Distance Estimation of an Unknown Person from a Portrait

Xavier P. Burgos-Artizzu[1,2], Matteo Ruggero Ronchi[2], and Pietro Perona[2]

[1] Technicolor - Cesson Sévigné, France
[2] California Institute of Technology, Pasadena, CA, USA
xavier.burgos@technicolor.com, {mronchi,perona}@caltech.edu

Abstract. We propose the first automated method for estimating distance from frontal pictures of unknown faces. Camera calibration is not necessary, nor is the reconstruction of a 3D representation of the shape of the head. Our method is based on estimating automatically the position of face and head landmarks in the image, and then using a regressor to estimate distance from such measurements. We collected and annotated a dataset of frontal portraits of 53 individuals spanning a number of attributes (sex, age, race, hair), each photographed from seven distances. We find that our proposed method outperforms humans performing the same task. We observe that different physiognomies will bias systematically the estimate of distance, i.e. some people look closer than others. We expire which landmarks are more important for this task.

Keywords: Camera-subject distance, Perspective distortion, Pose estimation, Face recognition.

1 Introduction

Consider a standard portrait of a person – either painted or photographed. Can one estimate the distance between the camera (or the eye of the painter) and the face of the sitter? Can one do so accurately even when the camera and the sitter are unknown? These questions are not just academic – we have four applications in mind. First, faces are present in most consumer pictures; if faces could provide a cue to distance, this would be useful for scene analysis. Second, psychologists have pointed out that the distance from which a portrait is captured affects its emotional valence [1]; therefore, estimating this distance from a given picture would provide a cue to automate the assessment of its emotional valence. Third, estimating the distance from which master paintings were produced will provide art historians with useful information on art practices throughout the ages [2]. The fourth potential application is forensics: inconsistency in the distance from which faces were photographed may help reveal photographic forgeries [3].

The most informative visual cues for distance are stereoscopic disparity [4], motion parallax [5],[6] and structured lighting [7, 8]. However, we are interested in the case of a static monocular brightness picture, such as a painting hanging in a museum or a photograph in a newspaper, where none of these cues is available.

D. Fleet et al. (Eds.): ECCV 2014, Part I, LNCS 8689, pp. 313–327, 2014.

Fig. 1. Portrait pictures of a subject taken from 7 different distances ranging between 60 cm (left-most image) and 480 cm (right-most image). The effect of perspective, improperly called 'perspective distortion', is clearly noticeable. In portraits taken from a closer distance (left) the nose and mouth appear bigger, the ears are partially occluded by the cheeks and the face appears longer. This systematic deformation in the image plane is related to distance. We explore whether, and how accurately, the distance from which the portrait was taken may be estimated from the image when both the person and the camera are unknown.

The most reliable remaining cue is object familiarity [9]; however, there are several obstacles to a straightforward use of this cue. First, if the camera is unknown one does not have calibration parameters, which rules out straightforward use of the distance of known points, such as the distance between the pupils. Second, when the sitter is unknown only statistical, rather than exact, knowledge of the 3D shape of the object is available. However, it is known that one image of a constellation of at least five 3D points whose mutual position is known is sufficient both for camera calibration and pose computation [10], and therefore one would expect that some useful signal is available, see Fig. 1.

In this paper we study the feasibility and accuracy of automatically estimating the distance of a person from a camera, using a single 2D frontal portrait image without requiring any prior knowledge on the camera used or the person being photographed. Our approach is to first detect automatically facial features and then estimate distance from their mutual positions in the image. Our main contributions are:

1. A novel approach for estimating the camera-head distance from a single 2D portrait photograph when both the camera and the sitter are unknown. Our method yields useful signal and outperforms humans by 16%, see Fig. 6.
2. The introduction of a new dataset of portraits, *Caltech Multi-Distance Portraits* (**CMDP**), composed of 53 subjects belonging to both sexes, a variety of ages, ethnic backgrounds and physiognomies. Each subject was photographed from seven different distances and each portrait manually labeled with 55 keypoints over the head and face. The dataset is available online.
3. In-depth analysis and discussion of the feasibility of the proposed approach. We study two different variants of the task and analyze what are the most important input visual cues. We compare our method's performance using machine estimated landmarks vs. ground-truth landmarks. Finally, we also compare with the performance of human observers. Interestingly, we found that the main source of error for both humans and our method is the variability of physiognomies.

2 Related Work

Estimating the pose of a human head from an image was explored in [11, 12]. The literature focuses on the estimation of the three degrees of freedom (DOF) - yaw, pitch and roll - under the assumption that the human head can be modeled as a disembodied rigid object. Knowledge of the intrinsic camera parameters or depth information is required.

