Preliminary results on nonparametric facial occlusion detection

Daniel López Sánchez[a] and Angélica González Arrieta[a]

[a]Computer Sciences Department, University of Salamanca, 1 Escuelas St., Salamanca, 37003 lopelh@gmail.com

KEYWORD

Face recognition; oclusion detection;

ABSTRACT

The problem of face recognition has been extensively studied in the available literature, however, some aspects of this field require further research. The design and implementation of face recognition systems that can efficiently handle unconstrained conditions (e.g. pose variations, illumination, partial occlusion...) is still an area under active research. This work focuses on the design of a new nonparametric occlusion detection technique. In addition, we present some preliminary results that indicate that the proposed technique might be useful to face recognition systems, allowing them to dynamically discard occluded face parts.

1. Introduction

This paper focuses on the design and implementation of new nonparametric facial occlusion detection algorithms. In the context of machine learning, and according to (Murphy, 2012), a model or algorithm is considered nonparametric when the number of parameters grow with the amount of training data. The proposed method is closely related to the nearest neighbour search family, and takes advantage of the local nature of feature extractors like Local Binary Patterns Histogram (LBPH)(Ojala et al., 1996). Our major goal is that the proposed method does not rely on the availability of training data with occlusion conditions, given that even when such information is present, we can not guarantee that the nature of the occlusion that the system will face in production stage will be the same.

In section 2, we provide a comprehensive review of the literature in the field of facial occlusion detection and face recognition under occlusion conditions. Section 3 focuses on the preprocessing of facial images and describes the proposed method. Experimental results that validate the proposed method are reported in section 4. Finally, in section 5, we discuss the results and propose the future work lines.

2. Related work

Although the problem of face recognition has been extensively studied in literature, only a few authors have focused on the problems of face occlusion detection and face recognition under occlusion conditions. In this section, some of the most relevant studies that address these problems are reviewed.

The author of (Ekenel, 2009) supports the idea that most of the accuracy loss that face recognition systems suffer under occlusion conditions is caused by the error induced by occlusion on the automatic alignment process. To address this problem, he proposes a technique that tries to minimize a distance metric between each training sample and a new observation whose label or class must be predicted; this method does so by brute-force testing a finite set of alignment variations. This method reported the best recognition rates for the ARFace database (Martinez, 1998). The major drawback of this technique is the computational complexity: the brute force search of the optimum alignment configuration implies the evaluation of hundreds of configurations for each test sample.

Other authors (Martínez, 2002; Tan et al., 2005) follow a similar approach to ours, they divide facial images in blocks based on an uniform grid. Then, they model those individual blocks or areas of the image respectively by means of the Principal Component Analysis (PCA) dimensionality reduction algorithm and a Self Organizing Map (SOM).

Most of the proposed approaches in literature include a previous step to recognition, namely the detection of facial regions or blocks affected by occlusion. Some authors propose using patches of facial images, manually labelled as occluded or non-occluded, to train a classifier to distinguish between occluded and non-occluded regions (Min et al., 2011). Although effective, this approach requires the availability of a training set of images that contain occlusion, and thus assumes that the nature (i.e. texture patters, distribution, location...) of the occlusion that the system will face in production stage will be the same as in the training set.

Some other papers adopt a color-based segmentation approach (Jia and Martinez, 2008). The major problem that this systems suffer is a high sensibility to illumination conditions. In addition, they do not take into consideration the inherent variability of human skin color, and do not consider the scenario where the occlusion is caused by an artefact of the same color as human skin.

More recently, numerous papers have been publish studying the applicability of deep learning models to the field of face recognition. These models take advantage of the overwhelming amounts of information that are becoming available in the recent years, as well as the computational power of massively parallel computation devices. Classification models trained with such techniques gain the ability to correctly classify facial images under unconstrained conditions (e.g. uneven illumination, pose and partial occlusion).

Currently, the state of the art on one of the most famous databases for face recognition evaluation, Labelled Faces in the Wild (LFW), is held by the researchers of Baidu company (Liu et al., 2015). Essentially, the authors propose an architecture where several deep convolutional neural networks are trained on patches of different face regions, then the activations of the final convolutional layer of each network are combined to compose the final face descriptor. Deep learning approaches are in general very effective and unleash unprecedent accuracy rates, however they require large databases with hundreds of images per individual and have prohibitive computational costs in both training and test phases.

