Conformal Prediction for Automatic Face Recognition

Charalambos Eliades PAMBOSELIADES@HOTMAIL.COM
Harris Papadopoulos H.PAPADOPOULOS@FREDERICK.AC.CY
Computer Science and Engineering Department, Frederick University,
7 Y. Frederickou St., Palouriotisa, Nicosia 1036, Cyprus

Editor: Alex Gammerman, Vladimir Vovk, Zhiyuan Luo, and Harris Papadopoulos

Abstract

Automatic Face Recognition (AFR) has been the subject of many research studies in the past two decades and has a wide range of applications. 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 for the detection of wanted persons or for performing post-processing in automatic annotation of photographs. This paper investigates the use of the Conformal Prediction (CP) framework for providing reliable confidence information for AFR. In particular we combine CP with two classifiers based on calculating similarities between images using Scale Invariant Feature Transformation (SIFT) features. We examine and compare the performance of several nonconformity measures for the particular task in terms of their accuracy and informational efficiency.

Keywords: Face Recognition, Scale Invariant Feature Transformation, Conformal Prediction, Confidence, Credibility, Prediction Regions, Uncontrolled Environment.

1. Introduction

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. 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 much 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. 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 probabilistically valid 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.

This work combines CP with two AFR techniques based on SIFT features, which have been shown to perform well in uncontrolled environments in the literature (Lowe, 2004). 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 particular AFR techniques, 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 Section 2 we provide an overview of related work on AFR and of previous work on obtaining confidence information for the particular task. Next, Section 3 gives a brief description of the general CP framework. In Section 4 we concentrate on the usage and calculation of SIFT features, while in Section 5 we discuss the two AFR techniques used as basis for the CPs proposed in this work. Section 6 details the developed NCMs and completes the description of the proposed CP approaches. Section 7 reports and discusses our experimental results. Finally, Section 8 gives our conclusions and plans for future work.

2. Related Work

The methods for AFR are commonly divided to holistic and feature-based ones. The holistic methods were popular mainly in the 90s. Such methods utilize for example Principal Component Analysis (PCA) (Turk and Pentland, 1991) or Fisher's Linear Discriminant (FLD) (Belhumeur et al., 1997) to project data from face space to a lower dimensional subspace.

The feature-based methods represent the face as a set of features and are usually more suitable for recent challenging AFR settings where the images are of uneven quality and show variances in appearance. A number of image descriptors have been used for 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. Other successful methods are for instance Patterns of Oriented Edge Magnitudes (POEM) (Vu et al., 2012) or Scale Invariant Feature Transform (SIFT) (Lowe, 2004). These approaches either divide the processed image using a rectangular grid and compute features for each region (Ahonen et al., 2004) or determine the feature points dynamically (Lenc and Král, 2016). The face representations are then compared against a gallery of known faces and the recognized person is determined by some distance measure or using the k-Nearest Neighbours (kNN) algorithm.

It is also worth to mention Artificial Neural Networks (ANN) that were used already in the work of Lawrence et al. (1997). 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. 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). Our work differs in the setting examined, but most importantly we additionally evaluate the informativeness of the outputs provided by CP and investigate the performance of alternative NCMs.

Other studies examining CMs in AFR include a pseudo 2-D Hidden Markov Model classifier with features created by the Discrete Cosine Transform (DCT) presented by 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.

3. Conformal Prediction

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

Let $A = \{(x_i, y_i)|i = 1, \ldots, N\}$ denote our training set, where x_i is an object given in the form of an input vector or matrix, $R = \{t_1, \ldots, t_c\}$ is the set of possible labels and $y_i \in R$ is the label of the corresponding input vector or matrix. Let $B = \{X_k|k = 1, \ldots, M\}$ 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 *non-conformity* 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 $C_{k,l}$. 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 NCM uses 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:

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

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 in (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 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. Sift Features

For the needs of this study we have extracted the (SIFT) features of images and we have used SIFT based methods in order to find similarity scores of a *Test* image with the *Gallery* images. The reason we have used SIFT features is because as mentioned before these features are invariant to image scaling, translation and rotation. Moreover, they are also partly invariant to changes in illumination and 3d camera viewpoint. Therefore, the use of SIFT features is beneficial for face recognition in real (uncontrolled) conditions where images may differ significantly. The SIFT algorithm consists of four steps: extrema detection, removal of key-points with low contrast, orientation assignment and descriptor calculation (Lowe, 2004).

4.1. Extrema Detection

Extrema detection constructs a Gaussian pyramid by applying in each level a Gaussian mask and then subsampling to reduce size. Each pixel in each level is the result of applying a Gaussian mask at the previous level and then subsample to get four images with 1/4th of the total number of pixels in the previous level. In each level Adjacent Gaussians are subtracted to produce the difference of Gaussians (DOG). This process is illustrated in

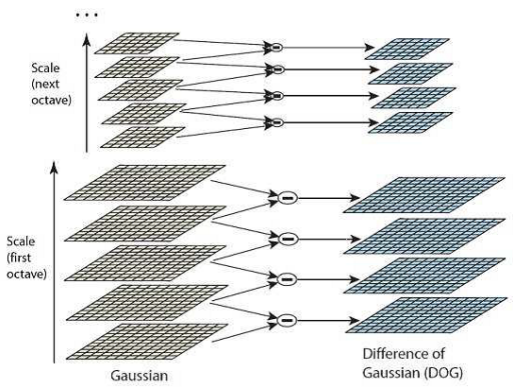


Figure 1: Difference of Gaussian filters at different scales (Lowe, 2004).