Psychophysics experiments [13, 14] show that human face recognition performance can be impaired by perspective transformation. As one might expect, the severity of this deficit depends on the difference between the amount of 'perspective distortion' at the learning and testing phases. They also established that both global perspective information and local image similarity features such as ears, eyes, mouth or nose play a fundamental role in this task. Their conclusion is that perspective distortion impairs face recognition, similarly to other visual cues such as lighting and head orientation. This poses the question of whether perspective distortion or, equivalently, distance may be estimated.

Psychologists [1] observed that portrait photographs taken from within personal space elicit lower investments in an economic trust game and lower ratings of social traits such as strength, attractiveness or trustworthiness. These findings could not be explained by width-to-height ratio, explicit knowledge of the camera distance or typicality of the presented faces, thus suggesting the existence of a facial cue influencing social judgments as a function of interpersonal distance. They suggest that there is an "optimal distance" at which portraits should be taken. This idea of choosing the optimal viewpoint and distance to subject is also known to be of great importance in traditional portraiture [2].

To our knowledge, Flores et al. [15] are the first to propose a method that recovers camera distance from a single image of a previously unseen subject. Their work is based on the Efficient Perspective n-Point algorithm (EPnP) [16], a non-iterative solution to the perspective n-point problem for pose estimation of a calibrated camera given n 3D-to-2D point correspondences. The main difference with our work is that this approach is based on explicit computation of 3D information; therefore it requires 3D models of heads. We argue this is an unnecessary complication. Moreover, Flores et al.'s method is not fully automated and requires hand-annotated landmarks on the test image.

In contrast with all prior work, we propose to train and test in image space without the need for 3D head features or pose information. Furthermore, no calibration or knowledge of camera parameters is needed. Finally, thanks to the recent improvement of automatic facial landmark estimation [17–21], our method is fully automated; it uses automatically estimated landmarks instead of manual annotations.

3 Caltech Multi-Distance Portraits Dataset

We collected a novel dataset, the *Caltech Multi-Distance Portraits* (**CMDP**). This collection is made of high quality frontal portraits of 53 individuals against

a blue background imaged from seven distances spanning the typical range of distances between photographer and subject: 60, 90, 120, 180, 240, 360, 480 cm, see Fig 1. For distances exceeding 5m, perspective projection approaches a parallel projection (the depth of a face is about 10cm), therefore no samples beyond 480cm were needed. Participants were selected among both genders, different ages and a variety of ethnicities, physiognomies, hair and facial hair styles, to make the dataset as heterogeneous and representative as possible.

Table 1. Diversity in the *Anonymous Portrait Faces* dataset. Individuals may belong to multiple categories.

Category	Number of Subjects	Percentage
African-American	4	7.5%
Asian	5	9.4%
Caucasian	36	68.2%
Latino	8	15%
Female	7	13.2%
Male	46	86.8%
With Facial Hair	13	24.5%
With Occlusions	11	20.7%

Pictures were collected with a Canon Rebel Xti DSLR camera mounting a 28-300mm L-series Canon zoom lens. Participants standing in front of a blue background were instructed to remain still and maintain a neutral expression. The photographer used a monopod to support the camera-lens assembly. The monopod was adjusted so that the height of the center of the lens would correspond to the bridge of the nose, between the eyes. Markings on the ground indicated seven distances. After taking each picture, the photographer moved the foot of the monopod to the next marking, adjusted the zoom to fill the frame of the picture with the face, and took the next picture. This procedure resulted in seven pictures (one per distance) being taken within 15-20 seconds. Images were then cropped and resampled to a common format. The lens was calibrated at different zoom settings to verify the amount of barrel distortion, which was found to be very small at all settings, and thus left uncorrected. Lens calibration was then discarded and not used further in our experiments.

As the camera approaches the subject the relationship of the size of the picture of the main parts of the face changes (Fig. 1). It is important to clarify that this 'perspective distortion' is not a lens error (this was verified, as explained in the previous paragraph): it arises from the projection of the three dimensional world into a two dimensional image and is easily observable with our own eyes. We could have used any other lens, including one with fixed focal length, or a pinhole camera, and there would have been no difference in the amount of 'perspective distortion' measured at a given distance (that is, assuming that

the lens has no internal flaws or distortion). Using a lens with a shorter focal length and wider field of view will result in a coarser pixel sampling of the face, but the perspective geometry and proportions would only depend on distance, or, equivalently, on the visual angle subtended by the face. Regardless of the lens used, crops of two images taken from the same distance would be identical, apart from sampling resolution. We used a zoom lens to obtain maximum pixel resolution at all distances.

