Age Estimation Using Local Binary Pattern Kernel Density Estimate

Juha Ylioinas, Abdenour Hadid, Xiaopeng Hong, and Matti Pietikäinen

Center for Machine Vision Research, P.O. Box 4500,
FI-90014 University of Oulu, Finland

Abstract. We propose a novel kernel method for constructing local binary pattern statistics for facial representation in human age estimation. For age estimation, we make use of the *de facto* support vector regression technique. The main contributions of our work include (i) evaluation of a pose correction method based on simple image flipping and (ii) a comparison of two local binary pattern based facial representations, namely a spatially enhanced histogram and a novel kernel density estimate. Our single- and cross-database experiments indicate that the kernel density estimate based representation yields better estimation accuracy than the corresponding histogram one, which we regard as a very interesting finding. In overall, the constructed age estimation system provides comparable performance against the state-of-the-art methods. We are using a well-defined evaluation protocol allowing a fair comparison of our results.

1 Introduction

Determining the exact age of a person in a given image is a challenging task even for a human observer. For as long as possible, the assignment is done based on the overall hints available. Clearly, there are many factors affecting the final judgement including clothes, posture, and so on. What if the judgement must be based exclusively on target's face which is the case in the branch of face recognition? Naturally, the problem turns out to be even more troublesome.

Automatic estimation of human age based on facial images is an understudied problem. The lack of studies can be explained by many factors including the shortage of ground-truth age information in many existing face databases. As a result, investigation has mainly been focused on solving two to four class problems, where data has been roughly divided into groups containing child, young, adult and old human targets. Until recently, efforts have given birth to two well-known databases named as `FG-NET` [1] and `Images of Groups` [2]. The announcement of these two databases has challenged the research community to investigate the very complex recognition problem of human aging.

Automatic age classification or estimation aims to assign a label to a face regarding the exact age or the age group it belongs. The latter is the case of many early studies in age estimation, but the recent efforts in collecting new databases has seen the birth of interesting corpuses providing the labels of exact ages of the target's which has further driven the focus on more exact age estimation.

A. Petrosino (Ed.): ICIAP 2013, Part I, LNCS 8156, pp. 141–150, 2013.
© Springer-Verlag Berlin Heidelberg 2013

As one of the recent dimensions of facial image analysis, automatic age estimation is useful in many applications such as more affective Human-Computer Interaction, video surveillance, forensics, audience measurement and reporting, and in age invariant face recognition [3].

Facial image based age estimation is challenging because of the appearance of a particular face varies due to changes in pose, expressions, illumination, and other factors such as make-up, occlusions (like eye glasses), image degradations caused by blur and noise, and so on. In addition to these there are variations that are due to, for example, living environment, lifestyle, and genes. Because of the diverse nature of facial aging process it is extremely difficult to find a model for this process.

Existing solutions for age estimation from facial images fundamentally differ in (i) the face representation and (ii) the classification scheme. Many face image representation methods have been studied such as anthropometric models, active appearance models (AAM), aging pattern subspace, and age manifold. An extensive review of age representation methods can be found in [3]. Regarding age classification schemes, the existing methods are based on either pure classification or regression analysis. Perhaps, among the pioneering studies on age classification are those proposed by Kwon and Vitoria Lobo [4], Lanitis et al. [5], and Guo et al. [6]. More recent methods include the ones proposed by Guo et al. [6] and Ruiz et al. [7] which treat age recognition as a regression problem.

The significance of face alignment is crucial in face recognition [8]. Recently, Vu and Caplier [9] noticed that by horizontally flipping the facial image the unpleasant effects due to pose variation can be mitigated in face matching. Considering the real-life nature simulated by the latest face databases and the inherent real-life-like set-up in face databases collected from the internet, the role of face alignment will remain important in face analysis.

In this paper, we propose a novel method for facial representation tackling the problem of human age estimation. Our proposal is an alternative to histograms and it is based on a kernel method that we use for constructing local binary pattern statistics for facial representation. To the best of our knowledge, we are the first ones to use the proposed kernel method for representing faces in age estimation. The outline of the paper is as follows. We first describe the modules of our facial age estimation pipeline. Then, in the experimental section we provide single- and cross-database evaluations proving the stated efficiency of the proposed kernel method against basic histograms. Finally, we provide some discussion about the advantages of our proposal and make the concluding remarks.

