Machine Learning (2019) 108:511–534
https://doi.org/10.1007/s10994-018-5756-7

Automatic face recognition with well-calibrated confidence measures

Charalambos Eliades[1] · Ladislav Lenc[2,3] · Pavel Král[2,3] · Harris Papadopoulos[1]

Received: 26 February 2018 / Accepted: 1 August 2018 / Published online: 20 August 2018

Abstract

Automatic face recognition (AFR) has gained the attention of many institutes and researchers in the past two decades due to its wide range of applications. This attention resulted in the development of a variety of techniques for the particular task with a high recognition accuracy when the environment is well-controlled. In the case of moderately controlled or fully uncontrolled environments however, the performance of most techniques is dramatically reduced due to the much higher difficulty of the task. As a result, the provision of some kind of indication of the likelihood of a recognition being correct is a desirable property of AFR techniques in many applications, such as the detection of wanted persons or the automatic annotation of photographs. This work investigates the application of the conformal prediction (CP) framework for extending the output of AFR techniques with well-calibrated measures of confidence. In particular we combine CP with one classifier based on patterns of oriented edge magnitudes descriptors, one classifier based on scale invariant feature transform descriptors, and a weighted combination of the similarities computed by the two. We examine and compare the performance of five nonconformity measures for the particular task in terms of their accuracy and informational efficiency.

Keywords Face recognition · Scale invariant feature transform · POEM · Conformal prediction · Confidence · Credibility · Prediction regions · Uncontrolled environment

Editors: Alexander Gammerman, Vladimir Vovk, Henrik Bostrom, and Lars Carlsson.

✉ Harris Papadopoulos
h.papadopoulos@frederick.ac.cy

Ladislav Lenc
llenc@kiv.zcu.cz

Pavel Král
pkral@kiv.zcu.cz

[1] Department of Computer Science and Engineering, Frederick University, 7 Y. Frederickou St., Palouriotisa, 1036 Nicosia, Cyprus

[2] Department of Computer Science and Engineering, Faculty of Applied Sciences, University of West Bohemia, Plzeň, Czech Republic

[3] NTIS - New Technologies for the Information Society, Faculty of Applied Sciences, University of West Bohemia, Plzeň, Czech Republic

1 Introduction

Automatic Face Recognition (AFR) refers to the use of a computer for the identification of a person from a digital photograph given a collection of digital photographs belonging to a number of different people, called a gallery. Nowadays AFR can be seen as one of the most progressive biometric authentication methods and represents a key task in several commercial or law enforcement applications such as surveillance of wanted persons, access control to restricted areas and automatic annotation of photos in photo sharing applications or social networks. Given the importance of such applications, the particular task has been the subject of many studies and many techniques have been proposed in the literature for it. For well-controlled environments (sufficiently aligned faces, similar face pose and lighting conditions, etc.) there are a number of approaches with a high recognition accuracy. However, in moderately controlled or fully uncontrolled environments the performance of most techniques is significantly lower (Kral and Lenc 2015).

Considering the difficulty of the task in moderately controlled or fully uncontrolled environments together with the rather large number of candidate outputs (all people in the gallery), some way of quantifying the uncertainty involved in each recognition would be very beneficial to many AFR applications. Therefore this work examines the utilization of a Machine Learning framework, called Conformal Prediction (Vovk et al. 2005), for quantifying uncertainty in AFR. CP can be used for complementing the predictions of conventional Machine Learning techniques with well-calibrated measures of confidence without assuming anything stronger than that the data is exchangeable. In the particular case, CP can provide either a confidence measure that indicates the likelihood of each recognition being correct, or produce a prediction set that is guaranteed to satisfy a given confidence level, thus narrowing down the possible candidates for each photograph with a guarantee on the frequency at which the true candidate will not be considered. Consequently the produced prediction sets can significantly reduce the workload of a manual identification process.

This paper is an extension of Eliades and Papadopoulos (2017) in which CP was combined with SIFT based classifiers and its performance was examined on the AT&T and Unconstrained Facial Images (UFI) corpora. Here we additionally consider POEM descriptors, which have also been shown to perform well in uncontrolled environments (Vu et al. 2012). In particular we utilize CP for extending one POEM based classifier, one SIFT based classifier and one classifier based on the weighted combination of the similarities computed by the two. The performance of the three techniques is examined on the Labeled Faces in the Wild (LFW) and UFI corpora, which both contain images taken in an uncontrolled environment as opposed to the AT&T corpus used in Eliades and Papadopoulos (2017). Furthermore they are both much larger than the AT&T corpus.

The combination of CP with some conventional technique, called the *underlying algorithm* of the CP, is performed through what is called a *Nonconformity Measure* (NCM), which utilizes the conventional technique to assess how different an object is from the known objects in the training set (Shafer and Vovk 2008). Though validity is guaranteed regardless of the NCM used, this measure affects the informativeness of the CP outputs. We develop and examine the performance of a number of NCMs for the three AFR techniques mentioned above, and in fact any technique based on calculating similarities between images, in terms of their accuracy and informational efficiency. The obtained results show that the proposed approaches provide high accuracy and well-calibrated confidence measures that can be useful in practice.

The rest of this paper is structured as follows. In Sect. 2 we provide an overview of related work on AFR and of previous work on obtaining confidence information for the particular task. Next, Sect. 3 gives a brief description of the general CP framework. Section 4 outlines the usage and calculation of POEM and SIFT features and details the conventional AFR techniques used as basis for the CPs proposed in this work. Section 5 defines the developed NCMs and completes the description of the proposed CP approaches. Section 6 describes the two corpora used for evaluating the proposed approaches, presents the experimental setting and performance measures used in our evaluation, and reports and discusses the obtained experimental results. Finally, Sect. 7 gives our conclusions and plans for future work.

2 Related work

The methods utilized for AFR can be divided into holistic and feature-based ones. We focus in the following text mainly on the feature-based methods, because they achieve significantly better results particularly in recent challenging AFR settings where the images are of uneven quality and show variances in appearance. The common idea of feature-based methods is the representation of the face as a set of features. In the identification scenario, the face representations are compared against a gallery of known faces and the recognized person is determined by some distance measure using the k-Nearest Neighbors (k-NN) algorithm.

A number of image descriptors have been used for creating the face representation. We can mention the popular Local Binary Patterns (LBP) (Ahonen et al. 2004) and many of its variants such as Local Ternary Patterns (LTP) (Tan and Triggs 2007) etc. Another successful method is the Patterns of Oriented Edge Magnitudes (POEM) (Vu et al. 2012) which is based on image gradients. A more recent example of gradient-based descriptor was proposed in Huang and Yin (2017). The authors propose Binary Gradient Patterns (BGP). The descriptors encode the local structures into a set of binary strings. This descriptor surpasses the performance of a variety of other descriptors on several standard corpora in the identification scenario. It reaches comparable accuracy as Deep Neural Network (DNN) based methods on the face verification task. These approaches usually divide the processed image using a rectangular grid and compute features for each region (Ahonen et al. 2004). An alternative way of feature extraction was proposed in Lenc and Král (2016) where the feature points are found dynamically and may differ for each image. Scale Invariant Feature Transform (SIFT) (Lowe 2004) is another popular descriptor used in many image processing tasks. The method includes both feature point detection and feature extraction tasks. Nanni et al. (2017) use an ensemble technique to combine face descriptors. The method is based on two descriptors: POEM and Monogenic Binary Coding (MBC). Each classifier in the ensemble determines the similarities that are subsequently fused by averaging (score-level fusion). The method is evaluated on the face recognition technology (FERET) and LFW standard corpora. The obtained results are slightly better than those of the descriptors used separately. Multi-Directional Multi-Level Dual-Cross Patterns (MDML-DCPs) are proposed by Ding et al. (2016). It is based on the first derivative of the Gaussian operator. The features are computed from the whole image as well as from image regions. The authors report that the descriptor outperforms state-of-the-art descriptors on several widely used datasets. Another efficient face recognition approach is proposed by Ding and Tao (2017). This method is particularly efficient when recognizing face images with arbitrary pose variations. A dense grid of 3D facial landmarks are projected to each 2D face image. Then, an optimal warp is estimated based on homography to correct texture deformation caused by pose variations.

The reconstructed frontal-view patches are then used for face recognition with common face descriptors. Experimental results demonstrate that this approach performs well on both constrained and unconstrained scenarios.