3.1 Annotating CMDP

All images in the dataset were manually annotated with 55 facial landmarks distributed over and along the face and head contour, see Fig. 2(a). The location of our landmarks is very different from landmark positions typically used in the literature, more focused towards the center and bottom of the face, as for example Multi-pie [22] format, Fig. 2(b). We purposely wanted to have landmarks around the head contour (in green) and all around the face (in red), to sample a larger area of the face.

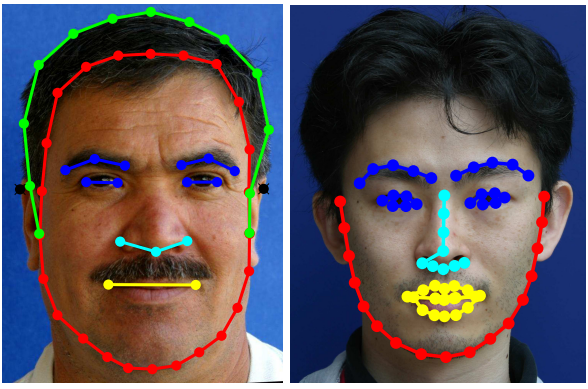


Fig. 2. Our 55 face landmarks (left) compared to the 68 facial keypoints in Multi-pie format [22] (right). With landmarks around the hair line and top of the forehead our landmarks allow to test whether these regions provide useful signal, despite their intrinsic variability.

The dataset was annotated by three different human annotators. Portraits from the same subject were always annotated by the same annotator in sequence, minimizing the variance in the location of landmarks between pictures at different distances. To check consistency of annotations, we doubly annotated several images from different subjects. Annotators are very consistent, showing an average disagreement between them less than 3% of the interocular distance, and not varying much across distances, see Fig 4.

4 Problem Formulation

The goal is to estimate the camera-head distance from a single 2D portrait photograph when both the camera and the sitter are unknown. From this initial problem formulation, we derive two different tasks:

1. Sorting the seven images belonging to a single previously unseen subject according to their distance.
2. Estimating the distance from which a single image of a previously unseen subject was taken.

While the difference between the two might seem subtle, it affects the entire procedure. Firstly, from a machine learning point of view, the former is a classification task, while the latter is a pure regression problem, meaning that feature normalization schemes and error metrics will be different in each case.

Secondly, pure regression is a much harder task. Since the person has never been seen before, it is difficult to account for his/her physiognomy. For example, a person with a round face or a big nose will often appear closer than a squared face with a small nose, see Fig. 3.

Fig. 3. Estimating the relative distance of previously unseen subjects is a difficult task. Consider these portraits. Their physiognomy confuses human annotators, which have a tendency to pick the left image (d=240cm) as the closest one, while the right hand side one (d=180cm) was closer.

In fact, while humans are able to perform the first task rather accurately, see Fig. 6, they are completely unable to perform the second task. Part of the reason is the well known fact that humans are better at relative judgments, rather than estimating absolute values. Another reason may be that having access to several pictures of the same subject allows to ignore physiognomy and focus on the important signal. For real-life applications the regression task is far more relevant; we use the ordering task exclusively to benchmark our method against human performance and guide our thoughts.

Error Metrics: In the re-ordering problem, we measure for each portrait the probability of being correctly classified into its distance category (from 1 to 7). In the regression task we measure both the Pearson correlation coefficient (Corr) and the coefficient of determination (R^2) between prediction and ground truth distance on all 7 images of the test subject:

$$Corr(sbj) = \frac{COV(gt(sbj), pred(sbj))}{std(gt(sbj)) * std(pred(sbj))}, R^2(sbj) = 1 - \frac{(gt(sbj) - pred(sbj))^2}{\left(gt(sbj) - \overline{gt}\right)^2}$$

Where the terms *gt(sbj)* and *pred(sbj)* are respectively the ground-truth and predicted distances of each picture belonging to the subject being evaluated and $\overline{gt}$ is the average of all ground truth distances.

5 Method

We use the position of the face's landmarks to capture the 2D shape of the face and therefore measure how much it changes with distance. Input landmarks can be both the result of manual annotations or the output of a landmark estimation algorithm. After computing the facial landmarks, we apply a supervised learning approach, see Sec 5.2. A subset of the subjects in the dataset are used to train a regressor capable of mapping the shape of their face at different distances to the ground truth distances. Then, the performance of the learned regressor is evaluated on the remaining subjects in the dataset according to each of the tasks defined in the previous Section.