2 Our Age Estimation Pipeline

We built a system containing modules for face alignment and facial representation generated by statistics of local features. Our face alignment consists of facial shape and pose normalizations by a similarity transformation and a simple image flipping, respectively. For facial representation, we make use of an

established spatially enhanced method based on local binary pattern (LBP) distributions. We compare two methods to estimate LBP distributions, namely the histogram method and kernel density estimation. Finally, we train a support vector regressor for age estimation.

We geometrically normalize faces based on both eyes and corners of the mouth by a similarity transformation. The motivation for using the similarity arises from the assumption that in the most usual cases the image to be aligned contains a subject directly facing the camera. Obviously, in that kind of case the input face may need some rotation, scaling, and perhaps, only a bit of translation. The face pose is subsequently corrected by flipping the image so that the facial normal is always directed to the left or right side relative to the view of a camera. From the age estimation point of view, the rigid similarity transformation and the image flipping operation are both pleasing as they retain the original shape and texture to large extent.

In our feature extraction and representation module we are using a local binary pattern variant called completed local binary pattern (CLBP) [10]. Compared to the conventional definition of LBP, the method provides two measures for local texture description, one for binary patterns and one for measuring the contrast of them. The CLBP method is explained in more detail in the next section. For the facial representation, we make use of a widely applied spatially enhanced method based on local feature distributions proposed by Ahonen et al. [12]. We compare two methods to estimate LBP distributions, namely the histogram method and the normal kernel method proposed by Aitchison and Aitken [13].

For estimating the age of a target person in a given image, we used the extracted facial representations as inputs to a support vector regressor (SVR) with a non-linear radial basis function kernel. The parameters of the SVR were determined using a grid search. We used the publicly available LIBSVM library [1].

3 Facial Representation by LBP Statistics

The local binary pattern (LBP) operator is a simple yet powerful gray-scale invariant texture primitive [14]. The original form of the operator works in a 3×3 neighborhood, using the center value as a threshold to label each pixel and considering the result as a binary number.

Later on, a more generic form of the operator was proposed using circular sampling and bilinear interpolation, which allowed to use any size of neighborhoods. Another extension addressed the importance of different binary patterns. This so called uniform patterns ($u2$) was inspired by the fact that some binary patterns occur more frequently than others. Uniform patterns are those that contain at most two bitwise transitions from 0 to 1 or vice versa. In general, the success of the method in image description can be seen in many variants and extensions proposed in the literature.

[1] `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

One of the most powerful extensions of LBP is so called completed modeling of local binary pattern (CLBP) [10]. In that local neighborhood is decomposed into two complementary components, the difference signs and the difference magnitudes. The sign component is coded using the conventional LBP operator defined as

$$\text{CLBP\_S}_{P,R} = \sum_{p=0}^{P-1} t(g_p - g_c)2^p, \tag{1}$$

where g_c corresponds to the gray value of the center pixel (x_c, y_c), g_p refers to gray values of P equally spaced pixels on a circle of radius R, and t defines a thresholding function with $t(x) = 1$ if $x \geq 0$ and $t(x) = 0$ otherwise. The magitude component ($\text{CLBP\_M}_{P,R}$) is coded replacing the threshold function in Eq.1 to $t(m_p, c)$, where m_p is is the magnitude of local pixel difference and c is a predetermined threshold value usually set as the mean value of local pixel differences in the whole image. As the magnitude operator encodes the difference in local pixel intensities, it gives a measure of contrast. The key idea of CLBP is to gain more comprehensive image representation by combining these two complementary descriptions.

After turning the input into two separate sign and magnitude labeled images, a histogram can be built by

$$H(i) = \sum_{x,y} \delta_{l,i}, \quad i = 0 \ldots 2^P - 1, \tag{2}$$

where l is a labeled pixel at a position (x, y), 2^P is the number of different labels produced by the operator, and $\delta_{l,i}$ is the Kronecker delta function. Broadly speaking, a histogram is an estimate of the probability distribution. In the context of LBPs, it contains information about the distribution of the local micropatterns, such as edges, spots and flat areas, over the whole image [12].

