correlation affinity matrix \( A \) analyzed on \( n \) images. \( c \) can be considered as a criterion of the sparse format selection \( m \). Maximized minimum correlation \( \hat{c} = \max_{m \in \mathcal{M}} (c_m) \).

This problem resembles K-center facility problem [17]. A classical K-center problem can be described as: Given a city with \( M \) locations, find the best \( k \) locations to build \( k \) facilities, so that the farthest distance from location to its nearest facility has to be as small as possible.

In our problem, the locations in the city can be considered as all the landmarks in the dense annotation format \( \mathcal{M} \). The \( k \) facilities can be considered as the landmarks selected in our sparse format \( m \). The distance between the landmark \( i \) and \( j \) can be considered as \( 1 - A_{i,j} \). In fact, a high correlation between two landmarks signifies that the distance between two landmarks is small.

K-center problem is NP-hard. Fortunately, this problem can be efficiently solved by mixed-integer programming using Gurobi [25], a powerful mathematical optimization solver. We present the canonical form of this problem:

Minimize \( z \), with subject to:

\[
\sum_j x_{ij} = 1 \quad \forall i
\]

\[
\sum_j y_j = m
\]

\[
x_{ij} \leq y_j \quad \forall i, j
\]

\[
(1 - A_{i,j})x_{ij} \leq z \quad \forall i, j
\]

\[
x_{ij} \in \{0, 1\} \quad \forall i, j
\]

\[
y_j \in \{0, 1\} \quad \forall j
\]

\( x_{ij} \) indicates that landmark \( i \) is inferred from the position of landmark \( j \). \( y_j = 1 \) indicates that the landmark \( j \) is selected in the sparse format. \( \sum_j x_{ij} = 1 \) ensures that all the landmarks are inferred from another landmark. \( \sum_j y_j = m \) ensures that there are \( m \) landmarks selected in the sparse format. \( x_{ij} \leq y_j \) ensures that landmark \( i \) can be inferred from landmark \( j \) only when landmark \( j \) is selected.
Table 1: NME(%) performance comparison of the few-shot learning task in Fig. 8 by using existing formats and searched sparse format (denoted as ours). The settings of M and m is consistent with the columns in Fig. 10. Ratio represents the percentage of densely annotated images, which is the value of n/N.

![Figure 11: Relationship between m and maximized minimum correlation $\hat{c}$. For each m, we run our searching method 10 times and plot the mean and variance of $\hat{c}$.](image1)

in the sparse format. Finally, the maximized minimum correlation can be obtained by $\hat{c} = 1 - z$. This optimization can be finished in just several seconds on a normal PC.

**Experiments:** We demonstrate several sparse formats searched by our method on the dense formats of 300W [35], 300W inner (exclude the face contour and eyebrows) and WFLW [46] (see Fig. 10). Note that the searched formats can be different each time depending on the data (n images) sampled from the entire dataset (N images). We also list some existing sparse formats: MAFL [54], LFW [44], AFLW [28] and COFW [4] with same annotation budget m as comparison. The advantage of our method is that we are able to distribute the annotation budget (m landmarks to annotate) evenly on each component based on the difficulty to learn. Compared to the heuristic choices made by common knowledge, our choice is completely data-driven.

In Tab. 1, we present the performance comparison on this task. We find that our sparse format achieves comparable performance compared to MAFL format and LFW format on 300W Inner. When the landmarks on the face contour are included in the learning (on 300W Full & WFLW Full), our format demonstrates more significant improvement compared to AFLW format and COFW format. We also noticed that our searched format is more advantageous with fewer densely annotated images. NME difference between our format and pre-defined format is larger when trained with ratio of 5% and 10%.

To further investigate the relationship between the annotation budget m and the maximized minimum correlation $\hat{c}$ on different dense formats, we run our sparse format search-

![Figure 12: NME and maximized minimum correlation $\hat{c}$ with different annotation budget using our sparse format. Tested on 300W full format, M=68, m=5-20.](image2)

ing method on each dataset with incremental m. The relationship between $\hat{c}$ and m is shown in Fig. 11. We also demonstrate the values of $c$ when using existing sparse formats. Our search method is able to find a bigger $c$ compared to the existing ones, which can result in better performance on the few-shot learning task. This figure is useful for us to choose an appropriate annotation budget m. For example, on the 300W full format, we find a significant improvement on $\hat{c}$ by including 9 landmarks in the sparse format. It indicates that it is more advantageous to set the annotation budget m to 9 than 8 because the performance can probably be largely improved by adding only 1 annotation budget.