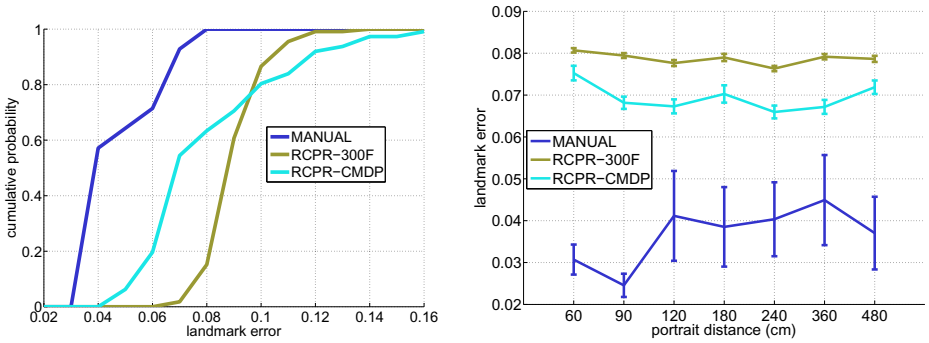
5.1 Facial Landmarks

Encouraged by the recent success of facial landmark estimation approaches, we decided to benchmark its feasibility for this task. We use *Random Cascaded Pose Regression (RCPR)* [20], due to its performance, speed and availability of code.

We trained RCPR on 70% of the individuals in our CMDP dataset (259 images in total), with the same parameters as in the original publication. When applied to the remaining 30% of our dataset, RCPR yields an average landmark error of 6.9% and a 16% failure rate, see Fig 4(a). Errors are measured as the average landmark distance to ground-truth, normalized as a percentage with respect to interocular distance. A failure is an average error above 10%, as in [20].

We also trained RCPR on the more exhaustive 300-Faces-in-the wild dataset [21] which contains more than 2K faces taken from previously existing datasets and re-annotated following Multi-Pie 68 landmarks [22] convention, see Fig 2(b). To compare its result on our test images, we only evaluate it on the 22 set of landmarks our convention shares with Multi-Pie format. This version of RCPR, applied to the same 30% subset of subjects achieves an average error of 7.8%, but with a much lower failure rate (4%), see Fig 4(a).

Both RCPR versions are still far from human performance, struggling slightly more with faces from both distance extremes, less common in face recognition datasets, see Fig 4(b). The distribution of errors by landmarks reveals that RCPR trained on CMDP struggles particularly with the head contour and the ears due to their inherent variability, while RCPR trained on 300F struggles with the nose and eyebrows. However, both versions still have a low number of failure cases and therefore these issues affect only slightly the final performance of distance estimation when compared with using ground-truth landmarks, as shown in next Section.



(a)Cumulative probability of errors (b)Error as a function of face distance

Fig. 4. Landmark estimation error. Human annotators are very consistent, showing a low disagreement (3% average), and not varying much across distances. Training RCPR using images from CMDP achieves good average performance except for its high number of failures (16%), struggling in close-range images. RCPR trained on 300-Faces yields slightly worse average performance while with a lower number of failures (4%). A failure is an average error above 10%, as in [20].

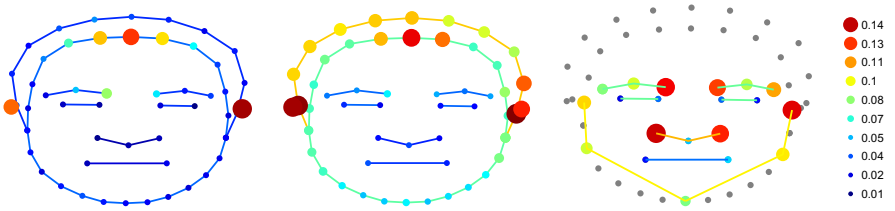
5.2 Proposed Approach

After collecting the facial landmarks, we use them as input to learn a regressor that maps face shapes to their ground-truth distances. More specifically, shape $\mathcal{S}$ is represented as a series of P landmark locations $\mathcal{S} = [(x_p, y_p)|p \in 1..P \wedge x, y \in \mathbb{R}]$. For each subject $i \in 1..N$ we dispose of seven different shape vectors associated to each one of the $d \in 1..7$ different distance images, S_d^i. The goal is to learn a robust mapping from each one of the seven shapes to their respective distance: $f : \mathcal{S}_d^i \mapsto \mathbb{R}$.