We must also mention Artificial Neural Networks (ANN) that were used already in the work of Lawrence et al. (1997). Some of the methods use neural networks for feature extraction. One approach for descriptor construction is described by Wen et al. (2016). This work utilizes a Convolutional Neural Network (CNN) to learn discriminative features. A novel loss function is proposed. It achieves state-of-the-art accuracy on LFW and MegaFace Challenge datasets. Another learning approach for image descriptors is proposed by Lu et al. (2015). Compact Binary Face Descriptor (CBFD) is another method that projects pixel difference vectors into low-dimensional binary vectors. It is done in an unsupervised manner. Many other ANN approaches also emerged with the recent boom of "Deep Learning" (Parkhi et al. 2015).

Confidence measures (CMs) have not been used in the field of AFR very often. However, given the uncontrolled nature of the images used nowadays, it can be an invaluable tool for the evaluation of the recognition result. It is beneficial in a wide range of applications because the provision of information on "how good is the recognition result" is of high importance. Studies examining CMs in AFR include a pseudo 2-D Hidden Markov Model classifier with features created by the Discrete Cosine Transform (DCT) presented in Eickeler et al. (2000). The authors propose three CMs based on the posterior probabilities and two others based on ranking the results. They experimentally show that the posterior class probability gives better results for the recognition error detection task. An ensemble of simple CMs was proposed by Kral and Lenc (2015). The authors utilize four measures that are subsequently combined using an Artificial Neural Network. The measures are based on posterior class probability and predictor features. The techniques presented by Eickeler et al. (2000) and Kral and Lenc (2015) however do not provide any guarantees on their CMs.

CP has previously been applied to AFR by Li and Wechsler (2005) for rejecting unknown individuals and identifying difficult to recognize faces in the open set setting. The same authors also applied CP to the recognition by parts setting in Li and Wechsler (2009). Even though Li and Wechsler (2005, 2009) applied CP to AFR, they did not evaluate the informativeness of the p values and prediction sets produced by CP. Our work differs in the setting examined and the underlying techniques utilized, but most importantly we additionally evaluate the informativeness of the outputs provided by CP using the criteria defined by Vovk et al. (2016) and investigate the performance of alternative NCMs.

3 Conformal prediction

This section gives a brief description of the main principles of CP. For more details see Vovk et al. (2005).

Let $A = \{(x_i, y_i)|i = 1, \dots, N\}$ denote our training set, where x_i is an object given in the form of an input vector or matrix, $R = \{t_1, \dots, t_c\}$ is the set of possible labels and $y_i \in R$ is the label of the corresponding input vector or matrix. Let $B = \{X_k|k = 1, \dots, M\}$ denote our test set, where X_k is a test instance (vector or matrix). We define as $C_{k,l} = A \cup \{(X_k, t_l)\}$, where $t_l \in R$, the training set extended with the test example X_k together with candidate label t_l. These sets will lead us to assessing predictions with confidence measures and finding which candidate labels are possible for the test instance X_k given a desired confidence level.

A *nonconformity score* (NCS) is a numerical value assigned to each instance that indicates how unusual or strange a pair (x_s, y_s) is, based on the underlying algorithm, where $s \in$

$\{1, \ldots, N, new\}$ is the index of the sth element in $C_{k,l}$ and new corresponds to the test instance. In particular, the underlying algorithm is trained on the instances belonging to $C_{k,l}$, for each $l \in \{1, \ldots, c\}$ and $k \in \{1, \ldots, M\}$, and the *nonconformity measure* (NCM) utilizes the resulting model to assign a NCS $\alpha_s^{k,l}$ to each example in $C_{k,l}$.

For every test example k we have c sequences of NCS denoted as $H_{k,l}$. Every sequence is used to find the p value of a test example k with a candidate label t_l. Given a sequence $H_{k,l}$ of NCS $\alpha_s^{k,l}$ we can calculate how likely a test instance (X_k, t_l) is with the function:

$$p_k(t_l) = \frac{\left|\left\{\alpha_s^{k,l} \in H_{k,l} | \alpha_s^{k,l} \geq \alpha_{new}^{k,l}\right\}\right|}{N+1}, \tag{1}$$

where $\alpha_{new}^{k,l}$ is the NCS of the k^{th} example in the test set with candidate label t_l.

Given a pair (X_k, t_l) with a p value of δ this means that this example will be generated with at most δ frequency, under the assumption that the examples are exchangeable, proven by Vovk et al. (2005).

After all p values have been calculated they can be used for producing prediction sets that satisfy a preset confidence level $1 - \delta$ (δ is called the significance level). Given the significance level δ, a CP will output the prediction set:

$$\{t_l | p_k(t_l) > \delta\}.$$

We would like prediction sets to be as small as possible. The size of the prediction sets depends on the quality of the p values and consequently on the NCM used.

If we want only a single prediction, or *forced prediction*, the CP outputs the label t_r with

$$r = \arg\max{}_{l=1,\ldots,c}\, p_k(t_l),$$

in other words the t_l with the highest p value. This prediction is complemented with measures of *confidence* and *credibility*. Confidence is defined as one minus the second largest p value. Confidence is a measure that indicates the likelihood of a predicted classification compared to all the other possible classifications. Credibility is defined as the largest p value. Low credibility means that either the data violate the exchangeability assumption or the particular test example is very different from the training set examples.

4 Automatic face recognition

In this work two efficient AFR techniques are used and combined with CP. We focus on the face identification task where a gallery of known people is available and the task is to find the identity of an unknown test face. We utilize a feature based face recognition method with two techniques of feature extraction.

Each person is represented as a set of feature vectors constructed in specified image locations. Therefore, the first step of our algorithm consists of the automatic detection of key-points (the most representative points) in the face images. Then, we calculate the face representation (feature vectors) in such points using two popular techniques, namely POEM and SIFT. If the gallery contains more images of one person, we use the so called "composed model" where features extracted from all images belonging to the given person are put together and create a single representation. The last step is the recognition itself where the face representations are compared with the gallery images in order to identify the person. This procedure is depicted in Fig. 1.

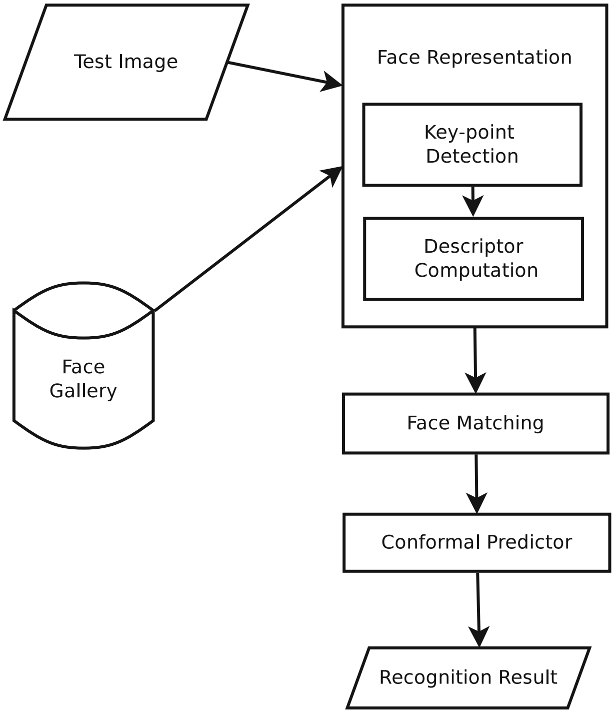


Fig. 1 Scheme of the proposed system combining face recognition and conformal prediction

4.1 Notation

In describing the techniques we use the following notation:

- We define as G_i the set of feature vectors (composed model) of person i in the training set.
- We define as T_k the set of feature vectors for image k in the test set.
- $q_m \in T_k$ is a feature vector belonging to the set T_k and $Q_n \in G_i$ is a feature vector belonging to the set G_i.
- $< x, y >$ denotes the dot product of vectors x and y in the Euclidean space.
- $|y|$ is the Euclidean norm if y is a vector of real values.
- $|C|$ is the cardinality of set C.

The methods of feature extraction are described in Sects. 4.2 and 4.3. The matching method used to compare the face representations is described in detail in Sect. 4.4, while the combined classifier is described in Sect. 4.5. The similarity scores provided by each technique are then used to define the AFR-CP nonconformity measures.

4.2 POEM features

4.2.1 Key-point detection

We use the key-point extraction method utilized in Lenc and Král (2014). A set of N_G Gabor filters of different orientations and wavelengths is applied to the original image and then the key-points are determined from the filter responses. The filtered images are scanned using a square sliding window W of size $w \times w$. The window center (r_0, c_0) is considered to be a key-point iff:

$$R_j(r_0, c_0) = \max_{(r,c)\in W} R_j(r, c), \tag{2}$$

$$R_j(r_0, c_0) > \frac{1}{wi * hi} \sum_{r=1}^{wi} \sum_{c=1}^{hi} R_j(r, c), \tag{3}$$

where R_j is the response of filter j (result of filtering the original image using filter j) $j = 1, \ldots, N_G$ (N_G is the number of Gabor filters) and wi and hi are the image width and height respectively. The key-point thus must have a value larger than all points in the defined neighborhood and at the same time higher than the average value of all pixels in the response j.