In addition to a histogram, there is another widely used estimator for probability distributions, namely a kernel density estimator. In our case, however, common estimators in a continuous domain such as the Parzen-Rosenblatt window method is not applicable because LBPs essentially are variables in a multidimensional binary space. Fortunately, there is a kernel method, originally proposed by Aitchison and Aitken [13], that is suitable for estimating the probability distribution of LBP-like random variables. The kernel is given by

$$K_h(l|l') = h^{P-d(l,l')}(1-h)^{d(l,l')}, \tag{3}$$

where l and l' are both P-dimensional binary variables, $d(l, l')$ is the Hamming distance between them, and $h \in [\frac{1}{2}, 1)$ is a bandwidth parameter. Finally, using the given kernel, instead of a histogram, one is then able to estimate the LBP probability distribution by

$$\hat{f}_h(i) = \sum_{x,y} K_h(l|i), \quad i = 0 \ldots 2^P - 1. \tag{4}$$

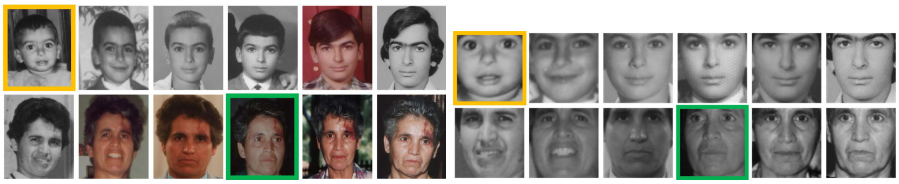
Both the histogram H and the kernel density estimate $\hat{f}_h$ can be further normalized so that they sum up to one. Using the kernel function $K_h(l|l')$ one is able to distribute the same probability mass among several bins instead of putting a probability mass equal to one to a single bin, like in a histogram. The determining factor is the Hamming distance between the given label l and the possible entry i in the statistic, so that the smaller the Hamming distance the larger the probability mass given to the bin. Findings about the possible benefits using the kernel density estimate instead of a histogram are discussed in the experiments.

To retain spatial information, both the histogram and the kernel density estimate can be used to form a spatially enhanced statistic of the whole face as Ahonen et al. described in [12].

4 Experimental Results

To gain insight into the effectiveness of our age estimation system, we conducted experiments following the guidelines of the BeFIT (`http://fipa.cs.kit.edu/befit/`) standards for benchmarking age estimation methods. In addition, we performed a cross-database evaluation where two distinct face databases are used separately as a training and testing sets.

Setup. For the single-database benchmarks, we considered the `FG-NET` database which contains 1,002 uncontrolled images from 82 subjects. There are 6-18 samples per subject with ages from 0 to 69. The database also provides 68 landmarks on each face image. Fig. 1 shows some exemplars from the database, which further highlight the typical conditions of the images containing varying illumination, expression, and pose.



(a) (b)

Fig. 1. (a) Original `FG-NET` images and (b) corresponding geometrically normalized and pose corrected faces samples. Framed samples illustrate pose correction by image flipping.

We use Leave-One-Person-Out (LOPO) for testing our proposals in the `FG-NET` database. This means that the images of one person are used as the test set and those of the others are used as the training set. After going through all 82 folds each subject has been used once as a the test set. The final result is then an average of the results of each fold.

As described earlier, the step before representing the faces is to first align them to reduce the effect of scale, rotation, translation variations. Hence, to get rid of the most of the variation, clearly visible in the images above, all face images are geometrically normalized by a similarity transformation with respect to both eyes and corners of the mouth using their coordinates provided by the database. We use a 76×76 pixels size of model to which all face images are fitted using the similarity calculated by point correspondences between the input and the model points. Before feature extraction, we further normalize all faces with respect to a facial normal by an image flipping operation. As visible in the faces above, they contain also severe pose variation which can be alleviated by forcing the pose to one of either right or left sides by flipping the image. The pose correction was performed by subjective conclusions of the target's facial normal based on, for example, if the right cheek is entirely visible (with respect to the camera), whereas the left shows less, it can be assumed that the facial normal is directed more to the left.