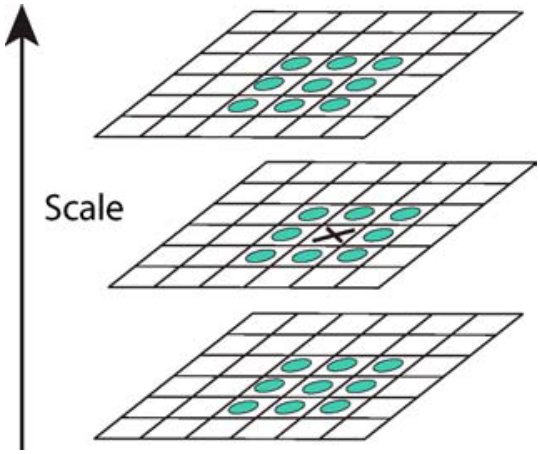


Figure 2: Maxima and minima of the difference-of-Gaussian images are detected by comparing a pixel (marked with X) to its 26 Neighbors in 3x3 regions at the current and adjacent scales (marked with circles) (Lowe, 2004).

Figure 1. Each pixel is then compared with eight neighbors in the current image and 9 neighbors in the scale above and below. A pixel is selected (key-point) if it is larger or smaller than all of the 26 neighbors compared. This comparison is illustrated in Figure 2.

4.2. Low Contrast Key-point Removal

Key-points that are low in contrast and are poorly localized along an edge will be excluded because they are sensitive to noise and therefore generally less reliable than the ones with high contrast.

4.3. 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, each value is weighted by the gradient magnitude.

4.4. Descriptor Calculation

The previous operations assigned to each key-point location, scale and orientation in a way that invariance to those parameters is ensured. The final step consists in the creation of descriptors for the local image region that is 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.

5. Automatic Face Recognition Techniques

In this work two conventional AFR techniques were considered and combined with CP. Both these techniques compare a test image with the available ones in our gallery and output a similarity score. We use the similarity scores provided by each technique to define the AFR-CP nonconformity measures.

Before application of the two techniques two preprocessing steps are applied to all images. The 1^{st} step is to detect the eyes in the images and then rotate them in such way that the eyes alignment will be parallel with the x-axis. Eye detection has been performed with the Viola Jones algorithm implemented in MATLAB (2016) using the VISION package. The 2^{nd} step is to extract the SIFT features from each image. The SIFT features have been extracted using the vlfeat library[1]. For more details see (Lowe, 2004).

In describing the techniques we use the following notation:

- We define as $\{(G_1, y_1), \ldots, (G_N, y_N)\}$ the gallery instances, where G_i are the SIFT features of image i in the gallery and $y_i \in \{t_1, \ldots, t_c\}$ is the person in the image.
- We define as $\{X_1, \ldots, X_M\}$ the sift features of the test images, where X_k corresponds to the SIFT features of image k.
- $q_m \in X_k$ *and* $Q_n \in G_i$ are vectors of SIFT features.
- $< x, z >$ denotes the dot product of vectors x and z in the Euclidean space.
- $|x|$ is the Euclidean norm if x is a vector of real values.
- $|C|$ is the cardinality of set C.
- L is the number of the best similarity scores of a class that we use in each algorithm to classify a person.

1. `http://www.vlfeat.org/matlab/vl_sift.html`

- $dT1$ is the distance threshold parameter of Algorithm 1.

The following subsections provide a description of the two conventional techniques. The general approach is based on calculating the similarities of images based on their feature distances and cosines. Given a test image we find its similarities with the images in the training set and classify it as the person who's images have the greatest similarity with the test image.

5.1. Partial Kepenekci Technique

Here we use part of the technique called Kepenekci Method described in (Lenc and Kral, 2015). The original Kepenekci method uses the percentage of the relevant vectors and their cosines. The Partial Kepenekci Technique (PKT) we use here uses only the percentage of the relevant vectors, since cosines are utilized in the 2^{nd} technique we consider.

Given the set of features X_k of an image k we find the similarity of the set X_k with each set G_i. Let $F = \{q_m : |Q_n - q_m| \leq dT1\}$ contain all the vectors of X_k that have distance to at least one vector of G_i that is less than or equal to $dT1$. The set above is computed with the build-in MATLAB (2016) function rangesearch.

The similarity measure of a test image k with a gallery image i is defined as

$$S_i^k = \frac{|F|}{|X_k|}. \tag{2}$$

In other words as the portion of the vectors in X_k that are within distance $dT1$ of at least one vector of G_i. The above similarity measure is similar to the one used in (Lenc and Kral, 2015). The only difference is that in the approach followed by Lenc and Kral (2015) the similarity measure is defined as $S_i^k = \frac{|F|}{|G_i|}$ and the set F contains all the vectors of G_i that have distance to at least one vector of X_k that is less than or equal to $dT1$.