Shape Vector Normalization: Due to the heterogeneity of face physiognomies contained in our dataset, a prior normalization step of the face shapes is crucial to learn a robust mapping. First, we standardize all portraits using the individual shape vectors $\mathcal{S}_d^i$ cropping the image around the face and removing scale and rotation variations. Then, we propose two different normalization schemes for each one of the tasks defined previously.

In the re-ordering task, we can compute the average subject face shape across all seven distances ($\overline{\mathcal{S}}^i = \frac{1}{7}\sum_{d=1}^{7} \mathcal{S}_d^i$) and use it to normalize each shape, subtracting the mean from the landmark's position ($\mathcal{S'}^i_d = \mathcal{S}_d^i - \overline{\mathcal{S}}^i$). This filters out the variations in the shape of the face due to the physiognomy of the individual, leaving only the changes due to perspective distortion.

In the case of the regression task, at test time we only have access to one shape $\mathcal{S}_d^i$ at a time. During training, however, we can compute the average shape for each distance ($\overline{\mathcal{S}}_d = \frac{1}{N}\sum_{i=1}^{N} \mathcal{S}_d^i$). These average shapes can then be used to codify the current shape as the concatenation of the differences between $\mathcal{S}_d^i$ and each one of the d average faces $\overline{\mathcal{S}}_d$: ($\mathcal{S'}^i_d = \langle \mathcal{S}_d^i - \overline{\mathcal{S}}_{d=1}, .., \mathcal{S}_d^i - \overline{\mathcal{S}}_{d=7} \rangle$).



(a)Human errors (b)RCPR-CMDP errors (c)RCPR-300F errors

Fig. 5. Individual landmark errors (blue=low average error, red=high average error). (a) Humans concentrate their disagreement on the forehead (telling where it ends is somewhat subjective) and ears (which can be occluded by hair and excessive distortion). (b) RCPR trained on our faces struggles with hair and face contours. (c) RCPR trained on 300W-Faces struggles the most with the eyebrows and chin. Gray points signify non-existence of the fiducial due to use of Multi-Pie convention.

The effect of each one of these normalization schemes on performance is presented in Fig 9(b), compared also to no normalization at all. Each step improves performance significantly. It is evident that being able to average out with respect to the subject's shape makes a big difference, even compared to our distance normalization scheme.

Inverse distance: In practice, inverse distance is preferred to avoid the saturation of the signal after a certain value of distance (i.e. the difference in the measured distortion becomes negligible with respect to the change in distance).

Learning algorithm: We train a multivariate linear regressor to learn the mapping from the normalized shapes ${\mathcal{S}'}_d^i$ onto the inverse distance of a face as a weighted linear combination of the P landmark locations: $(\sum_{p=1}^{P} \mathbf{w_p} {\mathcal{S}'}_d^i(x_p, y_p))$. We tried several other regression/classification methods but none improved results w.r. to simple linear regression. This may be due to the the relatively small number of training examples, see Supp. Material for more info.

Regression vs. classification: For the classification task, we sort the values the regressor outputs for each of the 7 images belonging to the same subject and compare it against ground-truth distance ordering.

6 Results

We now discuss the results of our method on the re-ordering and pure regression tasks. We benchmark three variants of our method depending on the nature of the input landmarks: using 1) Ground-truth landmarks (MANUAL), 2) Landmarks from RCPR trained on our CMDP images (RCPR-CMDP) and 3) Landmarks from RCPR trained on 300-Faces in the wild (RCPR-300F).

All reported results are obtained using 70% of the subjects for training and the remaining 30% for testing (the same train/test set as that used to train RCPR-CMDP), except in Fig 9 where cross-validation runs are used. Variance is shown

as standard errors. In Sec 6.3 we show examples of how subject physiognomy affects performance. Further analysis is available in Supp. Material.

6.1 Re-ordering Task

Figure 6 shows the performance on the re-ordering task. Apart from the three variants of our method, we also plot the result obtained by humans asked to perform the exact same task. We developed a specific GUI and asked a group of 5 people of different levels of computer vision expertise to sort a random permutation of all 7 pictures of a subject based on their conveyed distance. Each person annotated at least 10 different subjects (70 images in total).

The ground-truth landmarks based variant (MANUAL) outperforms human performance by 16%, while the automatic based ones (RCPR-CMDP and RCPR-300F) are slightly behind by 3% and 25% respectively. Closer faces appear to be much easier to classify than distant ones because of their unusual and disproportioned geometry. This has been confirmed by the human subjects of the study, stating their difficulty in telling apart images in the middle distance-range.

We find these results very encouraging. Our best variant outperforms human capabilities in the classification task, correctly reordering an average of 81% of the faces when random chance is merely 15%. The same method using machine estimated landmarks still classifies correctly 62% of the images, and could very likely be improved just by increasing the availability of training examples.