Once normalized, local features are extracted from uniformly distributed patches across the face. The face image is first divided into a set of L overlapping patches of a size 13×13 pixels, each patch overlapping its vertical and horizontal neighbors by 4 units. With a face image of size 76×76, this results in a total of 64 patches. The completed local binary pattern (CLBP) sign and magnitude operators are then used for feature extraction. CLBP has shown to be very powerful means for texture description, it has already been shown to suit well for facial representation in age group classification [11]. Our facial representation is based on local statistics of local features so we compared two sign and magnitude based statistics namely a histogram and a kernel density estimate using the normal kernel proposed in [13]. As the normal kernel requires to find a value for the smoothing bandwidth parameter h we were enforced to perform parameter tunings for both sign and magnitude component statistics.

After constructing the statistics for each patch we applied feature selection to reduce the patch-specific feature dimensions. Thus, for the sign component we used the *u2*-mapping and for the magnitude the *ri*-mapping. For the sign we considered only uniform patterns excluding the final 59th bin that is for non-uniform patterns. Further, as we believe that contrast is a rotation invariant property, it is well-founded to use rotation invariant magnitude component. For both measures, we operated on an $(8, 2)$-neighborhood. Using these settings the sign component yields a 58-dimensional, whereas the magnitude only a 36-dimensional feature vector. Finally, after concatenation of the sign and magnitude components we have 94-dimensional feature vectors for each of L patches. Thus, in our case, the final feature vector size is $L \times 94$, in the case of concatenated CLBP_S_M histogram $64 \times 94 = 6016$.

For estimating the age of a person in a given test image, we used the extracted facial representations as inputs to an SVR using a non-linear radial basis function kernel. The parameters of the SVR were determined using a grid search. According to the BeFIT standards, the performance of age estimation was then measured by the mean absolute error (MAE). The MAE is a mean of the absolute errors

between the estimates and the true ages, MAE $= \sum_{k=1}^{N} |\hat{a}_k - a_k| / N$, where $\hat{a}_k$ and a_k are the estimate and the true age of the sample image k, and the N is the total number of samples.

Results. In our experiments we have two test variables namely facial representation and the face pose correction. The conditions of facial representation vary between six different methods, whereas for pose correction we compare the effect of using the original faces and manually pose corrected faces. If the pose was corrected then it was done both during training and testing. For each setting, we compare the following representations: spatially enhanced conventional LBP sign and magnitude histograms (H + S and H + M), their kernel density versions (K + S and K + M), combined spatially enhanced sign and magnitude histograms (H + S_M), and its kernel density version (K + S_M).

At first we went through finding an optimal value of the smoothing parameter h, for both sign and magnitude kernel density estimates separately, in both original and pose corrected settings. We plot the MAE measure against the h in Fig. 2. The h that gave the highest MAE for the representation in each setting was then selected for further evaluation.

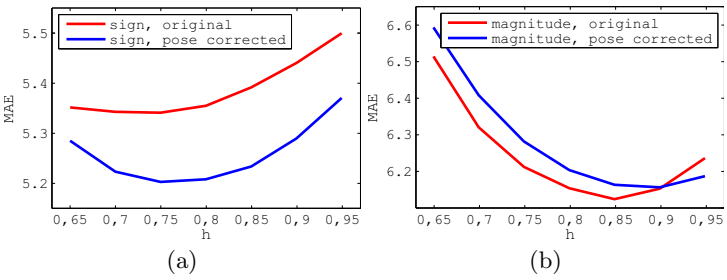


Fig. 2. The effect of h on MAE. (a) for sign and (b) for magnitude based representations in original and pose corrected settings.