5.2. Lenc-Kral Matching

This algorithm, called Lenc-Kral Matching (LKM), has been proposed in (Lenc and Kral, 2012). Given the set vectors of two images X_k *and* G_i we calculate the cosine similarity measure for each pair of vectors according to the formula

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

The algorithm consists of the following steps:

- Let $X_k = \{q_1, q_2, q_3, q_4, \ldots, q_{m'}\}$ *and* $G_i = \{Q_1, Q_2, Q_3, \ldots, Q_{n'}\}$, where m', n' is the number of SIFT features in X_k and G_i respectively.

- We find the following set of matrices

$$COS_i^k = \begin{matrix} cos(q_1\ , Q_1) & . & cos(q_1\ , Q_{n'}) \\ . & . & . \\ cos(q_{m'}\ , Q_1) & . & cos(q_{m'}\ , Q_{n'}) \end{matrix}$$

The matrix COS_i^k contains the cosines of each pair between the features of a test image and the features of the gallery image. It should be noted that the calculation of COS_i^k can be done by normalizing each vector q_m , Q_n in X_k and G_i respectively and simply multiplying the two matrices $G_i'.X_k$.

The similarity measure of a test image k and a gallery image i is defined as

$$S_i^k = mean(max(COS_i^k)), \tag{3}$$

where the $max(COS_i^k)$ is a vector that consists the maximum among the rows of COS_i^k.

In both techniques, a face k is recognized as the t_r with

$$r = \arg\max_j \ D_j^k, \tag{4}$$

where

$$D_j^k = \frac{1}{L}\sum_{i=1}^{L} \tilde{S}_i^{k,j}, \tag{5}$$

where $\tilde{S}^{k,j}$ are the S^k for the images corresponding to person t_j sorted in descending order. In other words, D_j^k is the mean of the L highest similarities of X_k to the images of person t_j in the gallery.

6. Nonconformity measures for FR-TCP

In this section we provide a description of the NCMs we have used in this study, these measures are based on the two classifiers described in Section 5. We have examined several NCMs to investigate which of them provides the most informative p-values. Recall from Section 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 TCP generates $C_{k,1}, \ldots, C_{k,c}$ and assigns a NCS to each example in each of the c sets. We denote as $z_s^{k,l}$ the sth element of $C_{k,l}$ and as $\alpha_s^{k,l}$ its NCS, with $s = 1, \ldots, N, new$.

For defining the NCM $\alpha_s^{k,l}$ we use the D_j^s calculated by each underlying AFR technique (see equation 5) with $\{z_i^{k,l} : i = 1, \ldots, s-1, s+1, N, new\}$ as gallery. In other words D_j^s is the mean of the L highest similarities of $z_s^{k,l}$ with all other elements of $C_{k,l}$ corresponding to person t_j. Our NCMs are defined in such way to contain at least one of two quantities: The first quantity, D_j^s where $y_s = t_j$, summarizes the similarity of the instance s with the other images of 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 1^{st} NCM for an image s corresponding to person t_j is defined as

$$\alpha_s^{k,l} = \max_{i \neq j}(D_i^s) - D_j^s, \tag{6}$$

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

The 2^{nd} NCM is defined as

$$\alpha_s^{k,l} = \frac{\max_{i \neq j}(D_i^s)}{D_j^s}, \tag{7}$$

where the same quantities as in (6) are used, but now subtraction is replaced by division.

The 3^{rd} NCM is defined as

$$\alpha_s^{k,l} = \underset{i \neq j}{\text{mean}}(D_i^s) - D_j^s, \tag{8}$$

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

The 4^{th} NCM is defined as

$$\alpha_s^{k,l} = \frac{\text{mean}_{j \neq i}(D_j^s)}{D_i^s}, \tag{9}$$

where the same quantities as in (8) are used, but now subtraction is replaced by division.

The 5^{th} NCM is defined as

$$\alpha_s^{k,l} = 1/(D_i^s), \tag{10}$$

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

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 Section 3.

7. Experiments and Results

In this section we detail the experiments and results of the proposed AFR-TCPs and of the two conventional AFR techniques used as underlying models for our TCPs on the Database of Faces of the AT&T Laboratories Cambridge[2] (Samaria and Harter, 1994) and on a subset of the UFI corpus (Lenc and Král, 2015). The AT&T dataset consists of ten different images for each of 40 distinct subjects, the size of each image is 92×112 pixels, with 256 grey levels per pixel. The UFI corpus subset consists of 40 persons with an average number of 9.1 images per person taken in an uncontrolled environment. The size of each image is 128×128 pixels, with 256 grey levels per pixel. Example images from the two datasets are shown in Figures 3 and 4.

7.1. Experimental Setting and performance Measures

Our experiments on the AT&T dataset were performed following a 10-fold cross-validation process (each fold consisting of one image per subject), while the parameters L and $dT1$ used for each fold were selected through nested cross-validation; i.e. by performing 9-fold cross-validation on the nine images per subject contained in each training set. For both techniques the parameter L was optimized by searching in the range of integers from 1 to 8, while the parameter $dT1$ for PKT was optimized by searching in the range $[0.5, 5]$ with a step of 0.5. To avoid random effects this process was repeated 100 times and all results reported here are the averages over all 100 repetitions.

2. Available at: http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

(a) (b) (c)

Figure 3: Example images from the AT&T dataset.

(a) (b) (c)

Figure 4: Example images from the UFI corpus.