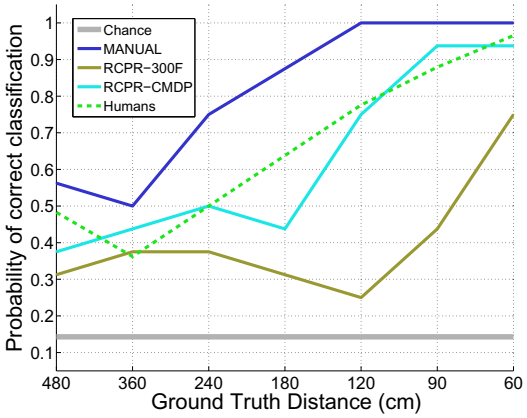
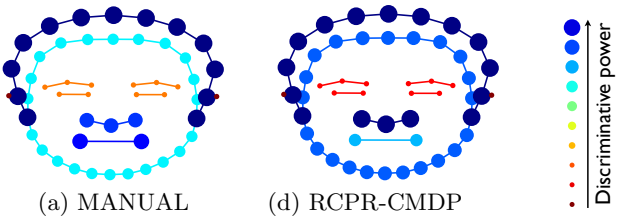


Fig. 6. Main results of our approach on the re-ordering task, measured as the probability of correctly ordering portraits of a subject according to their distance. Our methods using manual landmarks outperforms humans by 16%, while using RCPR-CMDP performance is virtually identical (lower by 3%).

Figure 7 shows which landmarks are most discriminative for the re-ordering task using both MANUAL and RCPR-CMDP input. We measure how well each landmark group (head contour, face contour, eyes, nose, mouth) compares to best performance when only that particular group is used. For both MANUAL and RCPR-CMDP, best results are achieved using the head contour and the nose, while the eyes seem to be the least useful.



(a) MANUAL (d) RCPR-CMDP

Fig. 7. Input landmarks discriminative power on the re-ordering task, measured as how well the method performs when incorporating those input landmarks into the learning, ranging from most discriminative (big blue dot) to least discriminative (small red dot). For both MANUAL and RCPR-CMDP, the most useful landmarks are the facial/head contours and the nose.

6.2 Regression Task

Figure 8 shows the results on the regression task. There is a strong correlation between ground-truth distances and predictions of our method. MANUAL achieves 75% correlation with a coefficient of determination of $R^2 = .5$, while RCPR-CMDP and RCPR-300F achieve 65% and 45% correlation and $R^2 = .48$ and .46 respectively. All variants seem to struggle more with the larger distances, as noticeable from the higher variance and greater distance to ground truth. This is an expected result considering the lower effect of perspective differences between two images taken from afar.

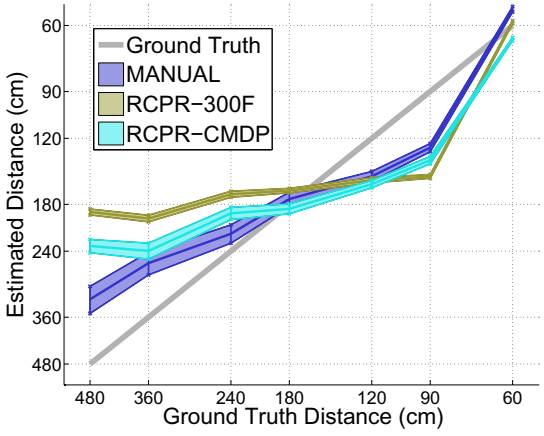
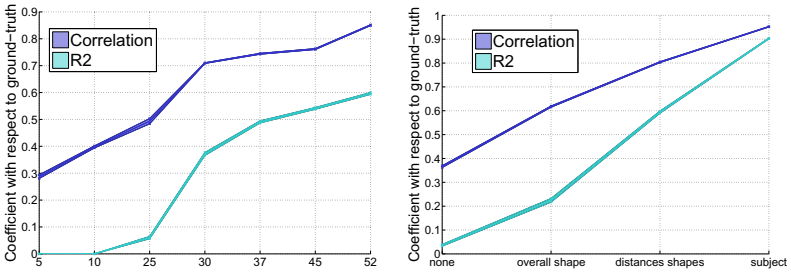


Fig. 8. Main results of our approach on the regression task, measured as the distance with ground truth distance. Using MANUAL landmarks achieves 75% correlation with a coefficient of determination of $R^2 = .5$, while RCPR-CMDP and RCPR-300F achieve 65% and 45% correlation and $R^2 = .48$ and .46 respectively.

