Mirror, mirror on the wall, tell me, is the error small?

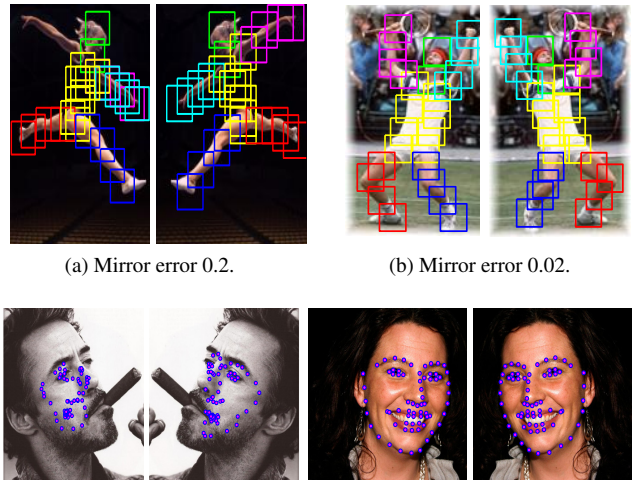
Heng Yang and Ioannis Patras
Queen Mary University of London
{heng.yang, I.Patras}@qmul.ac.uk

Abstract

Do object part localization methods produce bilaterally symmetric results on mirror images? Surprisingly not, even though state of the art methods augment the training set with mirrored images. In this paper we take a closer look into this issue. We first introduce the concept of mirrorability as the ability of a model to produce symmetric results in mirrored images and introduce a corresponding measure, namely the mirror error that is defined as the difference between the detection result on an image and the mirror of the detection result on its mirror image. We evaluate the mirrorability of several state of the art algorithms in two of the most intensively studied problems, namely human pose estimation and face alignment. Our experiments lead to several interesting findings: 1) Most of state of the art methods struggle to preserve the mirror symmetry, despite the fact that they do have very similar overall performance on the original and mirror images; 2) the low mirrorability is not caused by training or testing sample bias - all algorithms are trained on both the original images and their mirrored versions; 3) the mirror error is strongly correlated to the localization/alignment error (with correlation coefficients around 0.7). Since the mirror error is calculated without knowledge of the ground truth, we show two interesting applications - in the first it is used to guide the selection of difficult samples and in the second to give feedback in a popular Cascaded Pose Regression method for face alignment.

1. Introduction

The evolution of mirror (bilateral) symmetry has profoundly impacted animal evolution [7]. As a consequence, the overwhelming majority of modern animals (>99%), including humans, exhibit mirror symmetry. As shown in Fig. 1, the mirror of an image depicting such objects shows a meaningful version of the same objects. Taking face images as a concrete example, a mirrored version of a face image is perceived as the same face. In recent year, object (parts) localization has made significant progress and several methods have reported close-to-human performance. This includes localization of objects in images (e.g. pedestrian or face detection) or fine-grained localization of object parts (e.g. face parts localization, body parts localization, bird parts localization). Most of those methods augment the training set by mirroring the positive training samples. However, are these models able to give symmetric results on a mirror image during testing?



(a) Mirror error 0.2. (b) Mirror error 0.02. (c) Mirror error 0.6. (d) Mirror error 0.02.

Figure 1: Example pairs of localization results on original (left) and mirror (right) images. First row: Human Pose Estimation [27], second row: Face Alignment by RCPR [4]. The first column (a and c) shows large mirror error and the second (b and d) small mirror error. Can we evaluate the performance without knowing the ground truth?

In order to answer this question we first introduce the concept of mirrorability, i.e., the ability of an algorithm to give on a mirror image bilaterally symmetric results, and a quantitative measure called the mirror error. The latter is defined as the difference between the detection result on an image and the mirror of detection result on its mirror image. We evaluate the mirrorability of several state of the art algorithms in two representative problems (face alignment and human pose estimation) on several datasets. One

would expect that a model that has been trained on a dataset augmented with mirror images to give similar results on an image and its mirrored version. However, as can be seen in Fig. 1 first column, several state of the art methods in their corresponding problems sometimes struggle to give symmetric results in the mirror images. And for some samples the mirror error is quite large. By looking at the mirrorability of different approaches in human pose estimation and face alignment, we arrive at three interesting findings. First, most of the models struggle to preserve the mirrorability - the mirror error is present and sometimes significant; Second, the low mirrorability is not likely to be caused by training or testing sample bias - the training sets are augmented with mirrored images; Third, the mirror error of the samples is highly correlated with the corresponding ground truth error.

This last finding is significant since one of the *nice* properties of the proposed mirror error is that it is calculated 'blindly', i.e., without using the ground truth. We rely on this property in order to show two examples of how it could be used in practice. In the first one the mirror error is used as a guide for difficult samples selection in unlabelled data and in the second one it is used to provide feedback on a cascaded pose regression method for face alignment. In the former application, the samples selected based on the mirror error have shown high consistency across different methods and high consistency with the difficult samples selected based on the ground truth alignment error. In the latter application, the feedback mechanism is used in a multiple initializations scheme in order to detect failures - this leads to large improvements and state of the art results in face alignment.

To summarize, in this paper we make the following contributions:

- To the best of our knowledge, we are the first to look into the mirror symmetric performance of object part localization models.
- We introduce the concept of mirrorability and show how the corresponding measure, called mirror error, that we propose can be used in evaluating general object part localization methods.
- We evaluate the mirrorability of several algorithms in two domains (i.e. face alignment and body part localization) and report several interesting findings on the mirrorability.
- We show two applications of the mirrorability in the domain of face alignment.