In the case of the UFI corpus subset we have used 20 random divisions of the images into training and test set and averaged their results. For each division one image of each person was selected at random for the test set while the remaining images were used to form the training set. Thus the training set was comprised of 324 images in total and the test set was comprised of 40 images. The parameters of the two AFR techniques were selected from the same L and $dT1$ values mentioned above based on their leave-one-out performance on the training set; leave-one-out was used due to the fact that the number of images per subject varied. It should be noted that while the parameters were selected based on the performance of the conventional AFR techniques, the corresponding AFR-TCPs used the same parameters (as their underlying technique) for consistency.

Due to the fact that the accuracy itself is not a good indication for the choice of a NCM we used the 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 are:

- The S-criterion

$$\frac{1}{M}\sum_{l=1}^{c}\sum_{k=1}^{M} p_k(t_l), \tag{11}$$

where $p_k(t_l)$ is the p-value of the test example X_k with candidate label t_l as in equation (1). In effect the S-criterion is the average sum of all p-values.

- The N-criterion

$$\frac{1}{M}\sum_{k=1}^{M} |\{t_l | p_k(t_l) > \delta\}|, \tag{12}$$

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

The two Observed Criteria are:

- The OF-criterion

$$\frac{1}{(c-1)M}\sum_{k=1}^{M}\sum_{l,t_l\neq t_k} p_k(t_l), \tag{13}$$

which is the average of the p-values of the false labels.

- The OE-criterion

$$\frac{1}{M}\sum_{k=1}^{M} |\{t_l | p_k(t_l) > \delta, t_l \neq t_k\}|, \tag{14}$$

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.

7.2. AT&T Faces Results

7.2.1. Accuracy

Table 1 presents the accuracy of the two conventional AFR techniques, while Table 2 reports the accuracy of the corresponding AFR-TCP techniques along with the average confidence and credibility measures using the five NCMs defined in Section 6. The results reported in these tables show that all techniques perform very well on the particular dataset. The performance of the TCP approaches is slightly lower than that of their conventional counterparts, but this difference is extremely small to be of significance, especially considering the additional information provided by the proposed approaches. The average confidence measures reported in Table 2 are in all cases high reflecting the high certainty in the CP predictions. The differences between the five nonconformity measures are again too small to be of any significance.

Table 1: Average accuracy and standard deviation of each conventional AFR technique on the AT&T dataset.

Classifier	Mean Accuracy(%)	Std (%)
PKT	97.90	0.31
LKM	97.34	0.46

Table 2: Average accuracy, credibility and confidence of the AFR-TCPs on the AT&T dataset.

	Underlying	Nonconformity measure				
	Technique	(6)	(7)	(8)	(9)	(10)
Accuracy	PKT	97.69	97.65	97.65	97.66	97.71
	LKM	97.14	97.14	97.06	97.09	97.18
Average confidence	PKT	99.39	99.63	98.53	96.02	96.57
	LKM	99.58	99.58	97.71	97.67	91.94
Average credibility	PKT	57.40	54.95	57.59	57.24	55.96
	LKM	56.08	56.16	56.25	56.77	53.29

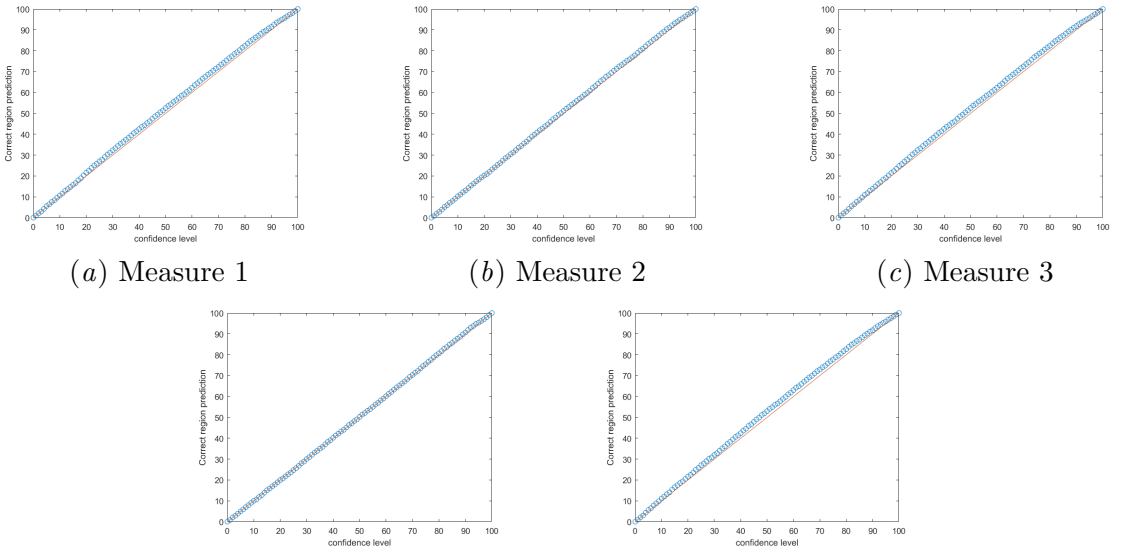
7.2.2. Empirical Validity

In this subsection we examine the empirical validity of the prediction regions produced by the proposed techniques. Figures 5 and 6 present the percentage of correct region predictions as a function of the confidence level for the five NCMs we used on top of the PKT and LKM techniques respectively. In both cases the plots follow (and are slightly above) the diagonal indicating that the produced region predictions are always well-calibrated (the accuracy is equal to or slightly higher than the required confidence level), as guaranteed by CP.