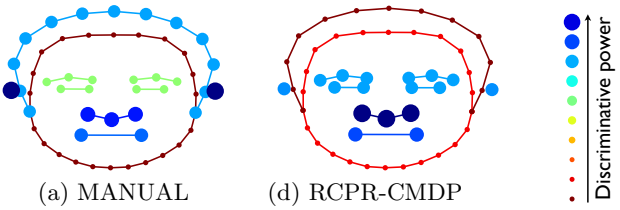
As expected, directly estimating the distance of an unknown face proved to be a harder task. Nonetheless, a correlation of 75% with ground-truth indicates that the method is learning well. Furthermore, increasing the amount of training data results in a peek of correlation up to 85%, see Figure 9(a), with no apparent saturation of performance, suggesting that with more data performance could be close to that desired for real-life applications. Overall, our experiments suggest that the distance of a face may be estimated from an uncalibrated 2D portrait.



(a) Subjects used for training (b) Normalization scheme

Fig. 9. Parameters evaluation. Results computed using cross-validation runs for robustness. (a) Result of increasing the number of training subjects on the regression task using MANUAL landmarks. With each added subject, the performance continues to grow with no saturation. (b) Result of the different normalization approaches presented in Sec. 5.2 using 52 training subjects in a leave-one-out cross validation scheme. Normalizing the shape of a subject's face using his own average shape across all distances achieves best performance.

Figure 10 shows which landmarks are most discriminative for the regression task using both MANUAL and RCPR-CMDP input. We measure how well each landmark group (head contour, face contour, eyes, nose, mouth) compares to best performance when only that particular group is used. For both MANUAL and RCPR-CMDP, most discriminative group is once again the nose. This finding agrees with human annotators, which consistently reported during the reordering psychophysics experiments the use of the deformation in a subject's nose as their main visual cue for the task.



(a) MANUAL (d) RCPR-CMDP

Fig. 10. Input landmarks discriminative power on the regression task, measured as how well the method performs when incorporating those input landmarks into the learning, ranging from most discriminative (big blue dot) to least discriminative (small red dot). For both MANUAL and RCPR-CMDP, the most useful group of landmarks is the nose. For MANUAL, the head contour is once again very discriminative.

Looking at Figures 7 and 10 together is very informative. They show that as we suspected, head and facial contours are extremely important for this task, which explains why RCPR using our landmark convention works far better that RCPR using Multi-Pie convention, which does not have landmarks around head contour. The differences between both figures tells us what parts of the face

vary the most across individuals, defining most important cues for physiognomy. Take the facial contour for instance: it switches from most useful in re-ordering task to least useful in regression. This is natural; if one knows the shape of a subject's face (re-ordering) it can be very useful to watch how it gets deformed by perspective. However, not being able to tell physiognomy apart from perspective (regression) makes those landmarks become useless.

6.3 Physiognomy

A final interesting observation regards physiognomy. Throughout all of the experiments we observed that physiognomy of people turned out to be one of the key factors for performance, both for human observers and for our algorithm. In fact, some people appear to be systematically closer than others exclusively due to the shape of their face, Figure 3. Therefore we discussed in Section 5 normalization schemes discarding physiognomy and preserving the signal from perspective distortion.

Accordingly, we have found that the accuracy of the method increases when we normalize using the subject's own average shape across the pictures at all distances, see Fig 9(b). However this subject-specific normalization is only applicable in the re-ordering task, where we can legitimately assume the availability of information on the subject. This has no practical bearing in the regression task where the person being portrayed is unknown.

We asked whether we could derive information on subject's physiognomy by observing the results of our method and if this could shed light on what specific attributes of a human face are most likely to bias distance estimates. We measured for all the faces in the dataset their average bias in the estimated distance over several runs with different training-test set combinations and show our findings in Figure 11. Besides a subjective feeling of roundness for the over estimated faces (judged closer by the algorithm) no evident pattern was found so far, see Supplementary Materials. Estimating physiognomy from a single picture is, thus, an open question.

Figure 12 shows an example output of our algorithm for a subject whose predicted distances are close to ground truth.

Fig. 11. Example of how physiognomy biases distance estimation. (TOP) Ten most under-estimated subjects. (BOTTOM) Ten most over-estimated subjects.

Fig. 12. Example output of the algorithm on the ordered images of a subject (ground-truth from left to right = 60, 90, 120, 180, 240, 360, 480 cm)

7 Conclusions

We proposed the first method for estimating automatically the distance from which a face was photographed. We assume that we have a single frontal photograph, where both the person and the camera are unknown. The method is based on two steps: first, estimating the position of a number of facial landmarks; second, estimating from their relative position the inverse distance by regression.