2. Mirrorability in Object Part Localization

2.1. Mirrorability concepts and definitions

We define mirrorability as the ability of a model/algorithm to preserve the mirror symmetry when applied on an image and its mirror image. In order to quantify it we introduce a measure called mirror error that is defined as the difference between a detection result on an image and the mirror of the result on its mirror image. Specifically, let us denote the shape of an object, for example a human or a face, by a set of K points, $\mathrm{X} = \{\mathrm{x}_k\}_{k=1}^{K}$, where x_k are the coordinates of the k-th point/part. The detection result on the original image is denoted by ${}^{\mathrm{q}}\mathrm{X} = \{{}^{\mathrm{q}}\mathrm{x}_k\}_{k=1}^{K}$ and the detection result on the mirror image is denoted by ${}^{\mathrm{p}}\mathrm{X} = \{{}^{\mathrm{p}}\mathrm{x}_k\}_{k=1}^{K}$. The mirror transformation of ${}^{\mathrm{p}}\mathrm{X}$ to the original image is denoted by ${}^{\mathrm{p}\to\mathrm{q}}\mathrm{X} = \{{}^{\mathrm{p}\to\mathrm{q}}\mathrm{x}_k\}_{k=1}^{K}$, where ${}^{\mathrm{p}\to\mathrm{q}}\mathrm{x}_k$ denotes the mirror result of the k-th part on the original image. Generally, a different index k' is used on the mirror image (e.g. a left eye in an image becomes a right eye in the mirror image). Therefore, the transformation consists of image coordinates transform and the part index mirror transform ($k' \to k$). The image coordinate transform is applied on the horizontal coordinate, that is ${}^{\mathrm{p}}x_k = \mathrm{w}_I - {}^{\mathrm{q}}x_k$, where w_I is the width of the image I and ${}^{\mathrm{p}}x_k$ is the x coordinate of the k point in the mirror image. The index re-assignment is based on the the mirror symmetric structure of a specific object, with an one-to-one mapping list where, for example, the left eye index is mapped to the right eye index. Formally, the mirror error of the k landmark (body joint or facial point) is defined as $||{}^{\mathrm{q}}\mathrm{x}_k - {}^{\mathrm{p}\to\mathrm{q}}\mathrm{x}_k||$, and the sample-wise mirror error as:

$$e_{\mathrm{m}} = \frac{1}{K}\sum_{k=1}^{K} ||{}^{\mathrm{q}}\mathrm{x}_k - {}^{\mathrm{p}\to\mathrm{q}}\mathrm{x}_k|| \tag{1}$$

The mirror error that is defined in the above equation has the following properties: First, a high mirror error reflects low mirrorability and vice visa; Second, it is symmetric, i.e., given a pair of mirror images it makes no difference which is considered to be the original; Third, and importantly, calculating the mirror error does not require ground truth information.

In a similar way we calculate the ground truth localization error ${}^{\mathrm{q}}e_{\mathrm{a}}$ as the difference between the detected locations and the ground truth locations of the facial landmarks or the human body joints. In order to be consistent and distinguish it from the mirror error we call it the alignment error. Formally,

$${}^{\mathrm{q}}e_{\mathrm{a}} = \frac{1}{K}\sum_{k=1}^{K} ||{}^{\mathrm{q}}\mathrm{x}_k - {}^{\mathrm{gt}}\mathrm{x}_k|| \tag{2}$$

where ${}^{\mathrm{gt}}\mathrm{x}_k$ is the ground truth location of the k-th point. In a similar way, we define the alignment error ${}^{\mathrm{p}}e_{\mathrm{a}}$ on the mirror image of the test sample. For simplicity in what follows when we use the term of alignment error e_a, we mean the alignment error in the original image.

Both Eq. 1 and Eq. 2 are absolute errors. In order to keep our analysis invariant to the size of the object in each image, we normalize them by the object size, i.e. s, the size of the body or the face. The size of the human body and the face are calculated in different ways and they are depicted when we use them.

2.2. Human pose estimation

Experiment setting In order to evaluate the mirroability of algorithms for human pose estimation, we focus on two representative methods, namely the Flexible Mixtures of Parts (FMP) method by Yang and Ramanan [27] and the Latent Tree Models (LTM) by Wang and Li [21]. The FMP is generally regarded as a benchmark method for human pose estimation and most of the recent methods are improved versions or variants of it. The one by Wang and Li [21] introduced latent variables in tree model learning that led to improvements. Both of them have provided source code which we used in our evaluation. Since it is not our main focus to improve the performance in a specific domain, we use popular state of the art approaches and evaluate them on standard datasets. We use three widely used datasets, namely the Leeds Sport Dataset (LSP), the Image Parse dataset [13] and the Buffy Stickmen dataset [6]. We use the default training/test split of the datasets. The number of test images on LSP, Parse and Buffy is 1000, 276 and 205 respectively. We trained both FMP and LTM models on LSP and only FMP model on Parse and Buffy. We emphasize that the training dataset is augmented with mirror images - this eliminates the training sample bias.