In Fig. 12, we plot the relationship between the performance of our few-shot learning tasks (NME %) and the maximized minimum correlation $\hat{c}$ with different m in our sparse format. We find that as the $\hat{c}$ goes up, the NME is decreased accordingly. It confirms our assumption that the performance of this few-shot learning task is strongly related to the $\hat{c}$ when using our sparse format.

6. Conclusions

We propose a correlation analysis as a simple yet effective tool to interpret the relationship among facial landmarks. Our analysis provides a new perspective which is completely different to the commonly used metric NME. Conducting this analysis on the output prediction, we gain some interesting insights on the three most important mod-
els in the last decade. We also propose a few-shot learning method to drastically reduce the cost of laborious manual dense annotation. Our methodology on the coordinate correlation can be further extended to 3D facial landmarks, hand/body pose, object landmarks and even the bounding boxes of object detection.

References

[1] Peter N Belhumeur, David W Jacobs, David J Kriegman, and Neeraj Kumar. Localizing parts of faces using a consensus of exemplars. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2930–2940, 2013. 2

[2] Volker Blanz, Thomas Vetter, et al. A morphable model for the synthesis of 3d faces. In *Siggraph*, volume 99, pages 187–194, 1999. 3

[3] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*, Oct 2017. 3, 4

[4] Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1513–1520, 2013. 2, 8

[5] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190, 2014. 3

[6] Xi Chen, Erjin Zhou, Yuchen Mo, Jiancheng Liu, and Zhimin Cao. Delving deep into coarse-to-fine framework for facial landmark localization. In *CVPR Workshops*, pages 2088–2095, 2017. 2, 3

[7] Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995. 3

[8] Arnaud Dapogny, Kevin Bailly, and Matthieu Cord. Decafa: Deep convolutional cascade for face alignment in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, October 2019. 3

[9] Changxing Ding and Dacheng Tao. A comprehensive survey on pose-invariant face recognition. *ACM Transactions on Intelligent Systems and Technology*, 7(3):37:1–37:42, Feb. 2016. 1

[10] Piotr Dollár, Peter Welinder, and Pietro Perona. Cascaded pose regression. In *2010 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1078–1085. IEEE, 2010. 3

[11] Xuanyi Dong and Yi Yang. Teacher supervises students how to learn from partially labeled images for facial landmark detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 3

[12] Stefan Duffner and Christophe Garcia. A connexionist approach for robust and precise facial feature detection in complex scenes. In *Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis*, pages 316–321, 2005. 3

[13] Haoqiang Fan and Erjin Zhou. Approaching human level facial landmark localization by deep learning. *Image and Vision Computing*, 47:27–35, 2016. 2, 3, 4

[14] Zhen-Hua Feng, Josef Kittler, Muhammad Awaais, Patrik Huber, and Xiao-Jun Wu. Wing loss for robust facial landmark localisation with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2235–2245, 2018. 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 5

[16] Zhenliang He, Meina Kan, Jie Zhang, Xilin Chen, and Shiguang Shan. A fully end-to-end cascaded cnn for facial landmark detection. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 200–207, 2017. 2

[17] Dorit S Hochbaum and David B Shmoys. A best possible heuristic for the k-center problem. *Mathematics of operations research*, 10(2):180–184, 1985. 7

[18] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936. 3

[19] Aaron S Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 1

[20] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874, 2014. 3, 4

[21] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, 2019. 3

[22] Marek Kowalski, Jacek Naruniec, and Tomasz Trzciński. Deep alignment network: A convolutional neural network for robust face alignment. In *Proceedings of the International Conference on Computer Vision & Pattern Recognition (CVPRW), Faces-in-the-wild Workshop/Challenge*, volume 3, page 6, 2017. 2

[23] J. Li, P. Zhou, Y. Chen, J. Zhao, S. Roy, Y. Shuicheng, J. Feng, and T. Sim. Task relation networks. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 932–940, Jan 2019. 3

[24] Zhiwei Liu, Xiangyu Zhu, Guosheng Hu, Haiyun Guo, Ming Tang, Zhen Lei, Neil M. Robertson, and Jinqiao Wang. Semantic alignment: Finding semantically consistent ground-truth for facial landmark detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[25] Gurobi Optimization LLC. Gurobi optimizer reference manual, 2019. 7

[26] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 2
[27] Jiangjing Lv, Xiaohu Shao, Junliang Xing, Cheng Cheng, and Xi Zhou. A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3317–3326, 2017. 2, 3