7.2.3. 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 Subsection 7.1 and proposed in (Vovk et al., 2016).

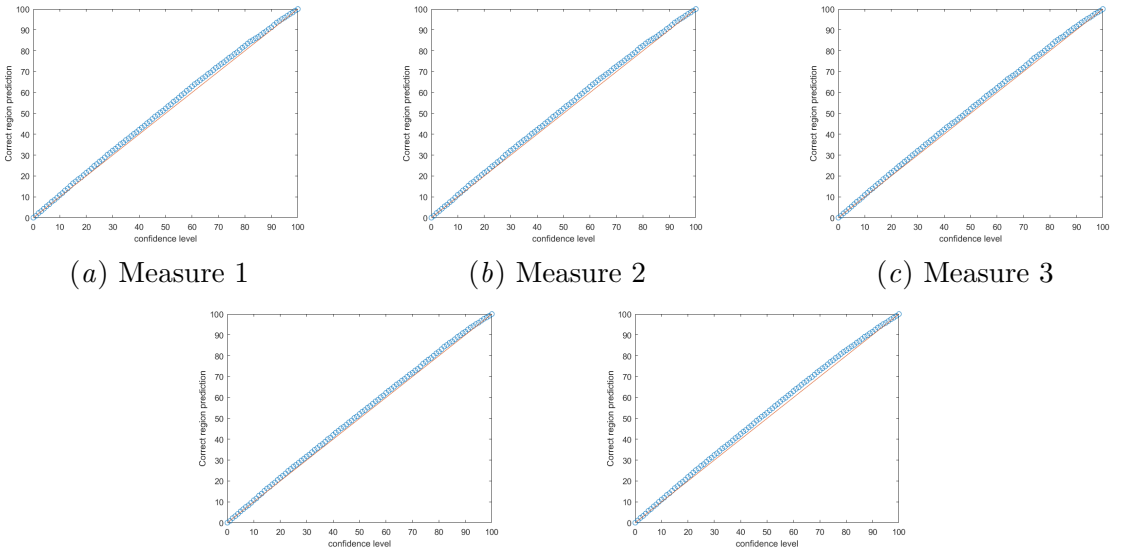
Table 3 presents the values of the two unobserved criteria for the AFR-TCPs with the five NCMs. Specifically the second column of the table contains the values of the S criterion, while the rest of the columns present the N criterion for the significance levels $0.01, 0.05, 0.1, 0.15$ and 0.2. In the 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 $0.01, 0.05, 0.10, 0.15, 0.2$.



(*a*) Measure 1 (*b*) Measure 2 (*c*) Measure 3

(*d*) Measure 4 (*e*) Measure 5

Figure 5: Percentage of correct region predictions of the five NCMs with the PKT underlying technique on the AT&T dataset.



(*a*) Measure 1 (*b*) Measure 2 (*c*) Measure 3

(*d*) Measure 4 (*e*) Measure 5

Figure 6: Percentage of correct region predictions of the five NCMs with the LKM underlying technique on the AT&T dataset.

Table 3: Unobserved criteria on the AT&T dataset.

Underlying Technique	NC Measure	S criterion	N criterion (per significance level) 0.01	0.05	0.10	0.15	0.20
PKT	(6)	0.63	1.41	0.96	0.92	0.87	0.82
	(7)	0.62	1.43	0.96	0.91	0.86	0.81
	(8)	0.73	5.43	1.15	0.93	0.88	0.83
	(9)	0.71	4.55	1.39	0.96	0.88	0.82
	(10)	1.01	14.94	2.28	1.13	0.94	0.85
LKM	(6)	0.63	1.34	0.97	0.91	0.87	0.82
	(7)	0.63	1.33	0.97	0.91	0.87	0.82
	(8)	0.71	4.95	1.13	0.92	0.87	0.82
	(9)	0.71	4.94	1.13	0.93	0.87	0.82
	(10)	1.49	24.52	6.08	2.15	1.23	0.94

Table 4: Observed criteria on the AT&T dataset.

Underlying Technique	NC Measure	OF criterion	OE criterion (per significance level) 0.01	0.05	0.10	0.15	0.20
PKT	(6)	0.0026	0.42	0.007	0.001	0	0
	(7)	0.0025	0.44	0.008	0	0	0
	(8)	0.0036	4.44	0.243	0.011	0.004	0
	(9)	0.0400	3.56	0.51	0.062	0.020	0.011
	(10)	0.0065	13.94	1.48	0.224	0.066	0.023
LKM	(6)	0.0025	0.35	0.0139	0	0	0
	(7)	0.0025	0.34	0.0140	0	0	0
	(8)	0.0036	3.96	0.200	0.009	0.004	0.0010
	(9)	0.0037	3.95	0.200	0.010	0.005	0.0013
	(10)	0.0121	23.53	5.732	1.320	0.380	0.1200

The results reported in the two tables show that the differences between the two underlying techniques are insignificant. The main comparison is between the performance of the different NCMs. The first two NCMs outperform the other three in all criteria, while there is no significant difference between the two. The values of the N and OE criterion for the NCMs (6) and (7) demonstrate the practical usefulness of the produced prediction regions since on average they contain less than 1.5 labels out of the possible 40 and less than 0.5 wrong labels out of the possible 39 for a confidence level as high as 99%. By lowering the confidence level to 95% we can be certain in a single label for almost all test examples. Note that the reason the N criterion has values below 1 is that there are some empty prediction regions, which are of no concern since we can add to them the classification with the highest p-value without affecting validity; empty prediction regions are part of the percentage of errors allowed at a given confidence level.