Overall performance difference We first compare the overall performance on the original test set and on the mirror set. We use the evaluation criterion proposed in [27] and also recommended in [1], namely the Percentage of Correct Keypoints (PCK). In order to calculate the PCK for each person a tightly-cropped bounding box is generated as the tightest box around the person in question that contains all of the ground truth keypoints. The size of the person is calculated as $s = \max(h, w)$, where h and w are the hight and width of the bounding box. This is used to normalize the absolute mirror error in Eq. 1 and the alignment error in Eq. 2. The results on Buffy, Parse and LSP are shown in Table 1, Table 2 and Table 3 respectively. As can be seen, there is no significant overall difference between the detection results on the original images and on their mirror images. The maximum difference of different methods on different datasets is around 1% while the average difference less than 1%.

Mirrorability The fact that the average performance on mirror images is similar to the average performance on the originals might be the root of the common belief that models produce more or less bilaterally symmetrical results. A closer inspection however reveals that this is not true. Let us first visualize the mirror error of individual body joints, i.e., $||{}^{q}x_k - {}^{p\to q}x_k||$ of both FMP and LTM on the LSP dataset. In Fig 2 we plot the mirror error (normalized by the body size in the example image) of the 1000 test images on each individual joint. As can be seen, there is a difference which in some cases it is quite large. For example on the elbows, feet and especially on the wrists ($\sim 18\%$ for FMP and $\sim 20\%$ for LTM). This result directly challenges the perception that the models give mirror symmetrical results. We reiterate that this is despite the fact that the overall performance is similar in the original and the mirror images and despite the fact that we have augmented the training set with the mirror images. This leads us to the conclusion that

Points	Head	Shou	Elbo	Wri	Hip	*Avg*
Original	96.9	97.3	91.1	80.8	79.6	89.1
Mirror	97.1	98.4	91.8	81.9	80.4	89.9

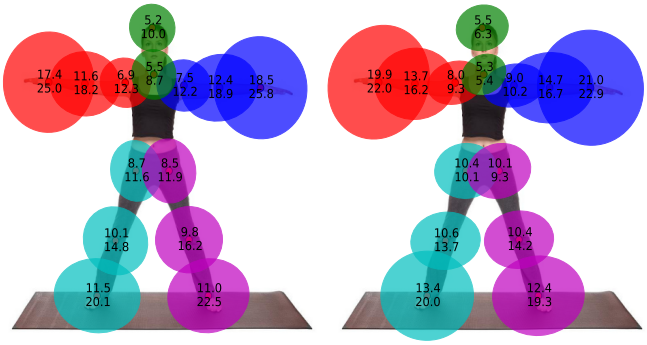
Table 1: PCK of FMP [27] on Buffy. A point is correct if the error is less than $0.2 * \max(h, w)$

Points	Head	Shou	Elbo	Wris	Hip	Knee	Ankle	*Avg*
Original	90.0	85.6	68.3	47.3	77.3	75.6	67.3	73.1
Mirror	90.0	86.1	67.6	46.3	76.8	74.6	68.5	72.8

Table 2: PCK of FMP [27] on Parse. A point is correct if the error is less than $0.1 * \max(h, w)$.

Points	Head	Shou	Elbo	Wris	Hip	Knee	Ankle	*Avg*
FMP Original	81.2	61.1	45.5	33.4	63.0	55.6	49.5	55.6
FMP Mirror	82.2	61.0	44.9	33.8	63.7	56.1	50.5	56.0
LTM Original	88.5	66.0	51.3	41.1	69.7	59.2	55.6	61.6
LTM Mirror	88.7	65.8	51.4	40.7	70.2	58.0	55.0	61.4

Table 3: PCK of FMP [27] and LTM [21] on LSP. A point is correct if the error is less than $0.1 * \max(h, w)$.



(a) Yang and Ramanan [27] (b) Wang and Li [21]

Figure 2: Visualization of mirror error (numbers on the upper) and alignment error (values on the lower) of body joints. The values are percentages of the body size. The radius of each ellipse represents the value of one standard deviation of the mirror error on the corresponding body joint.

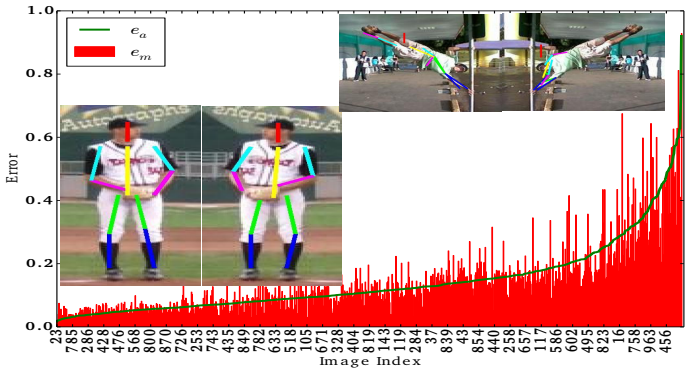
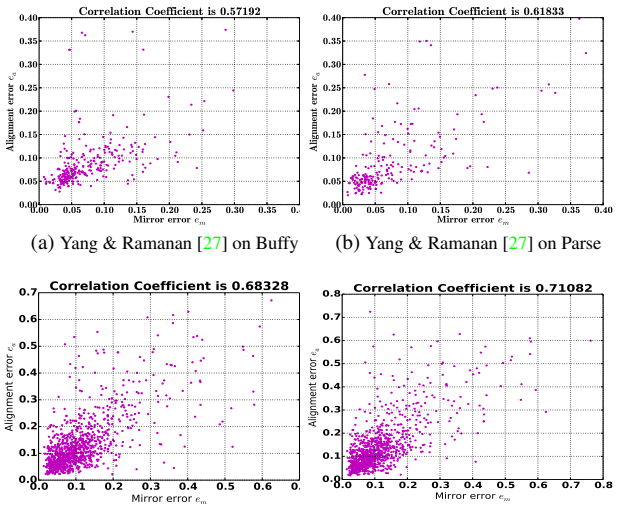


Figure 3: Mirror error and alignment error on LSP of LTM [21]. The x axis is the image indexes after sorting the alignment error in ascend. Two example images and their mirror images are shown, one with small mirror error and the other with large mirror error.



