Pattern Recognition Letters
journal homepage: www.elsevier.com

Evaluation of local descriptors and CNNs for non-adult detection in visual content

Modesto Castrillón-Santana[a,**], Javier Lorenzo-Navarro[a], Carlos M. Travieso-González[b], David Freire-Obregón[a], Jesús B. Alonso-Hernández[b]

[a] *SIANI, Universidad de Las Palmas de Gran Canaria (ULPGC), Spain*
[b] *IDETIC, Universidad de Las Palmas de Gran Canaria (ULPGC), Spain*

ABSTRACT

The recent evolution of storage devices, digital embedded cameras and the Internet have collaterally allowed sexual predators to take advantage of these technological breakthroughs to gather illegal media, which is exhibited uncensored through Peer-to-Peer file sharing networks. In this paper, we are particularly concerned about the increasing availability of Child Abuse Material. Therefore, we have explored alternatives to detect non-adults in visual content. Initially, different age estimations and underage detection techniques are reviewed by analyzing existing datasets. Finally, several local descriptors and convolutional neural networks for underage detection are evaluated. The experimental results obtained for a large dataset that combines collections such as FG-Net, Adience, GenderChildren, The Image of Groups and Boys2Men evidence the complementary information contained in both local descriptors and neural networks, as their fusion boosts the accuracy of non-adult detection to over 93%.

1. Introduction

According to INTERPOL, pornography refers to "*consensual sexual acts distributed (mostly) legally*". For this reason, instead of using the term child pornography, this organization suggests that *Child Abuse Material* (CAM) is the proper term to refer to multimedia material documenting the sexual abuse of a child.

Wolak et al. (2008) state that CAM possession is an unusual sex crime, it is a form of child sexual exploitation that requires no direct interaction with the victim. In fact, the crime is possessing CAM, which is contraband because it shows actual children being sexually abused or exploited. The children and adolescents pictured in this kind of material are generally unidentified victims of exploitation and often abused by the producers of it as Wolak et al. (2011) highlight. Moreover, Thompson (2009) points out that police forces are continuously involved in the task of detecting and removing CAM, identifying victims and offenders and applying the law.

Kawale and Patil (2014) confirm in their studies that this particular crime has increased since current technology facilitates the distribution and sharing of multimedia material. The growing capacity of computers to transmit (and store) images, together with the inexpensive availability of image and video acquisition in almost every device, such as smartphones, tablets or digital cameras may be expanding the risk of child exposure to these sexual offenders. Indeed, Jenkins (2009) states that a primary concern is *Peer-to-Peer* (P2P) file sharing networks, These allow users to bypass centralized servers and download CAM from networks to personal digital devices.

Liberatore et al. (2010) carried out an interesting survey on how law enforcement groups have developed sophisticated systems known by names such as Gridcop, RoundUp, and Ephex, which allow police forces to detect illegal activity by specific *Internet Protocol* (IP) addresses on a variety of P2P networks. However, using these systems is just a first investigation step. The massive presence of multimedia material on the Internet, still requires a strong effort to automate CAM detection. In this sense, the availability of automatic approaches to detect the presence of children in content involving sexual imagery would give extra leverage to the task of CAM detection. Such a system could be divided into two steps. The first one would be devoted to the detection of sexual visual material, making use

**Corresponding author: Tel.: +34-928-458-743; Fax: +34-928-458-711;
e-mail: modesto.castrillon@ulpgc.es (Modesto Castrillón-Santana
)

of techniques similar to parental filters, while the second would be in charge of determining the presence of children.

As a consequence of this filtering task, the detection of CAM would allow police forces to gather enough information to create a CAM-oriented database. In this regard, INTERPOL hosts the *International Child Sexual Exploitation image database* (ICSE DB), which is a powerful tool, that allows specialized investigators to share data with colleagues across the world. This database enables certified users in member countries to access the database directly and in real time, thereby providing immediate responses to queries related to child sexual exploitation investigations.

The first step, pornography detection, has been accomplished making use of skin detection as in Jones and Rehg (1998), or more recently combining different cues such as color, texture and shape also used in Sengamedu et al. (2011). Our proposal is devoted to partially solving the aforementioned second stage, describing an approach to classify an individual as adult or not. The aim is to save time for security forces when filtering material. This can be carried out using previous face detection and performing a classification based mainly on facial information.