4.2.2 K-means clustering

The number of points determined in the previous section is usually too high (hundreds) and the points are often concentrated near important facial parts. Moreover, a high number of the points increases significantly the computation complexity. Therefore, we propose to use clustering to identify only the most important points. This idea is supported by the fact that similar methods use less than 100 points and achieve very good results. We chose the k-means algorithm to cluster the key-points.

4.2.3 Descriptor calculation

The POEM descriptor was proposed by Vu et al. (2012). First the gradient in each pixel of the input image is computed. An approximation utilizing a simple convolution operator such as Sobel or Scharr is used to compute gradients in the x and y directions. These values are used for the computation of gradient orientation and magnitude.

The gradient orientations are then discretized. The number of orientations is denoted as d and is usually set to 3. Each pixel is now represented as a vector of length d. It is a histogram of gradient values in a small square neighborhood of a given pixel called *cell*. The recommended value for the cell size is 7 pixels (Vu et al. 2012). Figure 2 depicts the meaning of *cell* and *block* terms.

The final encoding is similar to the local binary patterns algorithm (LBP). It is done in a round neighborhood with diameter L called *block*. The algorithm assigns either a 0 or 1 value to the 8 neighboring pixels as:

$$B_i = \begin{cases} 0 \text{ if } g_i < g_c, \\ 1 \text{ if } g_i \geq g_c, \end{cases} \tag{4}$$

where B_i is the binary value assigned to the neighboring pixel $i \in \{1, .., 8\}$, g_i denotes the gray-level value of the neighboring pixel i and g_c is the gray-level value of the central pixel. The resulting values are then concatenated into an 8 bit number. Its decimal representation is used to create the feature vector. This is computed for each gradient orientation and thus the descriptor is d times longer than in the case of LBP.

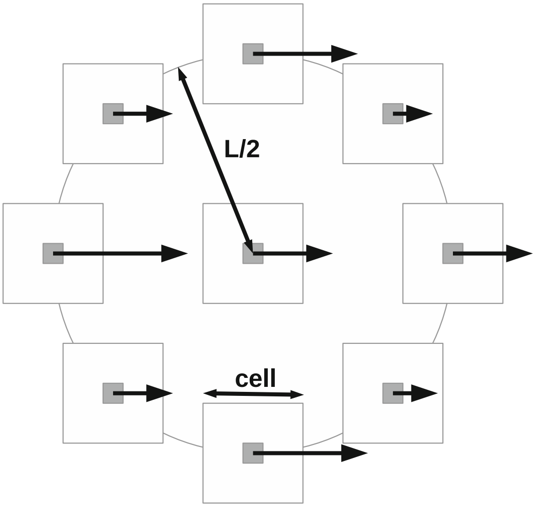


Fig. 2 POEM computation (Vu et al. 2012). The square regions around pixels represent the *cells* and the larger surroundings with diameter L are called *blocks*. Arrows represent the accumulated gradients for one direction

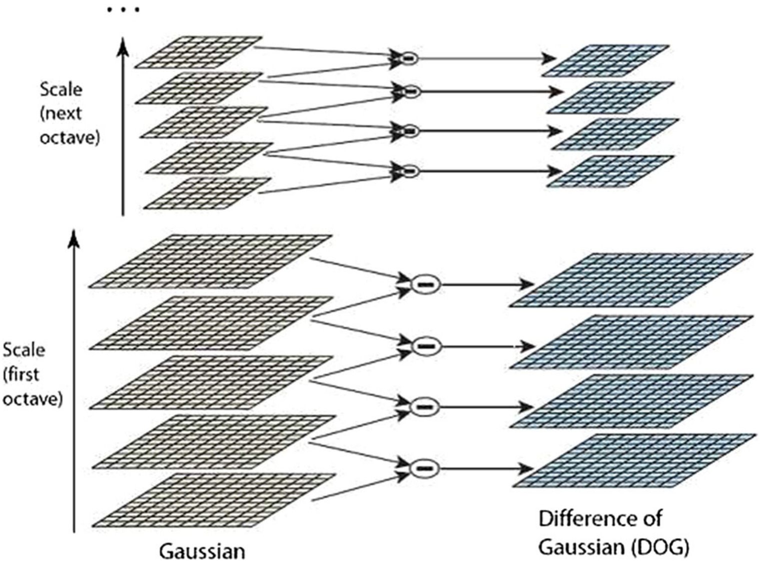


Fig. 3 Difference of Gaussian filters at different scales (Lowe 2004)

4.3 SIFT features

The SIFT algorithm (Lowe 2004) consists of four steps: extrema detection, removal of keypoints with low contrast, orientation assignment and descriptor calculation.

4.3.1 Extrema detection

The extrema detection process utilizes a Gaussian pyramid. The adjacent Gaussians are subtracted to produce the difference of Gaussians (DoG). This process is illustrated in Fig. 3.

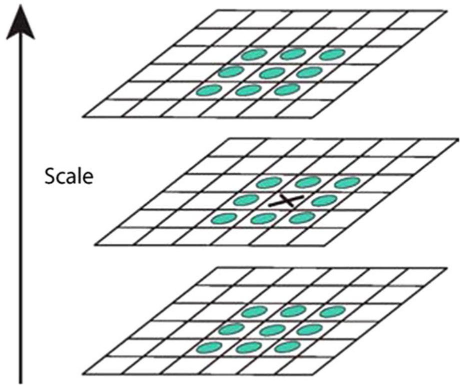


Fig. 4 Maxima and minima of the difference-of-Gaussian images are detected by comparing a pixel (marked with X) to its 26 Neighbors in 3 × 3 regions at the current and adjacent scales (marked with circles) (Lowe 2004)

Each pixel is then compared with 8 neighbors in the current scale and 9 neighbors at the scales above and below as depicted in Fig. 4. A pixel is selected as a key-point if it is larger or smaller than all of the 26 neighbors.

4.3.2 Low contrast key-point removal

If the magnitude of the intensity (i.e., without sign) at the current pixel in the DoG is less than a certain value, it is considered as a low contrast key-point. Key-points poorly localized along an edge are detected if some inequalities related to the Hessian Matrix and its corresponding eigenvalues are found, the derivatives are estimated by taking the differences of neighboring sample points.

4.3.3 Orientation assignment

A consistent orientation is assigned to each key point based on the local image properties ensuring invariance to image rotation. The calculation is based upon local gradient orientations in the neighborhood of the pixel, first a smoothed histogram of local gradient directions is created, then the peaks in the histogram are assigned.

4.3.4 Descriptor calculation

The final step consists in the creation of descriptors for the local image regions. The descriptors are highly distinctive and invariant as much as possible to changes in illumination and 3d camera viewpoint. The computation involves the 16 × 16 neighborhood of the key-point location. Gradient magnitudes and orientations are computed in each point of the neighborhood. Their values are weighted by a Gaussian window. For each sub-region of size 44 (16 regions), orientation histograms are created. Finally, a vector containing 128 (16 × 8) values is generated.

4.4 Lenc–Kral matching

This algorithm, called Lenc-Kral Matching (LKM), has been proposed in Lenc and Kral (2012). In this work we use it with two different similarity measures. SIFT features are mostly compared using cosine similarity (Lenc and Kral 2015a) defined by

$$cos(q_m, Q_n) = \frac{< q_m, Q_n >}{|q_m|.|Q_n|}. \tag{5}$$

POEM features are usually compared using χ^2 statistic or Histogram intersection (HI). Based on our preliminary experiments we chose HI which works slightly better. It is defined as follows:

$$HI(q_m, Q_n) = \sum_{i=1}^{n} \min(q_m(i), Q_n(i)). \tag{6}$$

For each feature vector q_m of the recognized face T_k we determine the most similar vector $Q^j_{max_n}$ of the gallery person G_j as:

$$Q^j_{max_n} = \arg\max_{G_j}(sim(q_m, Q_n)), \tag{7}$$

where sim denotes one of the similarity measures.