Table 5: Average accuracy of each conventional AFR technique on the UFI corpus subset.

Classifier	Training Set	Test Set
PKT	61.13	67.25
LKM	55.02	60.50

Table 6: Average accuracy, credibility and confidence of the AFR-TCPs on the UFI corpus subset.

	Underlying	Nonconformity measure				
	Technique	(6)	(7)	(8)	(9)	(10)
Accuracy	PKT	66.75	66.75	67.25	67.12	66.25
	LKM	60.25	60.25	60.25	60.00	59.87
Average credibility	PKT	59.87	59.49	60.69	61.52	67.45
	LKM	61.24	61.31	59.55	59.46	69.68
Average confidence	PKT	69.86	71.12	65.00	63.25	49.89
	LKM	63.29	63.21	62.77	62.88	42.91

7.3. UFI Corpus Subset Results

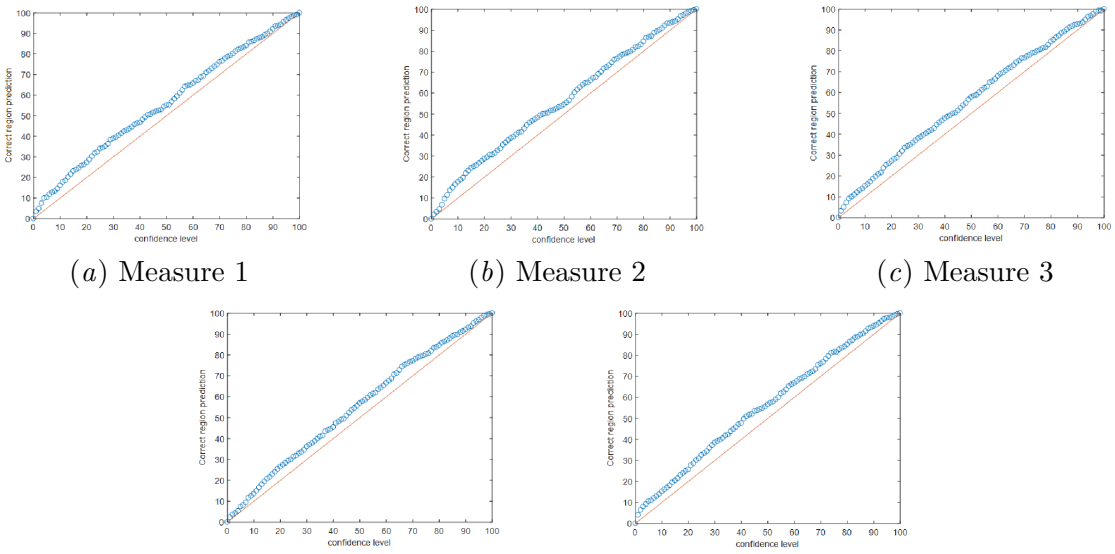
7.3.1. Accuracy

Table 5 presents the accuracy of the two conventional AFR techniques on the training and test sets of the UFI corpus subset, while Table 6 reports the accuracy of the corresponding AFR-TCP techniques along with the average confidence and credibility measures on the test set. Given the fact that the images of the UFI corpus were taken in an uncontrolled environment the performance of all the techniques is dramatically reduced compared to the one of the AT&T dataset. This lower performance indicates the need for quantifying the high uncertainty involved in uncontrolled environment face recognition, especially taking into consideration the large number of possible labels involved.

In comparing the results reported in the two tables we can observe that the PKT technique outperforms LKM. Again, in some cases the performance of the AFR-TCPs is slightly lower than that of the conventional AFR techniques on which they are based, but this difference is extremely small. The much lower average confidence measures of the AFR-TCPs suggest once again that the uncertainty involved in the particular task is very high. A comparison between the performance of the five NCMs shows that in terms of the values reported in Table 6 there is no significance difference among them.