(a) Yang & Ramanan [27] on Buffy (b) Yang & Ramanan [27] on Parse

(c) Yang and Ramanan [27] on LSP (d) Wang and Li [21] on LSP

Figure 4: Correlation between the alignment error and mirror error. The correlation coefficients are shown above the figures.

the low mirrorability (i.e. large mirror error) is not the result of sample bias.

It is interesting to observe in Fig. 2 that the joints with large average mirror error are usually the most challenging to localize, that is they are the ones with the higher alignment error. This seems to indicate that there is correlation between the mirror error and the alignment error. In Fig. 3, as an example, we show the mirror error vs. the sorted sample-wise alignment error of LTM on LSP dataset. It is clear that the mirror error tends to increase as the image alignment error increases. Two examples of pairs of images are shown in Fig. 3 and the correlation between the sample-wise mirror error and the alignment error are shown them in Fig. 4. On all three datasets the mirror error has shown a strong correlation to the alignment error. For the smaller datasets, Buffy and Parse the correlation coefficient is around 0.6. On the larger LSP dataset, the correlation coefficient of both LTM and FMP is around 0.7. We can conclude that although the mirror error is calculated without knowledge of the ground truth, it is informative of the real alignment error in each sample.

2.3. Face alignment

Face alignment has been intensively studied and most of the recent methods [25, 30, 17] have reported close-to-human performance on face images "in the wild". Here, we look into the mirrorability of face alignment methods and how their error is correlated to the mirror error.

Experiment setting For our analysis we focus on the most challenging datasets collected in the wild, namely the 300W. It is created for Automatic Facial Landmark Detection in-the-Wild Challenge [15]. To this end, several popular data sets including LFPW [3], AFW [31] and HELEN [10] were re-annotated with 68 points mark-up and a new data set, called iBug, was added. We perform our analysis on a test set that comprises of the test images from HELEN (330 images), LFPW (224 images) and the images in the iBug subset (135 images), that is 689 images in total. The images in the iBug subset are extremely challenging due to the large head pose variations, faces that are partially outside the image and heavy occlusions. The test images are flipped horizontally to get the mirror images. We evaluate the performance of several recent state of the art methods, namely the Supervised Descent Method (**SDM**) [23], the Robust Cascaded Pose Regression (**RCPR**) [4], the Incremental Face Alignment (**IFA**) [2] and the Gaussian-Newton Deformable Part Model (**GN-DPM**) [20]. For SDM, IFA and GN-DPM, only the trained models and the code for testing is available - we use those to directly apply them on the test images. As stated in the corresponding papers, the IFA and GN-DPM were trained on the 300W dataset and the SDM model was trained using a much larger dataset. SDM, IFA and GN-DPM only detect the 49 inner facial points - our analysis on those methods is therefore based on those points only. For RCPR, for which the code for training is available, we retrain the model on the training images of 300W for the full 68 facial points mark-up. All those methods build on the result of a face detector - since most of them are sensitive to initialization, we carefully choose the *right* face detector for each one to get the best performance. More specifically, for the IFA and GN-DPM we use the 300W face bounding boxes and for SDM and RCPR we use the Viola-Jones bounding boxes, that is for each method we used the detector that it used during training. For the meth-

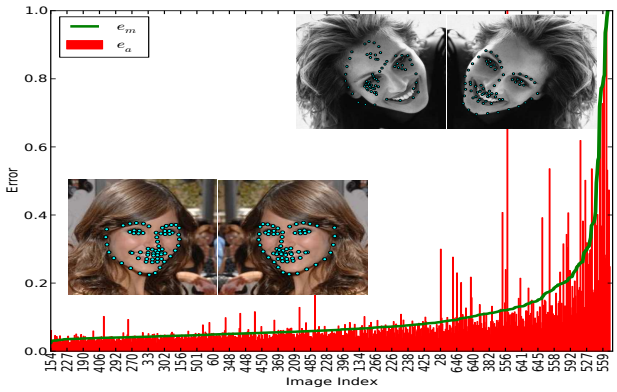


Figure 5: Mirror error and alignment error of RCPR [4] on 300W test images. Results are calculated over 68 facial points.