The sum of those similarities is computed as follows:

$$D^k_j = \sum_{n=1..|T_k|} Q^j_{max_n}, \tag{8}$$

where $|T_k|$ is the number of test image feature vectors. The recognized face is then determined by the following equation:

$$\hat{G}_j = \arg\max_{G_j}(D^k_j). \tag{9}$$

4.5 Combined classifier

We also examine a weighted combination of the similarities computed by the SIFT and POEM classifiers discussed in Sect. 4.4. Specifically, the similarities D^k_j computed by each classifier were mapped to the interval (0, 1) and their weighted average forms the combined similarity:

$$D^k_j = wD(SIFT)^k_j + (1 - w)D(POEM)^k_j, \tag{10}$$

where $D(SIFT)^k_j$ is the normalized output of (8) with (5) as similarity and $D(POEM)^k_j$ is the normalized output of (8) with (6) as similarity. The weight w was determined by examining the accuracy on the training set using leave-one-out cross-validation and selecting the one with best performance. After computing the combined similarities a person is recognized by applying Eq. (9).

5 Nonconformity measures for AFR-CP

In this section we complete the description of the proposed Automatic Face Recognition Conformal Predictor (AFR-CP) by defining the NCM's used, these measures are based on the three classifiers described in Sect. 4. We have examined several NCMs to investigate which of them provides the most informative p values. Recall from Sect. 3 that $C_{k,l} = A \cup \{(X_k, t_l)\}$, where $\{t_1, \ldots, t_c\}$ are the possible labels, corresponding to all persons in our gallery in this case. For each test example X_k CP generates $C_{k,1}, \ldots, C_{k,c}$ and assigns a NCS to each example in each of the c sets. We denote as $z^{k,l}_s$ the sth element of $C_{k,l}$ and as $\alpha^{k,l}_s$ its

NCS, with $s = 1, \ldots, N, new$ (where $z^{k,l}_{new}$ and $\alpha^{k,l}_{new}$ correspond to (X_k, t_l) and its NCS respectively).

When implementing (CP) for defining the NCM $\alpha^{k,l}_s$ we use the D^s_j calculated by the underlying AFR technique [see Eqs. (8) and (10)] with $\{z^{k,l}_i : i = 1, \ldots, s-1, s+1, \ldots, N, new\}$ as gallery, where $z^{k,l}_{new} = (X_k, t_l)$. In other words D^s_j is the similarity of $z^{k,l}_s$ with the person t_j. Our NCMs are defined in such way to contain at least one of two quantities: The first quantity, D^s_j where $y_s = t_j$, summarizes the similarity of the instance s with the same person, while the second quantity summarizes the similarity of the instance to all other persons. The bigger the NCS the more non-conforming the example and the lower the NCS the less non-conforming the example.

The following nonconformity measures have been used:

The 1st NCM is defined as

$$\alpha^{k,l}_s = \frac{1}{D^s_j}, \tag{11}$$

where only the similarity of s with person t_j is taken into account.

The 2nd NCM is defined as

$$\alpha^{k,l}_s = \text{mean}_{i \neq j}(D^s_i) - D^s_j, \tag{12}$$

where the first quantity represents the mean similarity of image s with all other persons excluding t_j.

The 3rd NCM is defined as

$$\alpha^{k,l}_s = \frac{\text{mean}_{i \neq j}(D^s_i)}{D^s_j}, \tag{13}$$

where the same quantities as in (12) are used, but now subtraction is replaced by division.

The 4th NCM for an image s corresponding to person t_j is defined as

$$\alpha^{k,l}_s = \max_{i \neq j}(D^s_i) - D^s_j, \tag{14}$$

where the first quantity represents the similarity of image s with the most similar of all other persons excluding t_j.

The 5th NCM is defined as

$$\alpha^{k,l}_s = \frac{\max_{i \neq j}(D^s_i)}{D^s_j}, \tag{15}$$

where the same quantities as in (14) are used, but now subtraction is replaced by division. It should be noted that when $s = new$ we use $y_s = t_l$ (the assumed class). After calculating the NCS we calculate p values and make predictions following the process described in Sect. 3.

6 Experiments and results

6.1 Corpora

6.1.1 Labeled faces in the wild

We use the cropped version of the well-known LFW corpus, so called *LFWcrop*.[1] This version was first utilized by Sanderson and Lovell (2009). The cropping is realized to ensure more

[1] http://conradsanderson.id.au/lfwcrop/.

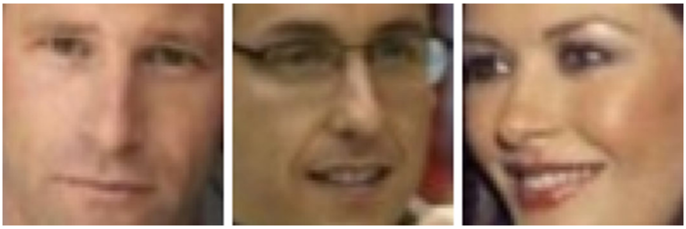


Fig. 5 Three example images from the LFWcrop dataset

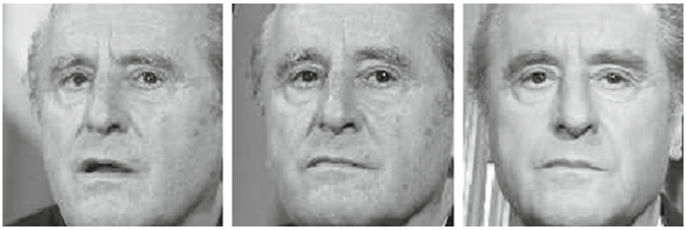


Fig. 6 Examples of one person from the UFI face corpus

standardized conditions for testing of face recognition approaches. The main reason for such preprocessing is the presence of a background in the original images that may add information and improve performance in some cases.

The extraction method places a bounding box around the faces and resizes the resulting area to 64 × 64 pixels. The bounding box is placed at the same location in every image.

We use the identification scenario proposed by Xu et al. (2014). It uses a subset of 86 people with 11–20 images per person. Seven images of each person are used for the gallery and the rest are used as the test set. The total number of images are 602 and 649 for the gallery and test set respectively. Figure 5 shows three example images from the LFW face dataset.

6.1.2 Unconstrained facial images

The UFI dataset was proposed by Lenc and Král (2015b). It is a real-world database created from photographs acquired by reporters of a news agency. It thus shows significant variances in the image quality, face orientation, face occlusion etc. The database is designated for the identification task. It comes with two image sets. The *Cropped images* dataset contains preprocessed faces extracted from photographs while the *Large images* includes a variable amount of background. We utilize the cropped version as the other partition is intended to be used with complete face recognition systems including the face localization stage.

The images have resolution of 128 × 128 pixels. The total number of individuals is 605. On average 7.1 images of each person are in the gallery set. The total number of gallery images is 4316. The test set contains just one image for every individual. Figure 6 shows three example images from one individual.

6.2 Experimental setting and performance measures

In this section we detail the experiments and results of the proposed AFR-CPs and of the three techniques used as their underlying models on the LFW and UFI face datasets described previously. Our experiments on both datasets were performed using the provided training and test sets.

In the case of the combined classifier the weights of the POEM and SIFT similarities were estimated using the leave-one-out cross-validation method on the training set on the conventional classifier. Leave-one-out was chosen due to the fact that the number of images per subject in the galleries varied.

Due to the fact that the accuracy itself is not a good indication for the choice of a NCM we used four probabilistic criteria for evaluating p values proposed in Vovk et al. (2016). These criteria are divided into two main categories called *Basic Criteria*, which do not take into account the true label, and *Observed Criteria*, which take into account the true label. The two Basic Criteria we used are:

- The S-criterion

$$\frac{1}{M}\sum_{l=1}^{c}\sum_{k=1}^{M} p_k(t_l), \tag{16}$$

  where $p_k(t_l)$ is the p value of the test example X_k with candidate label t_l. In effect the S-criterion is the average sum of all p values.
- The N-criterion

$$\frac{1}{M}\sum_{k=1}^{M} |\{t_l | p_k(t_l) > \delta\}|, \tag{17}$$

  which is the average size of the prediction sets with respect to a confidence level $1 - \delta$.

The two Observed Criteria we used are:

- The OF-criterion

$$\frac{1}{M}\sum_{k=1}^{M}\sum_{l, t_l \neq t_k} p_k(t_l), \tag{18}$$

  which corresponds to the average sum of the p values of the false labels.
- The OE-criterion

$$\frac{1}{M}\sum_{k=1}^{M} |\{t_l | p_k(t_l) > \delta, t_l \neq t_k\}|, \tag{19}$$

  which represents the average number of false labels included in the prediction sets, with respect to a confidence level $1 - \delta$.

For all criteria smaller values indicate more informative p values. Note that their output values are bounded below by zero.