7.3.2. Empirical Validity

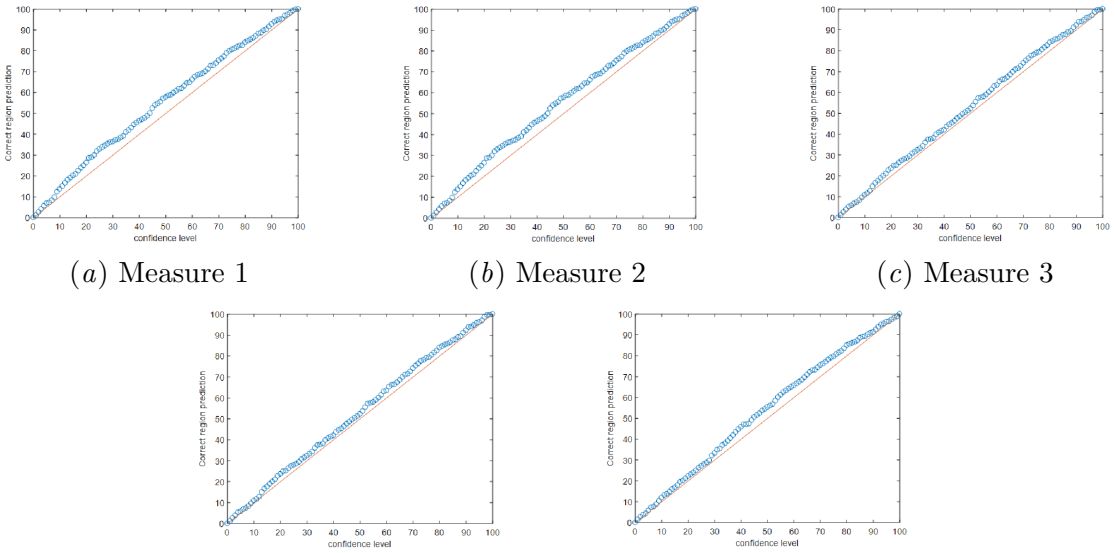
Figures 7 and 8 plot the percentage of correct region predictions against confidence level for each of the five NCMs of AFR-TCP combined with the PKT and LKM techniques respectively. The plots are in all cases slightly higher than the diagonal, confirming the



(*a*) Measure 1 (*b*) Measure 2 (*c*) Measure 3

(*d*) Measure 4 (*e*) Measure 5

Figure 7: Percentage of correct region predictions of the five NCMs with the PKT underlying technique on the UFI corpus subset.



(*a*) Measure 1 (*b*) Measure 2 (*c*) Measure 3

(*d*) Measure 4 (*e*) Measure 5

Figure 8: Percentage of correct region predictions of the five NCMs with the LKM underlying technique on the UFI corpus subset.

Table 7: Unobserved criteria on the UFI corpus subset.

Underlying Technique	NC Measure	S-criterion	N-criterion (per significance level) 0.01	0.05	0.10	0.15	0.20
PKT	(6)	0.0705	23.98	13.75	8.97	5.79	3.75
	(7)	0.0794	29.29	16.44	10.02	6.31	4.00
	(8)	0.1116	35.56	23.04	14.65	9.85	6.77
	(9)	0.1052	33.58	23.64	14.31	8.76	5.38
	(10)	0.1733	36.21	26.43	20.84	17.31	14.14
LKM	(6)	0.1097	34.36	21.68	13.90	10.03	7.07
	(7)	0.1103	34.35	21.67	14.06	10.13	7.11
	(8)	0.1176	34.49	22.95	16.06	11.08	7.61
	(9)	0.1174	34.5237	22.97	16.04	11.04	7.62
	(10)	0.2145	36.85	32.06	25.93	22.43	17.25

Table 8: Observed criteria on the UFI corpus subset.

Underlying Technique	NC Measure	OF-criterion	OE-criterion (per significance level) 0.01	0.05	0.10	0.15	0.20
PKT	(6)	0.3213	22.99	12.79	8.05	4.91	2.91
	(7)	0.3067	28.29	15.47	9.08	5.41	3.1462
	(8)	0.3701	34.57	22.08	13.72	8.96	5.92
	(9)	0.3991	32.58	22.67	13.39	7.87	4.53
	(10)	0.5377	35.22	25.46	19.9	16.41	13.28
LKM	(6)	0.3929	33.36	20.71	12.97	9.15	6.23
	(7)	0.3940	33.35	20.7	13.13	9.25	6.26
	(8)	0.3812	33.49	21.99	15.13	10.21	6.77
	(9)	0.3809	33.53	22.01	15.12	10.16	6.78
	(10)	0.5820	35.86	31.10	25.02	21.54	16.40

guarantee of CP that the accuracy of the produced prediction regions will be equal to or higher than the corresponding confidence level. The fact that these curves are higher than the diagonal, rather than following it more closely, indicates that there is room for improvement in the NCMs and underlying AFR techniques used.

7.3.3. Informational Efficiency

The most important evaluation and comparison of the different NCMs is in term of the informational efficiency of the corresponding AFR-TCPs. Tables 7 and 8 report the performance of the AFR-TCPs in terms of the two unobserved and the two observed efficiency criteria described in Subsection 7.1 respectively. The values reported in these tables suggest that, as in the case of the AT&T dataset, the NCMs (6) and (7) perform better with both underlying techniques. Overall the PTK underlying technique combined with (6) seems to

perform best (on all criteria except the OF criterion). As expected, given the uncontrolled nature of the images, the prediction regions produced by the TCP are much larger than the ones produced for the AT&T dataset. Still at the 95% confidence level the resulting prediction regions contain on average about one third of the possible persons in the corpus, while at the 80% confidence level, which is well above the obtained accuracy, they contain on average less than one tenth of the possible persons. This is arguably a good result considering the high difficulty of the task and the very low accuracy of conventional AFR techniques.