We find that the method is reasonably accurate. When using manually annotated landmarks as input, it outperforms relative depth judgments obtained from human observers. Furthermore, we find that performance does not suffer much when the method is fully automated with machine-based face landmark estimation. The fully automated method can estimate absolute distance, which human observers are unable to do. As expected, distance estimates beyond 3m, where perspective projection approaches parallel projection, are much noisier than distance estimates in the 0.5-2m range.

An interesting finding is that the main source of error is the variability of physiognomies. Some people appear to be systematically closer than others because their face is shaped differently. Once one normalizes for physiognomy the accuracy of the method increases about 30%; this has no practical bearing when the person being portrayed is unknown, and therefore it is impossible to normalize for physiognomy.

Recovering the distance of a face has a number of applications: as an additional cue to depth in scene analysis, as an indicator of the possible emotional valence of the picture [1], as a tool to study portraiture in classical paintings, and as a tool for forensic analysis of images [3]. Our experiments are encouraging, and are sufficient as a proof of principle to demonstrate feasibility. However, they indicate that accuracy would be significantly better if a much larger training set was available. It is intuitive that such a dataset should include a representative range of facial expressions, as well as a range of viewpoints.

Acknowledgments. This work is funded by ONR MURI Grant N00014-10-1-0933 and NASA Stennis NAS7.03001.

References

1. Bryan, R., Perona, P., Adolphs, R.: Perspective distortion from interpersonal distance is an implicit visual cue for social judgments of faces. PloS One 7(9), e45301 (2012)

2. Perona, P.: A new perspective on portraiture. Journal of Vision 7(9), 992 (2007)
3. Farid, H.: Image forgery detection. IEEE Signal Processing Magazine 26(2), 16–25 (2009)
4. Wheatstone, C.: Contributions to the physiology of vision. Part the first. On some remarkable, and hitherto unobserved, phenomena of binocular vision. Philosophical Transactions of the Royal Society of London 128, 371–394 (1838)
5. Gibson, E.J., Gibson, J.J., Smith, O.W., Flock, H.: Motion parallax as a determinant of perceived depth. Journal of Experimental Psychology 58(1), 40 (1959)
6. Rogers, B., Graham, M., et al.: Motion parallax as an independent cue for depth perception. Perception 8(2), 125–134 (1979)
7. Scharstein, D., Szeliski, R.: High-accuracy stereo depth maps using structured light. In: CVPR (2003)
8. Microsoft: Corp redmond wa. Kinect for xbox 360
9. Gogel, W.C.: The effect of object familiarity on the perception of size and distance. The Quarterly Journal of Experimental Psychology 21(3), 239–247 (1969)
10. Triggs, B.: Camera pose and calibration from 4 or 5 known 3D points. In: The Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 1, pp. 278–284. IEEE (1999)
11. Murphy-Chutorian, E., Trivedi, M.M.: Head pose estimation in computer vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(4), 607–626 (2009)
12. Fanelli, G., Dantone, M., Gall, J., Fossati, A., Van Gool, L.: Random forests for real time 3D face analysis. Int. J. Comput. Vision 101(3), 437–458 (2013)
13. Liu, C.H., Chaudhuri, A.: Face recognition with perspective transformation. Vision Research 43(23), 2393–2402 (2003)
14. Liu, C.H., Ward, J.: Face recognition in pictures is affected by perspective transformation but not by the centre of projection. Perception 35(12), 1637 (2006)
15. Flores, A., Christiansen, E., Kriegman, D., Belongie, S.: Camera distance from face images. In: Bebis, G., et al. (eds.) ISVC 2013, Part II. LNCS, vol. 8034, pp. 513–522. Springer, Heidelberg (2013)
16. Lepetit, V., Moreno-Noguer, F., Fua, P.: Epnp: An accurate o(n) solution to the pnp problem. International Journal of Computer Vision 81(2), 155–166 (2009)
17. Saragih, J., Lucey, S., Cohn, J.F.: Deformable model fitting by regularized landmark mean-shift. IJCV 2(91), 200–215 (2011)
18. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localiz. in the wild. In: CVPR (2012)
19. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. In: CVPR (2012)
20. Burgos-Artizzu, X.P., Perona, P., Dollár, P.: Robust face landmark estimation under occlusion. In: ICCV (2013)
21. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: The first facial landmark localization challenge. In: ICCV-Workshop (2013)
22. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. In: FG (2008)