6.3 LFWcrop corpus results

6.3.1 Accuracy

Table 1 presents the accuracy of the conventional AFR techniques on the LFW corpus in comparison with that of state-of-the-art techniques on the same dataset. Table 2 reports the accuracy of the corresponding AFR-CP techniques (using the conventional AFR techniques

Table 1 Accuracy of the three conventional AFR techniques on the LFW dataset in comparison with state-of-the-art techniques

Classifier	Accuracy (%)
HMLBP (Guo et al. 2010)	36.28
NCC (Marsico et al. 2013)	48.37
M-BNCC (Gaston et al. 2017)	65.27
POEM	55.16
SIFT	35.75
Combined	56.24

Table 2 Average accuracy, credibility and confidence of the AFR-CPs on the LFW dataset computed using non-conformity scores defined in Eqs. (11)–(15).

	Underlying technique	Nonconformity measure				
		(11)	(12)	(13)	(14)	(15)
Accuracy	POEM	55.47	55.32	55.32	55.32	55.32
	SIFT	34.21	35.90	35.90	35.75	35.90
	Combined	55.62	56.24	56.24	56.24	56.24
Average confidence	POEM	54.45	52.92	53.05	68.31	68.69
	SIFT	49.94	35.14	35.31	48.69	49.23
	Combined	53.32	53.65	53.45	70.76	71.01
Average credibility	POEM	60.06	64.17	63.99	67.53	67.52
	SIFT	51.13	73.42	72.60	77.37	77.47
	Combined	52.99	63.84	63.50	66.96	66.93

as underlying algorithms) along with the average confidence and credibility measures they produced using the five NCMs defined in Sect. 5. The results reported in these tables show that the use of the POEM descriptors significantly improves performance over the previously used SIFT descriptors. Additionally, a further improvement in accuracy is achieved by combining the similarities produced by the POEM-based and SIFT-based techniques. In fact, the accuracy of the POEM-based and combined classifiers is higher than that of two out of the three state-of-the-art techniques while being rather close to that of the best performing one.

In comparing the accuracy of the AFR-CP techniques (Table 2, first three rows) with that of their conventional counterparts (Table 1, last three rows), we observe that, with the exception of NCM (11), the former is equal to or in some cases even slightly better than the latter. In the case of NCM (11) the accuracy of the SIFT-based and combined classifiers is slightly lower than that of the conventional techniques. The reason for this is that the particular NCM only takes into account the similarity of the image to those of the correct/assumed person, as opposed to all other measures that include additional information about the similarity of the image to those of all other candidate persons.

Finally it is worth noting that the rather low accuracy even of the best performing techniques indicates the difficulty of the problem and consequently the need for quantifying the high uncertainty involved in uncontrolled environment face recognition, especially taking into consideration the large number of possible labels involved.

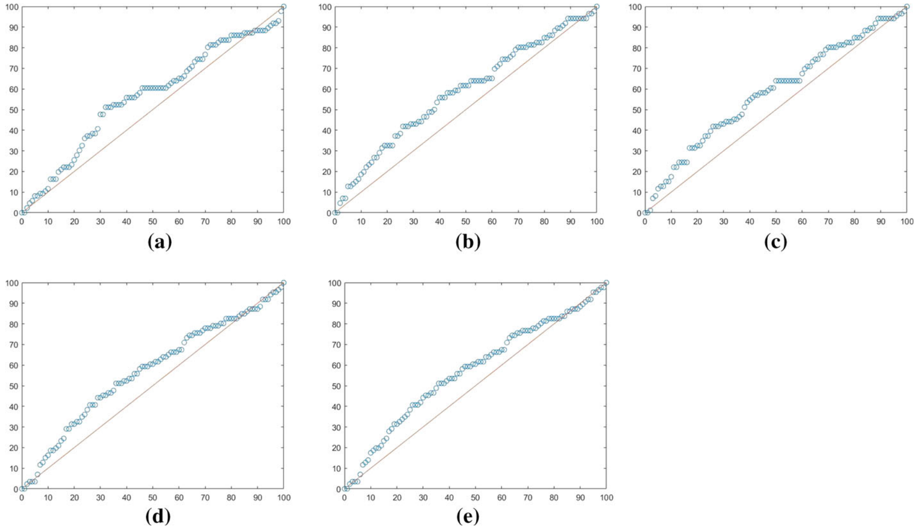


Fig. 7 Percentage of correct region predictions with respect to the confidence level of the CP combined with the five NCMs with the POEM-based underlying technique on the LFW dataset. The y-axis is the percentage of correct region predictions while the x-axis is the confidence level. **a** Measure (11). **b** Measure (12). **c** Measure (13). **d** Measure (14). **e** Measure (15)

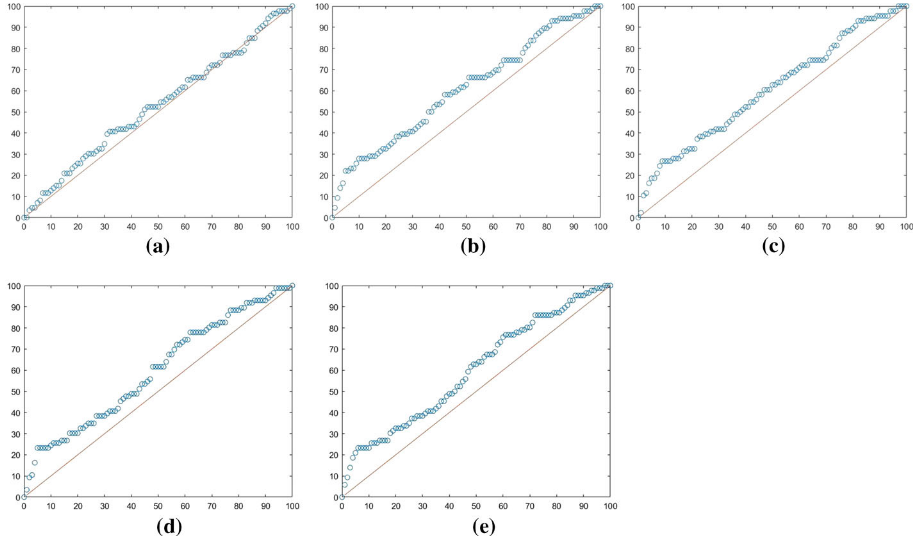


Fig. 8 Percentage of correct region predictions with respect to the confidence level of the CP combined with the five NCMs with the SIFT-based underlying technique on the LFW dataset. The y-axis is the percentage of correct region predictions while the x-axis is the confidence level. **a** Measure (11). **b** Measure (12). **c** Measure (13). **d** Measure (14). **e** Measure (15)

6.3.2 Empirical validity

In this subsection we examine the empirical validity of the prediction regions produced by the proposed techniques for the LFW corpus. Figures 7, 8 and 9 present the percentage of correct

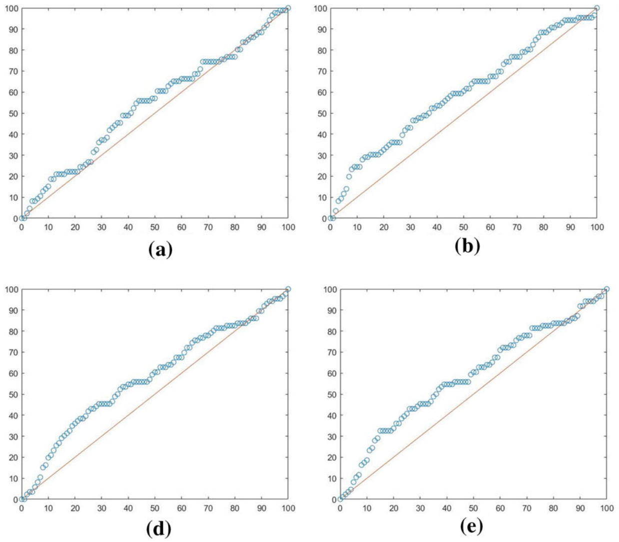


Fig. 9 Percentage of correct region predictions with respect to the confidence level of the CP combined with the five NCMs with the combined underlying technique on the LFW dataset. The y-axis is the percentage of correct region predictions while the x-axis is the confidence level. **a** Measure (11). **b** Measure (12). **c** Measure (13). **d** Measure (14). **e** Measure (15)

region predictions as a function of the confidence level for the five NCMs used on top of the POEM-based, SIFT-based and combined techniques respectively. In all cases the plots are very close (and slightly above) the diagonal indicating that the produced region predictions are well-calibrated (the accuracy is equal to or slightly higher than the required confidence level), as guaranteed by CP. It is worth noting that a small deviation from the diagonal is expected since our experiments were performed on the particular test set provided with the dataset. As test instances increase the plots will follow the diagonal even more closely.