Thus, the major contributions of this study are as follows: 1) to tackle the problem of filtering for CAM, 2) the design of a large collection of facial images for adult/non-adult classification, and, 3) the evaluation separately of local descriptors and *Convolutional Neural Networks* (CNNs), and finally, 4) the successful combination of both approaches to boost the system performance.

The paper is organized in four sections. The rest of Section 1 looks at age estimation related work. In Section 2, the local descriptors and the CNNs architecture are described. The collected dataset is fully detailed as well as the classification experiments are reported in Section 3. Finally, conclusions are drawn in Section 4.

1.1. Related work

Age is a soft biometric attribute that has recently received a great deal of attention by the computer vision community. Contrary to *strong* biometric traits which present distinctiveness and permanence to differentiate unequivocally any two individuals, soft biometric traits are referred to as ancillary information not unique, but useful to describe people in meaningful, non-overlapping categories, see Nixon et al. (2015). Furthermore, the facial age estimation task can be divided into three different problems:

- Classification problem as in Gao and Ai (2009); Guo et al. (2008). The age of an individual, considered as a label, is predicted and then grouped into one age group (e.g. child, young, adult, and so on).

- Regression problem as in Lanitis et al. (2004); Suo et al. (2008); Guo and Mu (2011). The goal is to obtain a precise age value.

- A hybrid of the two previous problems as in Lanitis et al. (2004); Takimoto et al. (2008); Guo et al. (2008). Also known as hierarchical age estimation, is a coarse-to-fine method used to find the age label in a pre-classified group of a dataset. As facial aging is perceived differently in different age groups, age estimation in a specific age group provides a more accurate result.

Over the last decade, apparent age estimation proposals are mostly based on facial information, which has been evaluated on datasets such as FG-Net (Panis et al. (2016)), MORPH (Ricanek and Tesafaye (2006)) or Adience (Eidinger et al. (2014)). However, these datasets do not include a wide collection of individuals close to the adult/non-adult classification border, as they have been designed for different purposes such as interaction or commercial applications. In this sense, we may refer to our recent work gathering Boys2Men, see Castrillón-Santana et al. (2016a,b), to our knowledge, the only dataset designed for this particular problem. Satta et al. (2014) described an interesting dataset focused on the gender classification of children, evidencing the presence of different features for such a task and suggesting the need for the integration of local context features to improve accuracy.

Moreover, Choi et al. (2011) considered two types of features used for predicting age from facial images: local and global features. The first type of features are taken from texture (hair and skin), while the second, global features, are extracted from the geometric properties (appearance and shape) of the facial elements. Choi et al. stated that local characteristics are commonly known to better classify people into age groups. On the other hand, these authors also argued that their global counterparts are better not only for estimating age, but also for identifying different individual traits, such as identity, emotions and race.

Another interesting proposal developed by He et al. (2016), reduced the problem to an age prediction problem over time evolution as a time-series sequence estimation analysis. For this purpose, a structure-aware method based on *Slow Feature Analysis* (SFA) was considered. This approach captures the structural information of a human face as the face evolves over time gradually by using SFA. This technique can be useful for child re-identification in different CAM-timelines.

Recent soft biometrics surveys as described y Nixon et al. (2015); Dantcheva et al. (2016) give results for both, age estimation and age group classification. Several techniques require a precise detection of face fiducial points. Studies based on local descriptors are relatively novel, and there are no extensive evaluations covering different alternatives. Moreover, recent great results based on deep learning for different Computer Vision problems, as seen in Krizhevsky et al. (2012), have also attracted the interest of soft biometric researchers to CNNs, see Lecun et al. (1998), in particular for gender and age estimation. We may refer to Levi and Hassner (2015) who present a CNN with three layers and two fully connected layers that obtains accuracies over 50% in the age problem.

Another approach is found in Almeida et al. (2016), where a tool was developed to identify persons under 20 years old, which is based on the Discrete Cosine Transform (DCT). FG-Net and Adience datasets were used to evaluate this approach, achieving up to 71.1% of success using almost 7,000 images as blind evaluation samples.