3. Proposed method

In this section we describe the proposed method for face occlusion detection (i.e. classification of facial regions as occluded or unoccluded). When a new image is presented to the system, it is processed in the following manner:

1. A region of interest (ROI) is decided by some face detection algorithm.
2. The face is aligned by means of a face alignment technique (i.e. the locations of several facial key points are estimated).
3. A pose normalization technique is applied to standardize the image to a size of 160 by 160 pixels.
4. An illumination normalization technique is applied, aiming to reduce the variance generated by uneven illumination conditions.
5. The chosen local feature extractor is applied to generate a descriptor of the facial image.
6. The proposed method for facial image occlusion detection is executed over the descriptor, classifying each local region of the image as occluded or not occluded.

The proposed method for occlusion detection is proposed in section .

3.1 Preprocessing

This sections describes the different image preprocessing steps that the proposed method applies before occlusion detection.

3.1.1 Face detection

The goal of face detection methods is to find an approximated region of interest for the human face present in the image. A linear classifier, namely a Support Vector Machine (SVM) with a linear kernel, is trained on Histogram of Oriented Gradients (HOG) features. The training is performed according to a hard negative mining scheme to minimize the amount of false positives. At test phase, the classifier is fed with patches of the image extracted by an sliding window, this is done at various scales and the results are used to compose the final region of interest.

For the experiments in this paper, the implementation of HOG object detector provided by Dlib C++ library (King, 2009) has been applied.

3.1.2 Face alignment

The process of face alignment or facial key point detection consists of automatically estimating the location of a number of facial key points (such as eyes, nose, mouth corners, chin...) based on the facial image and an estimated ROI. Currently, the most popular models are bases on the idea of cascade regression; here, several base regressors are trained to refine and improve the estimation of the previous one. In this paper, the chosen technique for face alignment is the one proposed by (Kazemi and Sullivan, 2014). The author proposes using ensembles of regression trees as the base regressor technique for the cascade.

This technique has been used for the face alignment step in all the experiments of this paper. The implementation provided by the Dlib C++ (King, 2009) library has been used.

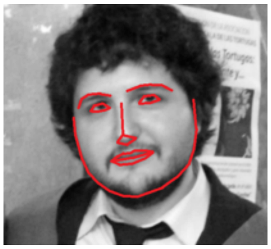

Figure 1: Automatic estimation of facial key points.

3.1.3 Pose normalization

A very simple approach is used for pose normalization. First the face is rotated upright based on the alignment results. Then the face region is cropped and the image is normalized to an standard size of 160 by 160 pixels.

3.1.4 Illumination normalization

Illumination normalization methods aim to reduce the amount of variation induced on images by illumination conditions, given that illumination characteristics do not contain useful information about the identity of the individual in the picture.

One of the most simple methods for illumination normalization is known as Histogram Equalization (HE). With this method, the pixel intensity values of the original image are mapped to a more uniform distribution in the valid range of values. To achieve this, the so called cumulative distribution function $H'(i)$ is used:

$$H'(i) = \sum_{j=0}^{i} H(j) \tag{1}$$

Once $H'(i)$ has been computed, it is normalized so its maximum value corresponds with the maximum valid intensity value for a pixel. Then, this function is used to compute the intensity values of pixels in the resulting image:

$$equalized(x, y) = H'(original(x, y)) \tag{2}$$

HE has been used to normalize the illumination in all the experiments of this paper.

3.2 Feature extraction: Local Binary Patterns

In practise, it is not useful to train classification models directly on pixel intensities. The principal reason for that is the hight dimensionality of this representation and the amount of needless information it contains. It has been proved that, for a fixed number of training samples, an increase in the dimensionality of those samples above some threshold will reduce the prediction capabilities (Hughes, 1968).

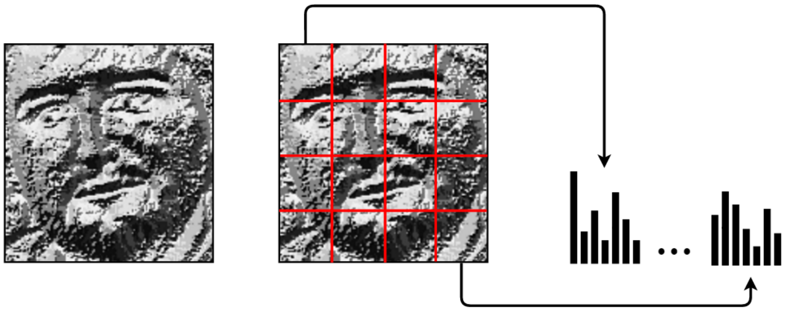

Figure 2: From left to right: LBP image, LBP image divided in blocks and LBPH descriptor obtained by concatenation of local histograms.