6.3.3 Informational efficiency

Since the purpose of this work is to provide additional information for each test example, here we examine the quality of the *p* values produced by the proposed approaches and consequently how informative the resulting prediction regions are. This is done following the informational efficiency criteria described in Sect. 6.2 and proposed by Vovk et al. (2016).

Table 3 presents the values of the two unobserved criteria on the LFW corpus for the AFR-CPs with the five NCMs. Specifically the second column of the table contains the values of the S criterion, while the rest of the columns present the N criterion for the significance levels 0.01, 0.05, 0.1, 0.15, 0.2 and 0.25. In the same manner Table 4 presents the values of the two observed criteria. The second column contains the values of the OF criterion, while the rest of the columns give the values of the OE criterion for the significance levels 0.01, 0.05, 0.10, 0.15, 0.2 and 0.25.

The results reported in the two tables show that in terms of informational efficiency there is no significant difference between the AFR-CPs with the POEM-based and combined classifiers as underlying techniques while they clearly outperform the SIFT-based AFR-CP. Furthermore a comparison between the results obtained with the five different NCMs shows that measures (14) and (15) produce the most informative *p* values according to all

Table 3 Unobserved criteria for the AFR-CPs on the LFW dataset using the NCMs defined in Eqs. (11)–(15)

Underlying technique	NCM	S criterion	N criterion (per significance level)					
			0.01	0.05	0.10	0.15	0.20	0.25
POEM	(11)	13.48	83.79	60.60	45.99	33.73	27.28	19.31
	(12)	8.48	67.72	44.98	27.39	18.67	13.42	9.19
	(13)	8.49	67.95	45.00	27.25	18.47	13.39	9.21
	(14)	6.09	48.07	31.53	17.99	12.99	9.34	7.39
	(15)	6.13	49.17	31.94	18.32	13.12	9.44	7.52
SIFT	(11)	40.28	85.12	81.26	75.81	70.69	65.95	63.05
	(12)	18.14	83.33	71.46	56.88	45.63	35.71	28.42
	(13)	18.40	83.77	70.77	56.81	45.87	36.14	29.19
	(14)	15.86	74.06	57.05	46.50	39.81	31.74	27.18
	(15)	16.64	77.52	61.57	48.66	40.30	34.16	28.76
Combined	(11)	29.25	84.21	77.47	65.23	58.39	51.55	46.02
	(12)	8.16	64.43	43.52	27.80	18.36	12.14	8.41
	(13)	8.23	63.61	43.55	28.38	18.95	12.27	8.63
	(14)	5.70	48.48	29.53	18.66	12.02	8.75	6.51
	(15)	5.84	49.68	30.01	19.40	12.65	9.43	6.55

Table 4 Observed criteria for the AFR-CPs on the LFW dataset using the NCMs defined in Eqs. (11)–(15)

Underlying technique	NCM	OF criterion	OE criterion (per significance level)					
			0.01	0.05	0.10	0.15	0.20	0.25
POEM	(11)	12.96	82.80	59.66	45.09	32.86	26.45	18.53
	(12)	7.94	66.73	44.03	26.47	17.80	12.56	8.39
	(13)	7.96	66.96	44.04	26.34	17.59	12.54	8.41
	(14)	5.55	47.09	30.58	17.07	12.10	8.48	6.57
	(15)	5.60	48.19	30.98	17.41	12.23	8.58	6.69
SIFT	(11)	39.78	84.13	80.31	74.92	69.84	65.17	62.28
	(12)	17.58	82.33	70.48	55.94	44.72	34.83	27.59
	(13)	17.84	82.77	69.80	55.86	44.96	35.27	28.35
	(14)	15.30	73.06	56.07	45.56	38.89	30.86	26.32
	(15)	16.09	76.53	60.59	47.72	39.38	33.28	27.90
Combined	(11)	28.74	83.22	76.5	64.34	57.53	50.75	45.26
	(12)	7.61	63.51	43.43	28.24	18.88	12.21	8.58
	(13)	7.68	62.63	42.60	27.45	18.07	11.42	7.84
	(14)	5.16	47.49	28.57	17.73	11.13	7.88	5.67
	(15)	5.29	48.69	29.05	18.46	11.75	8.55	5.71

four criteria. The values of the N and OE criteria for these NCMs demonstrate the practical usefulness of the produced prediction regions since when using the POEM-based or combined AFR-CPs together with these measures, the resulting prediction regions contain on average less than 20 out of the possible 86 persons in the gallery and less than 19 wrong labels out of the possible 85 for a confidence level as high as 90% (Tables 3 and 4 respectively, significance level 0.1). By lowering the confidence level to 75%, which is still well above the accuracy of the best state-of-the-art technique, the aforementioned prediction sets contain less than 8 labels on average and less than 7 wrong labels on average (Tables 3 and 4 respectively, significance level 0.25). This is arguably a good result considering the high difficulty of the task and the moderate accuracy of conventional state-of-the art AFR techniques.

6.4 UFI corpus results

6.4.1 Accuracy

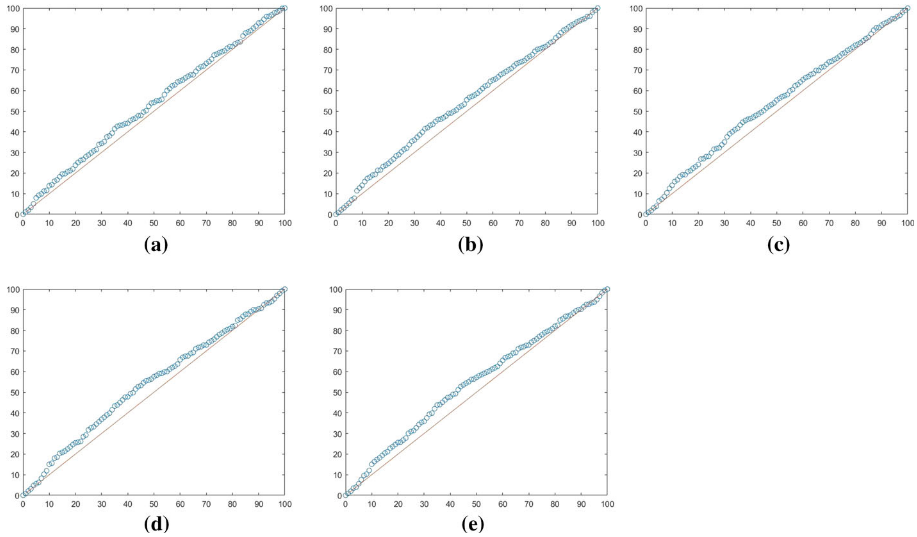
Table 5 presents the accuracy of the three conventional AFR techniques on the test set of the UFI corpus subset in comparison with that of state-of-the-art techniques on the same set. Table 6 reports the accuracy of the corresponding AFR-CP techniques along with the average confidence and credibility measures they produced using the five NCMs defined in Sect. 5. The results shown in these tables are consistent with the ones reported in Table 1 in that the POEM-based classifier performs much better than the SIFT based one while the combination of the two gives some further improvement. The accuracy of the POEM-based and combined classifiers is higher than that of two out of the three state-of-the-art techniques while having no significant difference with that of the best performing technique.

Table 5 Accuracy of the three conventional AFR techniques on the UFI dataset in comparison with state-of-the-art techniques

Classifier	Accuracy (%)
FS-LBP (Lenc and Král 2016)	63.31
POEM (Lenc and Král 2015c)	67.11
M-BNCC (Gaston et al. 2017)	74.55
POEM	71.07
SIFT	58.68
Combined	73.39

Table 6 Average accuracy, credibility and confidence of the AFR-CPs on the UFI dataset computed using non-conformity scores defined in Eqs. (11)–(15)

	Underlying technique	Nonconformity measure				
		(11)	(12)	(13)	(14)	(15)
Accuracy	POEM	70.91	71.07	71.07	71.07	71.07
	SIFT	57.19	58.84	58.51	58.68	58.68
	Combined	73.39	73.22	73.39	73.39	73.39
Average confidence	POEM	66.03	68.20	68.04	85.52	85.91
	SIFT	50.77	53.03	52.39	71.97	72.13
	Combined	56.91	69.50	68.44	86.86	87.05
Average credibility	POEM	57.36	58.60	58.56	60.81	60.74
	SIFT	51.44	62.33	61.68	64.66	64.80
	Combined	53.93	58.02	58.11	59.74	59.78



(a) **(b)** **(c)** **(d)** **(e)**

Fig. 10 Percentage of correct region predictions with respect to the confidence level of the CP combined with the five NCMs with the POEM underlying technique on the UFI corpus subset. The y-axis is the percentage of correct region predictions while the x-axis is the confidence level. **a** Measure 1. **b** Measure 2. **c** Measure 3. **d** Measure 4. **e** Measure 5

In comparing the accuracy of the AFR-CP techniques (Table 6, first three rows) with that of their conventional counterparts (Table 5, last three rows), we observe that in almost all cases they are equal. The only two exceptions to this are the POEM-based technique with NCM (11) and the combined technique with NCM (12) and in both cases the difference is very small to be of significance.