[28] Peter M. Roth Martin Koestinger, Paul Wohlhart and Horst Bischof. Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization. In Proc. First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies, 2011. 2, 8

[29] Brais Martinez, Michel F Valstar, Bihan Jiang, and Maja Pantic. Automatic analysis of facial actions: A survey. IEEE Transactions on Affective Computing, 2017. 1

[30] Daniel Merget, Matthias Rock, and Gerhard Rigoll. Robust facial landmark detection via a fully-convolutional local-global context network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 781–790, 2018. 2

[31] Ari Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In Advances in Neural Information Processing Systems, pages 5727–5736, 2018. 3

[32] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In Proceedings of the European conference on computer vision, pages 483–499. Springer, 2016. 3

[33] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In Advances in Neural Information Processing Systems, pages 6076–6085, 2017. 3

[34] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In Advances in Neural Information Processing Systems 30, pages 6076–6085. 2017. 3

[35] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In Proceedings of the IEEE International Conference on Computer Vision Workshops, pages 397–403, 2013. 1, 4, 8

[36] Zhiwen Shao, Shouhong Ding, Yiru Zhao, Qinchuan Zhang, and Lizhuang Ma. Learning deep representation from coarse to fine for face alignment. In IEEE International Conference on Multimedia and Expo (ICME), 2016. 3

[37] Zhiwen Shao, Hengliang Zhu, Xin Tan, Yangyang Hao, and Lizhuang Ma. Deep multi-center learning for face alignment. Neurocomputing, April 2019. 3

[38] Brandon M Smith and Li Zhang. Collaborative facial landmark localization for transferring annotations across datasets. In European Conference on Computer Vision, pages 78–93. Springer, 2014. 3

[39] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3476–3483, 2013. 3

[40] Ying* Tai, Yicong* Liang, Xiaoming Liu, Lei Duan, Jilin Li, Chengjie Wang, Feiyeue Huang, and Yu Chen. Towards highly accurate and stable face alignment for high-resolution videos. In The AAAI Conference on Artificial Intelligence (AAAI), 2019. 2

[41] Shuqiu Tan, Dongyi Chen, Chenggang Guo, and Zhiqi Huang. Extra facial landmark localization via global shape reconstruction. Computational intelligence and neuroscience, 2017. 2017. 3

[42] George Trigeorgis, Patrick Snape, Mihalis A Nicolaou, Epameinondas Antonakos, and Stefanos Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4177–4187, 2016. 2, 3

[43] Roberto Valle, Jose M. Buenaposada, Antonio Valdes, and Luis Baumllea. A deeply-initialized coarse-to-fine ensemble of regression trees for face alignment. In The European Conference on Computer Vision (ECCV), September 2018. 2

[44] Luc Van Gool. Real-time facial feature detection using conditional regression forests. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2578–2585, 2012. 2, 8

[45] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4724–4732, 2016. 2, 3

[46] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2129–2138, 2018. 1, 6, 8

[47] Yue Wu and Qiang Ji. Facial landmark detection: A literature survey. International Journal of Computer Vision, pages 1–28, 2017. 2

[48] Jiashun Xiao. Face alignment with an ensemble of gradient boosting trees. https://github.com/JiaShun-Xiao/face-alignment-ert-2D, 2019. 4

[49] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 532–539, 2013. 3

[50] Yongze Yan, Xavier Naturel, Thierry Chateau, Stefan Duffner, Christophe Garcia, and Christophe Blanc. A survey of deep facial landmark detection. In Reconnaissance des Formes, Image, Apprentissage et Perception RFIAP, 2018. 2

[51] A. R. Zamir, A. Sax, W. Shen, L. Guibas, J. Malik, and S. Savarese. Taskonomy: Disentangling task transfer learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3712–3722, June 2018. 3

[52] Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Leveraging datasets with varying annotations for face alignment via deep regression network. In Proceedings of the IEEE International Conference on Computer Vision, pages 3801–3809, 2015. 3

[53] Jie Zhang, Shiguang Shan, Meina Kan, and Xilin Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time
face alignment. In Proceedings of the European conference on computer vision, pages 1–16. Springer, 2014. 3

[54] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Learning deep representation for face alignment with auxiliary attributes. IEEE transactions on pattern analysis and machine intelligence, 38(5):918–930, 2016. 2, 8

[55] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. Transferring landmark annotations for cross-dataset face alignment. arXiv preprint arXiv:1409.0602, 2014. 3

[56] Shizhan Zhu, Cheng Li, Chen-Change Loy, and Xiaoou Tang. Unconstrained face alignment via cascaded compositional learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3409–3417, 2016. 4