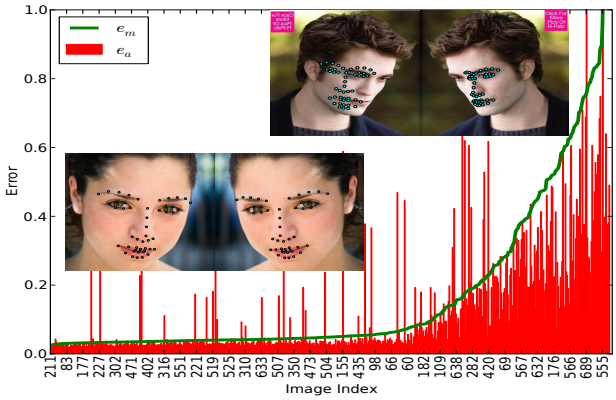
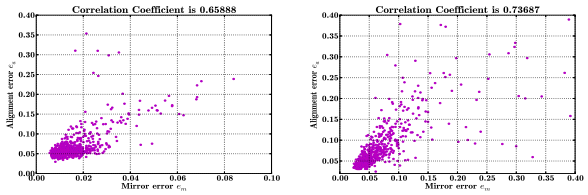


Figure 6: Mirror error and alignment error of GN-DPM [20] on 300W test images. Results are calculated over 49 inner facial points.

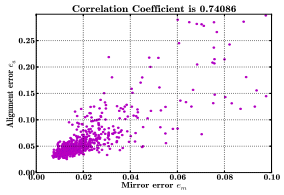
ods that use the Viola-Jones bounding boxes, we checked manually to verify that the detection is correct - for those face images on which the Viola-Jones face detector fails, we adjust the 300W bounding box to roughly approximate the Viola-Jones bounding box.

Mirrorability We calculated the mirror error and the alignment error for each of the 689 test samples in 300W for SDM, IFA, GN-DPM and RCPR. In Fig. 6 and Fig. 5 we show the errors for two of the algorithms, i.e., the GN-DPM and the RCPR. The former is a representative local-based method and the latter a representative holistic-based method. Similar results were obtained for SDM and IFA. In each figure, two pairs of example images are shown - one with low mirror error (lower left corner) and one with large mirror error (upper right corner). We sort the sample-wise alignment error in ascending order and plot it together with the corresponding sample mirror error. It is clear that although GN-DPM and the RCPR work in a very different way, for both the mirror error tends to increase as the

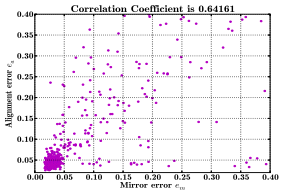


(a) SDM [23], 49P

(b) RCPR [4], 68P

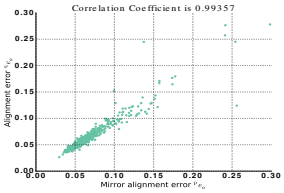


(c) IFA [2], 49P

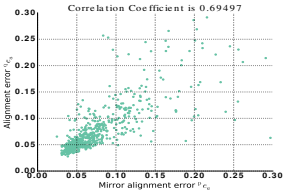


(d) GN-DPM [20], 49P

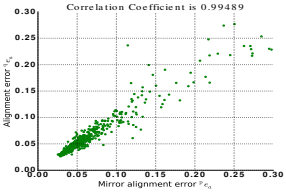
Figure 7: Correlation between the alignment error and the mirror error of various state of the art face alignment methods. The correlation coefficients are shown above the figures.



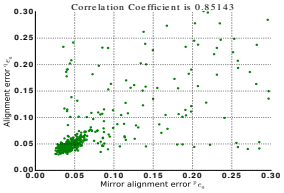
(a) SDM [23], 49P



(b) RCPR [4], 68P



(c) IFA [2], 49P



(d) GN-DPM [20], 49P

Figure 8: Correlation between alignment errors on original images and their mirror images.

alignment error increases. There are a few impulses in the lower range of the red curve, i.e., low ${}^{q}e_a$ and high e_m. This means that although the algorithm has small alignment error on the original samples it has large error on the mirror images, i.e., ${}^{q}e_a$ is high. There are three cases that result in high mirror error: 1) low ${}^{q}e_a$ and high ${}^{p}e_a$; 2) high ${}^{q}e_a$ and low ${}^{p}e_a$ (shown in Fig. 5 upper right corner); 3) high ${}^{q}e_a$ and high ${}^{p}e_a$ (shown in Fig. 6 upper right corner). In order to quantify this insight, we present the correlation between the mirror error and the alignment error in Fig. 7. In all of the four methods there is a strong correlation between the mirror error and the alignment error with correlation coeffi-

cients ranging from 0.64 to 0.74 - these are very high. We also plot the alignment error of the original images against that of their mirror images, i.e. ${}^{\mathrm{p}}e_a$ vs. ${}^{\mathrm{q}}e_a$, in Fig. 8. As can be seen, there is a strong correlation between the error on the mirror image and error on the original, with correlation coefficient from 0.69 to 0.99. We can conclude that 1) when one method fails to localize the landmarks accurately on one image, it is very likely it will fail on the mirror image as well; 2) the failures in original images and their mirror images are not symmetrically consistent, since the mirror error is also correlated to the alignment error. Finally, to assess the effect of randomness, we added small amount of Gaussian noises in the initializations either on the face bounding boxes (SDM and IFA) or on the initialization shape (GN-DPM and RCPR) the re-run the experiments starting from the new noise initializations in the original images and their mirror in the mirrored images. We obtained very similar results.

3. Mirrorability Applications

In the previous sections we have shown that one of the nice properties of the mirror error is that it is strongly correlated with the object alignment error, that is with the ground truth error. In this section we show how it can be used in two practical applications, namely for selecting difficult samples and for providing feedback in a cascaded face alignment method.