After finding the optimal h, we continued by evaluating sign and magnitude, and their combination using face samples without pose correction. The results, shown in Table 1, indicate that the combined representation, i.e. using both sign and magnitude components, clearly outperforms the separated versions. Evidently, as hinted in [10], the sign component is more discriminative compared to the magnitude. The phenomenom is repeated using kernel estimator instead of a histogram, as the combination of sign and magnitude outperforms separated versions, and sign yields better result than magnitude. The most interesting is to highlight the effect of using the kernel density estimator as in that mode the sign component turns out to outperform the combination of sign and magnitude in the histogram mode.

To answer whether pose correction by image flipping would enhance the age estimation performance, we performed further evaluation. Based on the results, age estimation accuracy improves compared to the preceding setting. Evidently,

Table 1. MAE measures for different CLBP-based face representations in two different settings

representation	original faces	pose corrected
$H + \mathrm{S}$	5.61	5.47
$H + \mathrm{M}$	6.36	6.29
$H + \mathrm{S\_M}$	5.39	5.23
$K + \mathrm{S}$	5.34	5.20
$K + \mathrm{M}$	6.12	6.16
$K + \mathrm{S\_M}$	5.20	5.09

the used block-based facial representations are sensitive to the off-plane rotations of faces. Intuitively thinking the idea makes sense as the facial representation methods used here are based on local measures of texture regions. Therefore, in a such demanding task of human age estimation, it is important that faces used in training and testing are well aligned with respect to each other.

What makes the kernel method more powerful than the histogram, can be partly explained by the limited sample size scenario at hand. Given a patch size used here (13×13 pixels) and applying LBP and subsequent histogramming results in extremely sparse descriptors which may cause problems in the learning phase given also the limited number of training instances. By distributing the confidence, brought by one detected LBP label, among many bins instead of a single bin, one is able to produce more robust representation. Secondly, we believe that by distributing the confidence we are able to tackle the problems due to hard quantization, a component of the LBP label calculation, while there are degradations in the image such as noise.

Comparison to State-of-the-Art. We compared our methods against the most relevant works in the literature that report results using the same BeFIT benchmarking standard with `FG-NET`. The results, shown in Table 3, indicate fairly comparable performance.

Table 2. MAE measures for the most relevant works in the literature

method	QM [15]	MLP [15]	RUN [16]	BM [17]	LARR [18]	BGRM [7]	BIFs [6]	**ours**
MAE	6.55	6.98	5.78	5.33	5.07	4.96	4.77	**5.09**

The advantage of our method is computational lightness and its algorithmic simplicity comparing, for example, to the methods in [7] and in [6] where the facial representations are based on the outputs of Gaussian or Gabor-like filter bank responses. Besides, in [7] they also apply LBP operators to encode the resulting Gaussian receptive maps into histograms which hints that the method might gain even higher accuracy by using our proposed kernel density estimate instead of a histogram.

Cross-Database Evaluation. To investigate the generalization ability of the trained age regressors, we also performed a cross-database evaluation. For the testing benchmark, we ended up using the `Images of Groups` (`IoG`) database as it is the only available corpus containing face samples representative of the same age groups and ethnic origins present in `FG-NET`. However, as `IoG` only provides rough age group groundtruth, we explored solely how the age model trained using all `FG-NET` faces manages to categorize the given test set into age groups assigning the regressor output to the specified age range.

`IoG` consists of 28,231 facial images collected from Flickr images, taken in uncontrolled conditions. Each face is labeled with an age category defining seven age groups as follows: 0-2, 3-7, 8-12, 13-19, 20-36, 37-65, and 66+. The grouping roughly corresponds to different life stages [2]. In our evaluation, we considered only faces having interocular distance more than 40 pixels. By that way, we collected a subset of 1495 face images. We further relaxed the experiment by reorganizing the age labels into child, teen, and adult classes setting 0-12, 13-19, and 20+, respectively. The setting yielded the following amount of samples per each age group: 546, 250, and 699. Finally, we went through all of the `IoG` faces performing same normalizations than in the previous experiment. For facial representation, we only considered sign component statistics.

Based on the results, given in Table 3, the pose correction did not seem to provide any meaningful improvement, but comparing the facial representations we found a clear margin in performance between the histogram and the kernel density estimate.