8. Conclusions

We examine the application of CP for AFR based on two techniques that calculate the similarity between images using SIFT features. 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. We have developed five NCMs for combining CP with two base AFR techniques and investigated their performance experimentally on the AT&T dataset and a subset of the UFI corpus; the latter consisting of images taken in an uncontrolled environment.

Our experimental results show that in terms of accuracy the proposed approaches are comparable with 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 CP 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 PTK underlying technique combined with (6) and (7) as NCM seems to perform best. The prediction sets produced by this approach for the AT&T dataset are small even for high confidence levels. In the case of the much more difficult UFI corpus, where the images are taken in an uncontrolled environment, the prediction sets are much larger. However considering the difficulty of the task combined with the large set of possible persons, the resulting prediction sets can be very useful in the the manual classification process by significantly reducing the number of candidate persons for each image.

Our future plans include examining alternatives and improvements to the conventional AFR techniques and NCMs used and investigating 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.

References

Timo Ahonen, Abdenour Hadid, and Matti Pietikäinen. Face recognition with local binary patterns. In Tomás Pajdla and Jiří Matas, editors, *Computer Vision - ECCV 2004: 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part I*, pages 469–481, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg. ISBN 978-3-540-24670-1. doi: 10.1007/978-3-540-24670-1_36. URL http://dx.doi.org/10.1007/978-3-540-24670-1_36.

Peter N. Belhumeur, João P. Hespanha, and David J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1997.

Stefan Eickeler, Mirco Jabs, and Gerhard Rigoll. Comparison of confidence measures for face recognition. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 257–262. IEEE, 2000.

Pavel Kral and Ladislav Lenc. Confidence measure for experimental automatic face recognition system. In Béatrice Duval, Jaap van den Herik, Stephane Loiseau, and Joaquim Filipe, editors, *Agents and Artificial Intelligence: 6th International Conference, ICAART 2014, Angers, France, March 6-8, 2014, Revised Selected Papers*, pages 362–378, Cham, 2015. Springer International Publishing. ISBN 978-3-319-25210-0. doi: 10.1007/978-3-319-25210-0_22. URL http://dx.doi.org/10.1007/978-3-319-25210-0_22.

Steve Lawrence, C Lee Giles, Ah Chung Tsoi, and Andrew D Back. Face recognition: A convolutional neural-network approach. *IEEE transactions on neural networks*, 8(1): 98–113, 1997.

L. Lenc and P. Král. Unconstrained Facial Images: Database for face recognition under real-world conditions. In *14th Mexican International Conference on Artificial Intelligence (MICAI 2015)*, Cuernavaca, Mexico, 25-31 October 2015 2015. Springer.

Ladislav Lenc and Pavel Kral. Novel matching methods for automatic face recognition using sift. In Lazaros Iliadis, Ilias Maglogiannis, and Harris Papadopoulos, editors, *Artificial Intelligence Applications and Innovations: 8th IFIP WG 12.5 International Conference, AIAI 2012, Halkidiki, Greece, September 27-30, 2012, Proceedings, Part I*, pages 254–263, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-33409-2. doi: 10.1007/978-3-642-33409-2_27. URL http://dx.doi.org/10.1007/978-3-642-33409-2_27.

Ladislav Lenc and Pavel Kral. Automatic face recognition system based on the sift features. *Computers & Electrical Engineering*, 46:256 – 272, 2015. ISSN 0045-7906. doi: http://dx.doi.org/10.1016/j.compeleceng.2015.01.014. URL http://www.sciencedirect.com/science/article/pii/S0045790615000208.

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

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

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

David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. ISSN 1573-1405. doi: 10.1023/B:VISI.

0000029664.99615.94. URL http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94.

MATLAB. Version 9.10.0 (r2016b), 2016.

Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *British Machine Vision Conference*, volume 1, 2015.

F. S. Samaria and A. C. Harter. Parameterisation of a stochastic model for human face identification. In *Proceedings of 1994 IEEE Workshop on Applications of Computer Vision*, pages 138–142. IEEE, Dec 1994. doi: 10.1109/ACV.1994.341300.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *J. Mach. Learn. Res.*, 9:371–421, June 2008. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1390681.1390693.

Xiaoyang Tan and Bill Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. In *International Workshop on Analysis and Modeling of Faces and Gestures*, pages 168–182. Springer, 2007.

Matthew A. Turk and Alex P. Pentland. Face recognition using eigenfaces. In *IEEE Computer Society Conference on In Computer Vision and Pattern Recognition*. Computer Vision and Pattern Recognition, 1991.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.

Vladimir Vovk, Valentina Fedorova, Ilia Nouretdinov, and Alexander Gammerman. Criteria of efficiency for conformal prediction. In Alexander Gammerman, Zhiyuan Luo, Jesús Vega, and Vladimir Vovk, editors, *Conformal and Probabilistic Prediction with Applications: 5th International Symposium, COPA 2016, Madrid, Spain, April 20-22, 2016, Proceedings*, pages 23–39, Cham, 2016. Springer International Publishing. ISBN 978-3-319-33395-3. doi: 10.1007/978-3-319-33395-3_2. URL http://dx.doi.org/10.1007/978-3-319-33395-3_2.

Ngoc-Son Vu, Hannah M Dee, and Alice Caplier. Face recognition using the poem descriptor. *Pattern Recognition*, 45(7):2478–2488, 2012.