The method chosen in this work is the local feature descriptor known as Local Binary Patterns (LBP) (Ojala et al., 1996). As explained in the following sections, its local nature will help to isolate the features extracted form occluded face regions. In addition to that, it is often adopted by its low computational complexity. More details about this feature extractor and its calculation are provided below.

The LBP descriptor, labels the pixels in an image by considering the difference of intensity with its neighbours. The use of a circular neighbourhood and bilinear interpolation of non integer pixel coordinates allows to compute de operator with ant radius and neighbour number (Ojala et al., 2002). $LBP_{P,R}$ denotes the LBP operator parametrized with P neighbours and a radius of length R.

In (Ojala et al., 2002) it was proven that when using the $LBP_{8,1}$ operator, 90% of the extracted patterns were uniform patterns (i.e. its binary representation contained at most two transitions between zero and one or vice versa). For this reason, they proposed a new LBP representation where every uniform pattern has its own label, but all the not uniform patterns were assigned a common label. From now on, we will refer to that operator as $LBP^{u}_{P,R}$.

Rather than directly training a classifier on this raw LBP descriptor, is a common previous step to reduce the dimensionality of the descriptor and at the same time capture the structural information by computing the histograms of local regions of the LBP image. Usually, the image is divided into several blocks by an uniform grid, and then the histograms are computed for each block, counting the number of occurrences of each pattern. Finally all the histograms are concatenated to form the final descriptor, which is known as Local binary pattern histograms (LBPH). Figure 2 shows the previously described process of LBPH descriptor extraction from a LBP image.

3.3 Occlusion detection

The fundamental proposal of this work consist of the definition of a technique capable of detecting LBP blocks that correspond to occluded face regions.

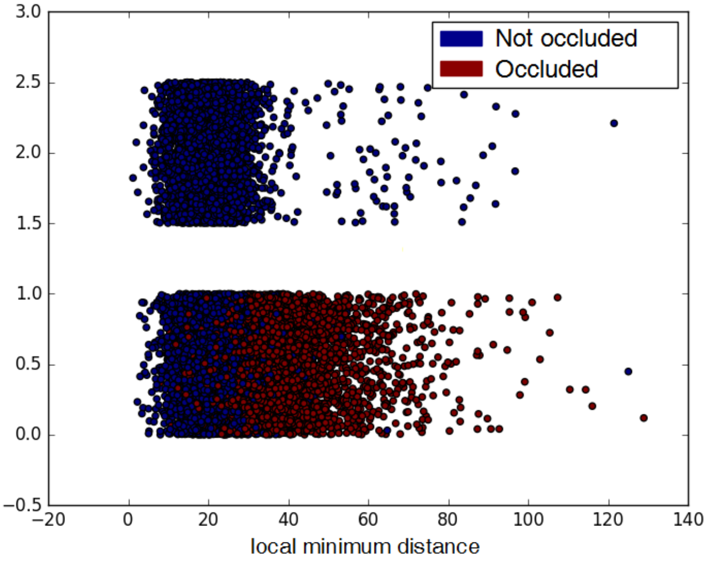

Figure 3: Distribution of local minimum distance for blocks form unclouded facial images (top) and from partially occluded facial images (bottom). Occluded blocks are drawn in red and not occluded blocks in blue. We added random noise in the vertical axis to avoid overlapping.

This proposal arises looking at the very nature of the LBPH descriptor. This descriptor keeps features isolated according to the block from which they were extracted. This allows the identification of occluded features, as they do not merge with valid features.

Our method is based in the idea of local minimum distance. We compute local minimum distance for the histogram of a block as the minimum distance found when comparing this histogram with every histogram of the training dataset that corresponds to the same facial region (i.e. the same coordinates in the LBPH grid of blocks). Then, the only assumption made by our detector about the nature of the occlusion is that blocks corresponding to occluded face regions have a greater local minimum distance values. To validate this hypothesis, we can visualize the distribution of local minimum distances for occluded and not occluded blocks extracted from two sets of images, the first without facial occlusion and the second presenting partial facial occlusion (see figure 3). As we can see, it is possible to find a suitable threshold that discards most of the occluded blocks, leaving out only an small fraction of the not occluded blocks.