Table 3. Age group classification performance on `IoG` using the sign component based histogram and kernel density estimate representations

representation	original faces (%)	pose corrected (%)
$H + \mathrm{S}$	56.99	56.19
$K + \mathrm{S}$	61.67	61.87

5 Conclusions

In this work we investigated human age estimation problem proposing a kernel method for constructing local binary pattern based statistical face representation. In our experiments, we compared our proposal to the widely used histogram based representation concluding that the kernel one yields much better accuracy. We validate our conclusion using single- and cross-database evaluations.

The motivation of our proposal arises from the limited-sample-size problem inherent to the spatially enhanced facial representation by histograms. While solving such a complex problem as human aging using local features, one is confined to very small image patches from which the aging trace might be possible to be captured. The problem with widely used local binary pattern histograms is then the resulting sparse nature of the representation. In our experiments we show that by using the proposed kernel estimator one is able to tackle this problem.

We also analysed the effect of performing face pose correction by image flipping. Based on the single-database experiments, pose correction provided significant performance improvement. However, in our cross-database experiment, pose correction did not provide any meaningful improvement which may be due to the simplified setting.

References

1. The FG-NET Aging Database, http://www.fgnet.rsunit.com/
2. Gallagher, A.C., Chen, T.: Understanding images of groups of people. In: CVPR 2009, pp. 256–263 (2009)
3. Fu, Y., Guo, G., Huang, T.S.: Age synthesis and estimation via faces: A survey. IEEE TPAMI 32(11), 1955–1976 (2010)
4. Kwon, H.Y., da Vitoria Lobo, N.: Age classification from facial images. In: CVPR 1994, pp. 762–767 (1994)
5. Lanitis, A., Taylor, C.J., Cootes, T.F.: Toward automatic simulation of aging effects on face images. IEEE TPAMI 24(4), 442–455 (2002)
6. Guo, G., Mu, G., Fu, Y., Huang, T.S.: Human age estimation using bio-inspired features. In: CVPR 2009, pp. 112–119 (2009)
7. Ruiz, J., Crowley, J., Lux, A.: "How old are you?": Age estimation with tensors of binary gaussian receptive maps. In: BMVC 2010, pp. 6.1–6.11 (2010)
8. Gross, R., Baker, S., Matthews, I., Kanade, T.: Face recognition across pose and illumination. In: Li, S.Z., Jain, A.K. (eds.) Handbook of face recognition, pp. 197–221. Springer, London (2011)
9. Vu, N.-S., Caplier, A.: Face recognition with patterns of oriented edge magnitudes. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part I. LNCS, vol. 6311, pp. 313–326. Springer, Heidelberg (2010)
10. Guo, Z., Zhang, L., Zhang, D.: A completed modeling of local binary pattern operator for texture classification. IEEE TIP 19(6), 1657–1663 (2010)
11. Ylioinas, J., Hadid, A., Pietikäinen, M.: Age classification in constrained conditions using LBP variants. In: ICPR 2012, pp. 1257–1260 (2012)
12. Ahonen, T., Hadid, A., Pietikäinen, M.: Face description with local binary patterns: Application to face recognition. IEEE TPAMI 28(12), 2037–2041 (2006)
13. Aitchison, J., Aitken, C.: Multivariate binary discrimination by the kernel method. Biometrika 63(3), 413–420 (1976)
14. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE TPAMI 24(7), 971–987 (2002)
15. Lanitis, A., Draganova, C., Christodoulou, C.: Comparing different classifiers for automatic age estimation. IEEE TSMCB 34(1), 621–628 (2004)
16. Yan, S., Wang, H., Tang, X., Huang, T.S.: Learning auto-structured regressor from uncertain nonnegative labels. In: ICCV 2007, pp. 1–8 (2007)
17. Yan, S., Wang, H., Tang, X., Liu, J., Huang, T.S.: Regression from uncertain labels and its applications to soft biometrics. IEEE TIFS 3(4), 698–708 (2008)
18. Guo, G., Yun, F., Dyer, C.R., Huang, T.S.: Image-based human age estimation by manifold learning and locally adjusted robust regression. IEEE TIP 17(7), 1178–1188 (2008)