6.4.2 Empirical validity

Figures 10, 11 and 12 examine the empirical validity of the prediction regions produced by the proposed techniques for the UFI corpus. Specifically, they plot the percentage of correct region predictions as a function of the confidence level for the five NCMs used on top of the POEM-based, SIFT-based and combined techniques respectively. Like in the case of the LFW corpus, in all cases the plots are very close to the diagonal indicating that the produced region predictions are well-calibrated, as guaranteed by CP. Again note that the small deviation from the diagonal is due to statistical fluctuations.

6.4.3 Informational efficiency

The most important evaluation and comparison of the different NCMs is in term of the informational efficiency of the corresponding AFR-CPs. Tables 7 and 8 report the performance of the AFR-CPs on the UFI corpus in terms of the two unobserved and the two observed efficiency criteria described in Sect. 6.2 respectively. The values reported in these tables suggest that, as in the case of the LFW dataset, the NCMs (14) and (15) perform best with all underlying techniques. Overall the p values produced by the AFR-CP for this corpus are more informative than the ones produced for the LFW corpus. This is due to the better performance of the three underlying techniques on this corpus. In particular the POEM-based

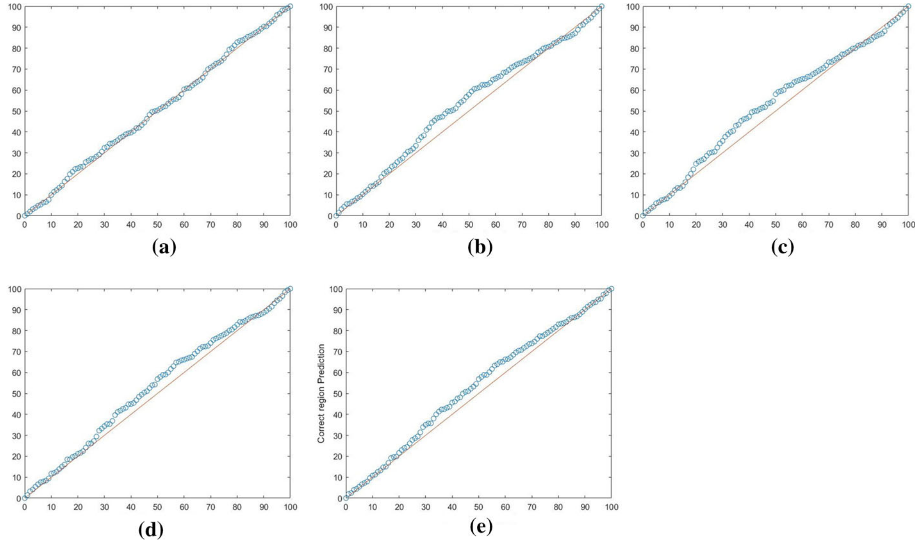


Fig. 11 Percentage of correct region predictions with respect to the confidence level of the CP combined with the five NCMs with the SIFT underlying technique on the UFI corpus subset. The y-axis is the percentage of correct region predictions while the x-axis is the confidence level. **a** Measure 1. **b** Measure 2. **c** Measure 3. **d** Measure 4. **e** Measure 5

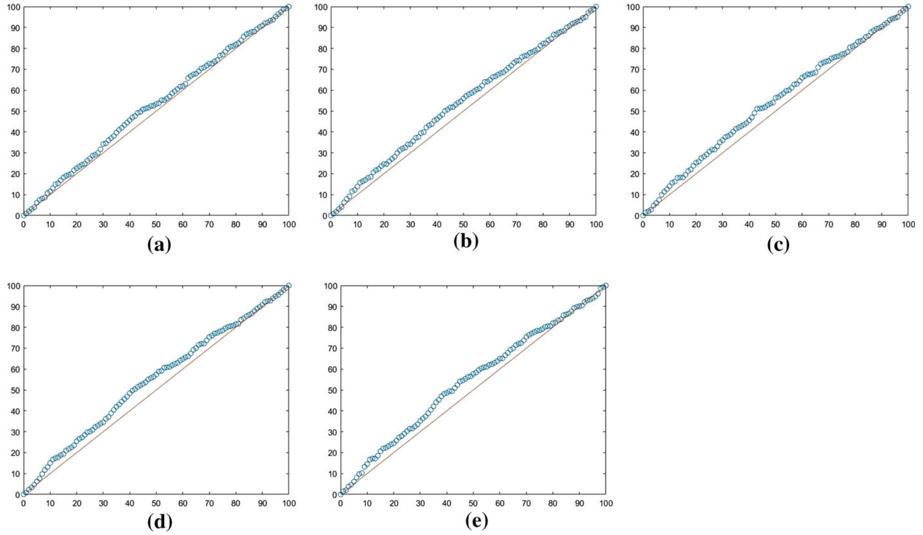


Fig. 12 Percentage of correct region predictions with respect to the confidence level of the CP combined with the five NCMs with the combined underlying technique on the UFI corpus subset.The y-axis is the percentage of correct region predictions while the x-axis is the confidence level. **a** Measure 1. **b** Measure 2. **c** Measure 3. **d** Measure 4. **e** Measure 5

and combined AFR-CPs with NCMs (14) and (15) for a confidence level of 90% (Table 7, significance level 0.1) produce prediction regions that contain on average less than 21 out of the possible 605 persons in the gallery (only 3.35% of all possible labels), while at the 80% confidence level (Table 7, significance level 0.2), which is still above the accuracy of

Table 7 Unobserved criteria on the UFI dataset using the NCMs defined in Eqs. (11)–(15)

Underlying technique	NCM	S criterion	N criterion (per significance level)					
			0.01	0.05	0.10	0.15	0.20	0.25
POEM	(11)	49.17	587.12	309.13	165.93	98.20	45.32	28.36
	(12)	24.58	459.02	138.88	53.61	23.43	9.66	4.44
	(13)	25.01	461.41	142.16	55.89	23.93	11.37	5.09
	(14)	11.65	245.87	56.38	20.80	10.10	3.79	1.90
	(15)	11.15	232.36	51.97	20.19	9.64	3.85	1.90
SIFT	(11)	275.37	596.68	568.44	532.60	500.82	464.83	421.81
	(12)	64.17	559.30	389.71	225.31	146.53	80.75	51.82
	(13)	70.02	566.97	391.45	230.34	146.95	100.68	70.30
	(14)	41.62	466.98	245.20	134.67	96.53	57.93	27.36
	(15)	45.39	487.33	259.44	166.01	112.48	67.12	33.89
Combined	(11)	181.17	582.58	497.83	438.56	369.35	324.37	282.34
	(12)	25.39	512.85	137.68	47.71	18.95	8.77	4.00
	(13)	26.33	503.07	140.48	49.33	21.70	9.95	5.55
	(14)	12.70	286.99	58.84	18.99	6.82	3.12	1.95
	(15)	12.77	282.00	68.02	18.88	7.79	3.25	2.00

Table 8 Observed criteria on the UFI dataset using the NCMs defined in Eqs. (11)–(15).

Underlying technique	NCM	OF criterion	OE criterion (per significance level)					
			0.01	0.05	0.10	0.15	0.20	0.25
POEM	(11)	48.64	586.13	308.17	165.01	97.32	44.50	27.57
	(12)	24.04	458.03	137.93	52.69	22.57	8.85	3.67
	(13)	24.47	460.42	141.21	54.98	23.07	10.55	4.31
	(14)	11.10	244.88	55.43	19.90	9.22	2.97	1.12
	(15)	10.60	231.37	51.04	19.29	8.78	3.03	1.12
SIFT	(11)	274.87	595.68	567.48	531.69	499.97	464.00	421.05
	(12)	63.64	558.31	388.78	224.44	145.68	79.94	51.06
	(13)	69.49	565.98	390.51	229.47	146.11	99.88	69.53
	(14)	41.09	465.99	244.26	133.78	95.67	57.10	26.58
	(15)	44.86	486.34	258.50	165.11	111.62	66.29	33.11
Combined	(11)	180.65	581.59	496.89	437.65	368.49	323.55	281.57
	(12)	24.85	511.86	136.74	46.80	18.09	7.96	3.23
	(13)	25.79	502.09	139.54	48.43	20.85	9.15	4.78
	(14)	12.17	285.99	57.91	18.08	5.96	2.31	1.16
	(15)	12.23	281.00	67.08	17.98	6.93	2.43	1.21

the best performing state-of-the-art technique, they contain on average less than 4 persons (only 0.66% of all possible labels). Given the very high number of possible persons in the gallery, one can appreciate the practical usefulness of these prediction regions.