Formally we propose the following: let S be a set of n LBPH descriptors of facial images, our training dataset:

$$S = \{x^{(i)}, i = 1, 2, ..., n\} \tag{3}$$

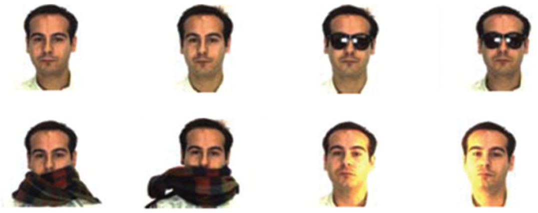

Figure 4: Sample images from the ARFace database.

Each descriptor $x^{(i)}$ consists of d features; each subgroup of p features corresponds to the histogram extracted from one block, and therefore the descriptor was constructed by concatenation of the histograms of d/p image blocks. let x a descriptor of similar characteristics, extracted from a test image, and suppose we want to predict if each of its blocks is occluded or not. First, we must compute the $n \times d/p$ matrix of local distances L. For each feature subgroup $j = 1, 2, 3, \cdots, d/p$ of each descriptor $i = 1, 2, \cdots, n$ the local distance is:

$$L_{i,j} = ||(x_{p(j-1)+1}, \cdots, x_{pj}) - (x^{(i)}_{p(j-1)+1}, \cdots, x^{(i)}_{pj})||^2 \tag{4}$$

Then we compute the occlusion mask $M \in \mathbb{R}^{d/p}$, a vector that has a value for each feature subgroup of the descriptor x; a one indicates the absence of occlusion whereas a value of zero means that the corresponding block is predicted to be occluded:

$$M_j = thr(\ min(\ col_j(L)\ )\ ) \qquad thr(x) = \begin{cases} 1 \ \ si \ \ x < threshold \\ 0 \ \ si \ \ x > threshold \end{cases} \tag{5}$$

The only hyper-parameter of the proposed method is the threshold value, it can be determined on a validation dataset. There is no need for occluded images in the validation dataset. The threshold must be fixed to the lowest value that, when used on this validation dataset, correctly predicts the absence of occlusion in every block. Therefore the proposed method is independent of the nature of the occlusion and does not require images with occlusion to be trained.

4. Experimental results

The goal of this section is to present some preliminary results on the effectiveness of the proposed facial occlusion detection method. The database used in our experiments is the *ARFace database* (Martinez, 1998). This database of facial images contains about 4.000 color images corresponding to 126 different individuals (70 men and 56 women). The images show frontal views of the faces of the individuals with different illumination conditions, facial expressions and occlusion conditions (e.g. with sunglasses and scarfs).

A subset of the images of the ARFace database was selected, several sets were collected to train and test the proposed method. The sets of images that we used are the following:

- Training set: This set consists of one image per individual, with neutral expression and even illumination conditions. The images correspond to the fist session.
- Validation set: This set consists of almost one image per individual, with neutral expression and even illumination conditions. The images correspond to the second session. This set was used to estimate the appropriate values for the hyperparameters of the proposed occlusion detection method.
- Scarf test set: This set consists of almost two images per individual, with neutral expression, even illumination conditions and partial occlusion due to wearing a scarf. The images correspond to the first and second session. This set was used to evaluate the occlusion detection accuracy of the proposed method. In addition, we manually labelled each block of the images as occluded or not occluded.
- Glasses test set: This set consists of almost two images per individual, with neutral expression, even illumination conditions and partial occlusion due to wearing a sunglasses. The images correspond to the first and second session. This set was used to evaluate the occlusion detection accuracy of the proposed method. In addition, we manually labelled each block of the images as occluded or not occluded.

The experiments were conducted as follows: first we trained the proposed method on the training set, and then evaluated a number of threshold values on the validations set. As mentioned before the validation set contains only not occluded images. As the threshold decreases the algorithm begins to erroneously classify not occluded blocks as occluded blocks, we fix the threshold to the lowest value that does produce misclassification of occlusion (40 in this case). This hyper-parameter selection procedure is shown in figure 5.

After that, the occlusion detection model was parametrized as described above and trained on the training set. Then we evaluated the occlusion detection accuracy for both the Scarf and Glasses test sets, obtaining accuracies of 77.8% and 80.2% respectively. In addition, we evaluated the model for a range of threshold values to provide additional insight into the nature of the algorithm; the results are shown in figure 6.

5. Discussion and future work

In this paper, a novel face occlusion detection method has been proposed. The training procedure of this method is independent from the type of occlusion that the system will have to face on test stage, and so it is not necessary to provide occluded images for the training phase. The proposed technique was evaluated on the well known ARFace database, and we found an appropriate threshold value by performing exhaustive search over a range of values.