3.1. Difficult samples selection

For any computer vision task, including face alignment, it is generally accepted that some samples are relatively more difficult than others, that is the error of the algorithm on them is higher. However, it is very difficult to estimate a measure of how well the algorithm has performed on a given sample without knowledge of the ground truth. Such a measure would be very useful, for example in order to select a proper alignment model for a given dataset or to select which samples to annotate in an Active Learning scheme. Here, we show how the mirror error can be used for selecting difficult samples in the problem of face alignment. In order to do so we apply several methods (IFA, SDM, GN-DPM, RCPR) on the test images of the 300W and get the detection results. Then we sort the normalized mirror error e_m in descending order and select the first M samples as being the most difficult ones. We denote this set as S_{e_m}.

In order to evaluate whether the samples that we have selected in this way are truly 'difficult' we measure the similarity between the set containing those M selected samples and the set S_{e_a} that contains the M samples that have the largest alignment error e_a for each method. We use a measure that we call consistency which we define as the fraction of the common samples between the two sets, that is

$$\rho = \frac{|S_1 \cap S_2|}{M} \tag{3}$$

where $|S_1 \cap S_2|$ is the size of the intersection of S_1 and S_2. For each method i, we calculate two sets each containing M samples, i.e., $S^i_{e_m}$ and $S^i_{e_a}$. We set the value of M to 150. The chance rate is $\frac{M}{N}$, where M is the number of selected and N is the size of the dataset - in our case is $\frac{150}{689} \approx 0.22$.

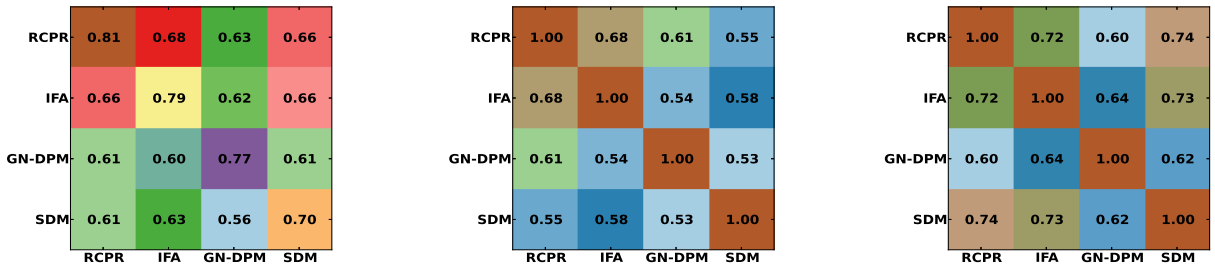
The pairwise consistency rate matrix of $S^i_{e_m}$ and $S^i_{e_a}$ is shown in Fig. 9a, where in a certain row we show the consistency between the $S^i_{e_m}$ of a certain method with the $S^i_{e_a}$ of all methods, including the method itself. Note that the diagonal does not contain ones, since $S^i_{e_m}$ are the M samples with the highest mirror error and $S^i_{e_a}$ the M samples with the highest alignment error. As it can be seen, the consistency between the two sets of samples for a specific method (i.e., the diagonal values) are all above 0.7 - the highest is 0.81 for RCPR. More interestingly, the consistency across different methods, i.e., the M samples selected according to e_a for a method in a certain row and the M samples selected according to e_m in a certain column is high, with values ranging from 0.56 to 0.68. This shows that the samples that we have selected are truly 'difficult', not only for the method employed in the selection process but also for the other face alignment methods. In other words this shows that the methods that we have examined have difficulties with the same images.

Second, we evaluate the consistency across different approaches, i.e., the consistency of 'difficult' samples found by different approaches. Thus, we calculate the pairwise consistency of $S^i_{e_m}$ of those methods as shown in Fig. 9b. The resulting values are clearly much higher than the chance value of 0.22. In Fig. 9c we depict the 'optimal' case where the ground truth, that is the alignment error itself, is used to calculate the pairwise consistency. We observe that the consistency calculated by our selection process is very close to the one calculated based on the ground truth. We can further conclude that:

- the difficulty of samples is shared by the different methods that we have examined.
- the difficult samples selected by the mirror error show high consistency across different approaches.

3.2. Feedback on cascaded face alignment

In recent years cascaded methods like SDM [23], IFA [2], CFAN [28] and RCPR [4] have shown promising results in face alignment on ordinary images and face sketches [26]. Although they differ in terms of the regressor and the features that they use in each iteration they all follow the same strategy. The methods start from one or several initializations of the face shape, that are often calculated from the



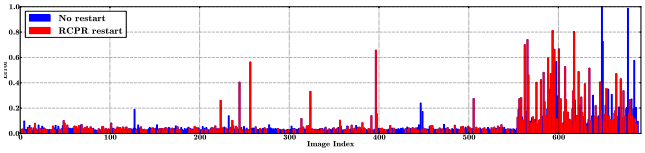
(a) ρ of $S_{e_a} \Leftrightarrow S_{e_m}$. (b) ρ of $S_{e_m} \Leftrightarrow S_{e_m}$. (c) ρ of $S_{e_a} \Leftrightarrow S_{e_a}$.

Figure 9: Consistency measure of 'difficult' samples detection, with $M = 150$.