7 Conclusions

We examined the application of conformal prediction in unconstrained environment AFR. Unlike most existing AFR approaches that output only a single prediction, the proposed CP approaches complement each of their predictions with probabilistically valid measures of confidence. The difficulty of the particular task as well as the large number of candidate labels (all persons in the gallery) indicate the usefulness of providing confidence information rather than just the most likely person in the gallery.

In particular we implemented CP on top of three AFR classifiers: one based on POEM descriptors, one based on SIFT descriptors and a weighted combination of the two. The combination of these classifiers with CP was performed through five different NCMs and the performance of the resulting AFR-CPs was investigated experimentally on two corpora consisting of images taken in an uncontrolled environment. Namely the performance of the proposed approaches was evaluated experimentally on the LFW and UFI corpora.

Our experimental results show that in terms of accuracy the proposed approaches are comparable with state-of-the-art conventional AFR techniques while having the added advantage of quantifying the uncertainty involved in each prediction. The empirical validity results demonstrate that the prediction regions produced by the AFR-CPs are always valid, i.e. having an accuracy equal to or higher than the desired confidence level. Based on the informational efficiency comparison of the produced p values the POEM-based and combined underlying techniques together with (14) and (15) as NCMs seem to perform best. Considering the difficulty of the task combined with the large set of possible persons, the resulting prediction sets can be very useful in the manual classification process by significantly reducing the number of candidate persons for each image.

The proposed CP method can be combined with any other underlying technique, while a better performing underlying technique will result in more informative CP outputs. Thus our future plans include examining alternatives to the conventional AFR techniques used here, such as Deep Neural Networks, and the investigation of the performance of CP on much larger datasets. Furthermore the examination of other AFR settings, such as the open set and recognition by parts, is also a future goal.

Acknowledgements This work has been partly supported by the project LO1506 of the Czech Ministry of Education, Youth and Sports.

References

Ahonen, T., Hadid, A., & Pietikinen, M. (2004). Face recognition with local binary patterns. In T. Pajdla & J. Matas (Eds.) *Computer Vision - ECCV 2004* (Vol. 3021, pp. 469–481). Lecture Notes in Computer Science. Berlin: Springer.

De Marsico, M., Nappi, M., Riccio, D., & Wechsler, H. (2013). Robust face recognition for uncontrolled pose and illumination changes. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *43*(1), 149–163.

Ding, C., Choi, J., Tao, D., & Davis, L. S. (2016). Multi-directional multi-level dual-cross patterns for robust face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(3), 518–531.

Ding, C., & Tao, D. (2017). Pose-invariant face recognition with homography-based normalization. *Pattern Recognition*, *66*, 144–152.

Eickeler, S., Jabs, M., & Rigoll, G. (2000). Comparison of confidence measures for face recognition. In *Proceedings of fourth IEEE international conference on automatic face and gesture recognition, 2000* (pp. 257–262). IEEE.

Eliades, C., & Papadopoulos, H. (2017). Conformal prediction for automatic face recognition. In: A. Gammerman, V. Vovk, Z. Luo, H. Papadopoulos (Eds.) *Proceedings of the sixth workshop on conformal and probabilistic prediction and applications, proceedings of machine learning research* (vol. 60, pp. 62–81). PMLR, Stockholm, Sweden. http://proceedings.mlr.press/v60/eliades17a.html.

Gaston, J., Ming, J., & Crookes, D. (2017). Unconstrained face identification with multi-scale block-based correlation. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 1477–1481). IEEE.

Guo, Z., Zhang, L., Zhang, D., & Mou, X. (2010). Hierarchical multiscale LBP for face and palmprint recognition. In *2010 17th IEEE international conference on image processing (ICIP)* (pp. 4521–4524). IEEE.

Huang, W., & Yin, H. (2017). Robust face recognition with structural binary gradient patterns. *Pattern Recognition*, *68*, 126–140.

Kral, P., & Lenc, L. (2015). Confidence measure for experimental automatic face recognition system. In B. Duval, J. van den Herik, S. Loiseau, & J. Filipe (Eds.) *Agents and artificial intelligence: 6th international conference, ICAART 2014, Angers, France, March 6–8, 2014, revised selected papers* (pp. 362–378). Cham: Springer International Publishing.

Lawrence, S., Giles, C. L., Tsoi, A. C., & Back, A. D. (1997). Face recognition: A convolutional neural-network approach. *IEEE Transactions on Neural Networks*, *8*(1), 98–113.

Lenc, L., & Kral, P. (2012). Novel matching methods for automatic face recognition using sift. In: L. Iliadis, I. Maglogiannis, H. Papadopoulos (Eds.) *Artificial intelligence applications and innovations: 8th IFIP WG 12.5 international conference, AIAI 2012, Halkidiki, Greece, September 27–30, 2012, Proceedings, Part I* (pp. 254–263). Berlin: Springer.

Lenc, L., & Král, P. (2014). Automatically detected feature positions for LBP based face recognition. In *IFIP international conference on artificial intelligence applications and innovations* (pp. 246–255). Springer.

Lenc, L., & Kral, P. (2015a). Automatic face recognition system based on the sift features. *Computers & Electrical Engineering 46*, 256–272.https://doi.org/10.1016/j.compeleceng.2015.01.014. http://www.sciencedirect.com/science/article/pii/S0045790615000208.

Lenc, L., & Král, P. (2015b). Unconstrained facial images: Database for face recognition under real-world conditions. In *Mexican international conference on artificial intelligence* (pp. 349–361). Springer.

Lenc, L., & Král, P. (2015c). Unconstrained facial images: Database for face recognition under real-world conditions. In *14th Mexican international conference on artificial intelligence (MICAI 2015)*. Cuernavaca, Mexico: Springer.

Lenc, L., & Král, P. (2016). Local binary pattern based face recognition with automatically detected fiducial points. *Integrated Computer-Aided Engineering*, *23*(2), 129–139.

Li, F., & Wechsler, H. (2005). Open set face recognition using transduction. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, *27*(11), 1686–1697.

Li, F., & Wechsler, H. (2009). Face authentication using recognition-by-parts, boosting and transduction. *International Journal of Pattern Recognition and Artificial Intelligence*, *23*(3), 545–573.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94.

Lu, J., Liong, V. E., Zhou, X., & Zhou, J. (2015). Learning compact binary face descriptor for face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(10), 2041–2056.

Nanni, L., Lumini, A., & Brahnam, S. (2017). Ensemble of texture descriptors for face recognition obtained by varying feature transforms and preprocessing approaches. *Applied Soft Computing*, *61*, 8–16.

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. *In: BMVC*, *1*, p. 6.

Sanderson, C., & Lovell, B.C. (2009). Multi-region probabilistic histograms for robust and scalable identity inference. In International conference on biometrics (pp. 199–208). Springer.

Shafer, G., & Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research 9*, 371–421. http://dl.acm.org/citation.cfm?id=1390681.1390693.

Tan, X., & Triggs, B. (2007). Enhanced local texture feature sets for face recognition under difficult lighting conditions. In *International workshop on analysis and modeling of faces and gestures* (pp. 168–182). Springer.

Vovk, V., Fedorova, V., Nouretdinov, I., & Gammerman, A. (2016). Criteria of efficiency for conformal prediction pp. 23–39. https://doi.org/10.1007/978-3-319-33395-3_2.

Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic learning in a random world*. Berlin: Springer.

Vu, N. S., Dee, H. M., & Caplier, A. (2012). Face recognition using the poem descriptor. *Pattern Recognition*, *45*(7), 2478–2488.

Wen, Y., Zhang, K., Li, Z., & Qiao, Y. (2016). A discriminative feature learning approach for deep face recognition. In *European conference on computer vision* (pp. 499–515). Springer.

Xu, Y., Fang, X., Li, X., Yang, J., You, J., Liu, H., et al. (2014). Data uncertainty in face recognition. *IEEE Transactions on Cybernetics*, *44*(10), 1950–1961.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.