This method could be applied to automatic face recognition and face verification systems, allowing them to detect the facial regions that are occluded and so discard this corrupted information before the recognition process takes place.

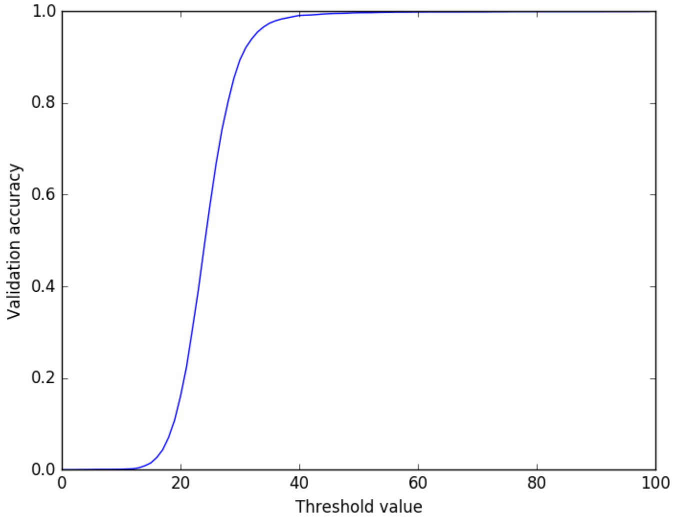

Figure 5: Validation occlusion detection accuracy for different threshold values.

6. References

Ekenel, H. K., 2009. *A robust face recognition algorithm for real-world applications*. Ph.D. thesis, Karlsruhe, Univ., Diss., 2009.

Hughes, G. P., 1968. On the mean accuracy of statistical pattern recognizers. *Information Theory, IEEE Transactions on*, 14(1):55–63.

Jia, H. and Martinez, A. M., 2008. Face recognition with occlusions in the training and testing sets. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–6. IEEE.

Kazemi, V. and Sullivan, J., 2014. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874.

King, D. E., 2009. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research*, 10:1755–1758.

Liu, J., Deng, Y., and Huang, C., 2015. Targeting ultimate accuracy: Face recognition via deep embedding. *arXiv preprint arXiv:1506.07310*.

Martinez, A. M., 1998. The AR face database. *CVC Technical Report*, 24.

Martínez, A. M., 2002. Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(6):748–763.

Min, R., Hadid, A., and Dugelay, J.-L., 2011. Improving the recognition of faces occluded by facial accessories. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference*

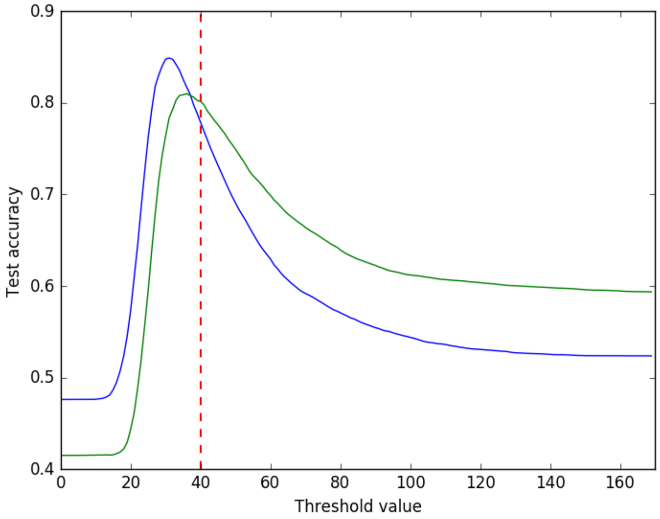

Figure 6: Occlusion detection accuracy on Glasses test set (green) and Scarf test set (blue) for different threshold values and the selected threshold value (dashed red).

on, pages 442–447. IEEE.

Murphy, K. P., 2012. *Machine learning: a probabilistic perspective*. MIT press.

Ojala, T., Pietikäinen, M., and Harwood, D., 1996. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1):51–59.

Ojala, T., Pietikäinen, M., and Mäenpää, T., 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971–987.

Tan, X., Chen, S., Zhou, Z.-H., and Zhang, F., 2005. Recognizing partially occluded, expression variant faces from single training image per person with SOM and soft k-NN ensemble. *Neural Networks, IEEE Transactions on*, 16(4):875–886.