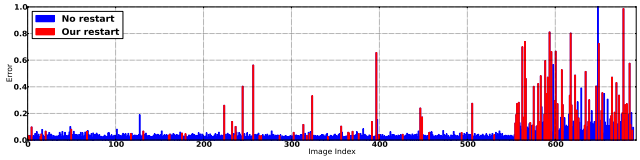
face bounding box, and then iteratively refine the estimation of the face shape by applying at each iteration a regressor that estimates the update of the shape. These methods are intrinsically sensitive to the initialization [4, 28] . As stated in [24], only initializations that are in a range of the optimal shape can converge to the correct solution. To address this problem, [5] proposed to use several random initializations and give the final estimate as the median of the solutions to which they convergence. However, having several randomly generated initializations does not guarantee that the correct solution is reached. The 'smart restart' proposed in [4] has improved the results to a certain degree. The scheme starts from different initializations and apply only 10% of the cascade. Then, the variance between the predictions is checked. If the variance is below a certain threshold, the remaining 90% of the cascade is applied as usual. Otherwise the process is restarted with a different set of initializations.

Here, we propose to use the mirror error as a feedback to close this *open* cascaded system. More specifically, for a given test image we first create its mirror image. Then we apply the RCPR model on the original test image and the mirror image and calculate the mirror error. If the mirror error is above a threshold we restart the process using different initializations, otherwise we keep the detection results. This procedure can be applied until the mirror error is below a threshold, or until a maximum number of iterations M is reached. In contrast to the original RCPR method that keeps only the results from the last set of initializations, we keep the one that has the smallest mirror error. This makes sense since new random initializations do not necessarily lead to better results than past initializations.

First we evaluate the effectiveness of our feedback scheme. Ideally, the restart will be initiated only when the current initialization is unable to lead to a *good* solution. Treating it as a two class classification problem we report results using a precision-recall based evaluation. A face alignment is considered to belong to the 'good' class if the mean alignment error is below 10% of the inter-ocular distance, otherwise, it is considered to belong to the 'bad' class



(a) Original RCPR restart scheme. Presion=0.25, Recall = 0.63.



(b) Our restart scheme. Precision = 0.65, Recall = 0.63.

Figure 10: Restart scheme of our method vs. RCPR [4] (best viewed in color).

- in the latter case a re-start is needed. The precision is the number of samples classified correctly as belonging to the 'bad' (positive) class divided by the total number of samples that are classified as belonging to the 'bad' class. Recall in this context is defined as the number of true positives divided by the total number of samples that belong to the bad class. For a fair comparison, we adjust our threshold on the mirror error (i.e. the threshold above which we restart the cascade with a different initialization) to get similar recall as the RCPR with smart re-start [4] gets using its default parameters. We note that our parameter can also be optimized by cross validation for better performance. As can be seen in Fig. 10, at a similar recall level, our proposed scheme has significantly higher precision (0.65 vs. 0.25) than that of RCPR 'smart re-start', this verifies that our method is more effective in selecting samples for which restarting initializations are needed.

Second, we evaluate the improvement in the face alignment that we obtain using our proposed feedback scheme. We compare to 1) RCPR without restart (RCPR-O), 2) RCPR with the smart restart of [4] (RCPR-S) and 3) other state of the art methods. We create two versions of our

Methods	*RCPR-F2*	*RCPR-F1*	RCPR-S	RCPR-O	SDM	IFA	GN-DPM	CFAN
49P	**5.35**	6.07	6.59	7.14	7.12	8.31	12.42	7.24
68P	**6.25**	7.11	7.42	7.73	-	-	-	7.72

Table 4: 49/68 facial landmark mean error comparison .

method. The first version, RCPR-F1, uses 5 initializations and at most two restarts - this allows direct comparison to the baseline method that uses the same number of initializations and restarts. The second version, RCPR-F2, uses 10 initializations and at most 4 times of restarts - this version produces better results and still has good runtime performance. We compare to SDM [23], IFA [2], GN-DPM [20] and CFAN [28] - all of those have publicly available software and report good results. The results of the comparison is shown in Table 4. We compare the normalized alignment error of the common 49 inner facial landmarks for all of these methods and the 68 facial landmarks whenever this is possible. On the challenging 300W test set, with our proposed feedback scheme, the RCPR method has the best performance compared to not only the original version of RCPR but also to all the other methods. Although good performance is obtained on the face alignment problem, we emphasize that the main focus of this work is to bring attention to the mirroability of object localization models.

4. Related Work

As a method that estimates the quality of the output of a vision system, our method is related to works like the meta-recognition [16], face recognition score analysis [22] and the recent failure alert [29] for failure prediction. Our method differs from those works in two prominent aspects (1) we focus on fine-grained object part localization problem while they focus on instance level recognition or detection. (2) we do not train any additional models for evaluation while all those methods rely on meta-systems. In the specific application of evaluating the performance of Human Pose Estimation, [9] proposed an evaluation algorithm, however, again such an evaluation requires a meta model and it only works for that specific application.

Our method is also very different from object/feature detection methods that exploit mirror symmetry as a constraint in model building [19, 12]. We note that our model does not assume that the detected object or shape appears symmetrically in an image - such an assumption clearly does not hold true for the articulated (human body) and deformable (face) objects that we are dealing with. None of the methods that we have exploited in this paper explicitly used the appearance symmetry in model learning. Our method only utilizes the mirror symmetry property to map the object parts between the original and mirror images.

Developing transformation invariant vision system has drawn much attention in the last decades. Examples are the rotation invariant face detection method [14] and the scale invariant feature transform (SIFT) [11], which handle efficiently several transformations including the mirror transformation. Recently, Gens and Domingos proposed the Deep Symmetry Networks [8] that use symmetry groups to represent variations - it is unclear though how the proposed method can be applied for object part localization. Szegedy *et al.* [18] has studied some intriguing properties of neural networks when dealing with certain artificial perturbations. Our method focuses on examining the performance of object part localization methods on one of the simplest transforms, i.e. mirror transformation, and drawing useful conclusions.

5. Conclusion and Discussion

In this work, we have investigated how state of the art object localization methods behave on mirror images in comparison to how they behave on the original ones. All of the methods that we have evaluated on two representative problems, struggle to get mirror symmetric results despite the fact that they were trained with datasets that were augmented with the mirror images.

In order to qualitatively analyse their behaviour, we introduced the concept of mirrorability and defined a measure called the mirror error. Our analysis let to some interesting findings in mirrorability, among which a high correlation between the mirror error and ground truth error. Further, since the ground truth is not needed to calculate the mirror error, we show two applications, namely difficult samples selection and cascaded face alignment feedback that aids a re-initialization scheme. It is also useful to simply take the mirror error as a measure for an object part localization method evaluation. We believe there are many other potential applications in particular in Active Learning.

The findings of this paper raise several interesting questions. Why some methods have shown better performance in terms of absolute mirror error, for example SDM is smaller and RCPR is bigger? Can the design of algorithms with low mirrorability error lead to algorithms with good ultimate performance? What kind of mirror error exhibit in different types of models: proctorial structures, deep neural networks, random forest models? We believe these are all interesting research problems for future work.

Acknowledgement

We thanks the anonymous reviewers for their constructive feedbacks. This work is partially supported by IP project REVERIE (FP-287723). H. Yang is sponsored by a CSC/QMUL joint scholarship and would like to thank Yichi Zhang, Negar Rostamzadeh, Mojtaba Khomami Abadi for useful discussion.

References

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 3

[2] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Incremental face alignment in the wild. In *CVPR*, 2014. 4, 5, 6, 8

[3] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *CVPR*, 2011. 4

[4] X. P. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. In *ICCV*, 2013. 1, 4, 5, 6, 7

[5] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. In *CVPR*, 2012. 7

[6] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari. 2d articulated human pose estimation and retrieval in (almost) unconstrained still images. *IJCV*, 99(2):190–214, 2012. 3

[7] J. R. Finnerty. Did internal transport, rather than directed locomotion, favor the evolution of bilateral symmetry in animals? *BioEssays*, 27(11):1174–1180, 2005. 1

[8] R. Gens and P. Domingos. Deep symmetry networks. In *NIPS*, 2014. 8

[9] N. Jammalamadaka, A. Zisserman, M. Eichner, V. Ferrari, and C. Jawahar. Has my algorithm succeeded? an evaluator for human pose estimators. In *ECCV*. Springer, 2012. 8

[10] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *ECCV*, 2012. 4

[11] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999. 8

[12] G. Loy and J.-O. Eklundh. Detecting symmetry and symmetric constellations of features. In *ECCV*. 2006. 8

[13] D. Ramanan. Learning to parse images of articulated bodies. In *NIPS*, 2006. 3

[14] H. A. Rowley, S. Baluja, and T. Kanade. Rotation invariant neural network-based face detection. In *CVPR*, 1998. 8

[15] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCV*, 2013. 4

[16] W. J. Scheirer, A. Rocha, R. J. Micheals, and T. E. Boult. Meta-recognition: The theory and practice of recognition score analysis. *T-PAMI*, 33(8):1689–1695, 2011. 8

[17] B. M. Smith, J. Brandt, Z. Lin, and L. Zhang. Nonparametric context modeling of local appearance for pose-and expression-robust facial landmark localization. In *CVPR*, 2014. 4

[18] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 8

[19] S. Tsogkas and I. Kokkinos. Learning-based symmetry detection in natural images. In *ECCV*, pages 41–54. Springer, 2012. 8

[20] G. Tzimiropoulos and M. Pantic. Gauss-newton deformable part models for face alignment in-the-wild. In *CVPR*, 2014. 4, 5, 8

[21] F. Wang and Y. Li. Beyond physical connections: Tree models in human pose estimation. In *CVPR*, 2013. 3, 4

[22] P. Wang, Q. Ji, and J. L. Wayman. Modeling and predicting face recognition system performance based on analysis of similarity scores. *TPAMI*, 29(4):665–670, 2007. 8

[23] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, 2013. 4, 5, 6, 8

[24] X. Xiong and F. De la Torre. Supervised descent method for solving nonlinear least squares problems in computer vision. *arXiv:1405.0601*, 2014. 7

[25] H. Yang and I. Patras. Sieving regression forests votes for facial feature detection in the wild. In *ICCV*, 2013. 4

[26] H. Yang, C. Zou, and I. Patras. Face sketch landmarks localization in the wild. *IEEE Signal Processing Letters*, 21(11):1321–1325, Nov 2014. 6

[27] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. *T-PAMI*, 35(12):2878–2890, 2013. 1, 3, 4

[28] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *ECCV*, 2014. 6, 7, 8

[29] P. Zhang, J. Wang, A. Farhadi, M. Hebert, and D. Parikh. Predicting failures of vision systems. In *CVPR*, 2014. 8

[30] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *ECCV*, 2014. 4

[31] X. Zhu and D. Ramanan. Face detection, pose estimation and landmark localization in the wild. In *CVPR*, 2012. 4