A Microsoft working group has developed an application

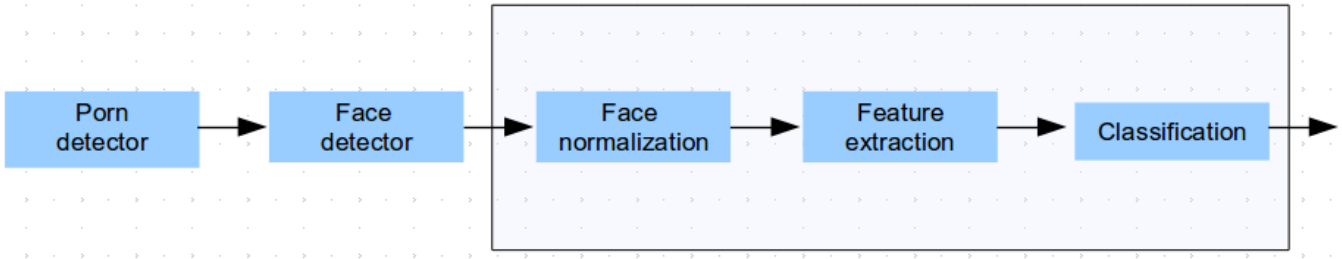

Fig. 1. CAM detection proposal

called *How old do I look?*, see Microsoft (2016). Depending on the image, the application obtains good results in age estimation, but wrong estimations have also been reported. Other websites that estimate age or age ranges are Detector (2016); KAIROS AR (2016). The first one from facial images and KAIROS using video images. Their developers confirm that these applications do not have totally reliable results.

Another well-known site is Face.com, see Facebook (2016), that was developed by an Israeli company that offers facial recognition to other web pages and companies, including Facebook. This system tries to estimate the age of a person by only analyzing a photo. Its application does not give an accurate result, but shows a range of approximate ages: the apparent age, and the estimated maximum. The real system effectiveness is unknown, since the company does not disclose any information about its software.

In this paper, we perform a study covering and comparing local descriptors, *Regions of Interest* (ROIs), CNNs and fusion strategies for the classification of facial images belonging to adults or non-adults.

2. Proposal

An outline of the proposed system is shown in Fig. 1. Our work is not focused on the detection of pornographic content, nor on face detection. In fact, we tackle exclusively the facial analysis modules, see the highlighted part.

In this section, two alternatives are described to tackle the adult/non-adult classification problem. The first one is based on local descriptors, where the possibility of fusing different alternatives and ROIs is evaluated. The second one explores the use of deep CNNs for this particular problem. Instead of estimating the precise age of a given individual, we have adopted a *simpler* binary approach to classify an individual as adult or non-adult.

2.1. Local descriptors

As stated in Section 1.1, facial cues can be defined as a standard reference for human faces used by scientists for facial analysis, e.g. to estimate the age of a person. In this regard, the aging process affects the facial structure and appearance of a person in many ways. The changes that occur are related to craniofacial morphology and face texture.

On the one hand, certain features of craniofacial morphology appear only in people of a certain age and change during the aging process (e.g. the facial skeleton growth). On the other hand, changes in the facial texture are defined as variations in face associated with skin and muscle elasticity, see Geng et al. (2008).

Fig. 2. Boys2Men sample showing the different ROIs evaluated with local descriptors. From left to right: head and shoulders (HS) (64×64 pixels), face (F) (59×65 pixels), periocular (P) (49×19 pixels), and mouth areas (M) (37×31 pixels).

To tackle the problem, we have considered four facial ROIs, see Fig. 2, with the aim of extracting features at different resolutions in order to capture the specific characteristics of each region. The analysis has included several texture based descriptors:

- The *Histogram of Oriented Gradients* (HOG) proposed by Dalal and Triggs (2005). This technique counts occurrences of gradient orientation in localized rectangular areas of an image.
- *Local Binary Patterns* (LBP), and in their uniform form (LBP^{u2}) proposed by Ahonen et al. (2006). These descriptors have proven to be useful for texture classification. They encode each pixel considering the magnitude relation with the surrounding neighbors. As a result, a binary code is obtained. Moreover, the LBP^{u2} variant reduces considerably the number of possible binary codes, taking into account only the most common ones.
- *Local Gradient Patterns* (LGP), proposed by Jun and Kim. (2012), contrary to LBP, consider the pixel gradient instead of its gray value. LGP representation is insensitive to global intensity variations like others such as the aforementioned LBP.
- *The Local Ternary Patterns* (LTP), proposed by Tan and Triggs (2010), are an extension of LBP. Unlike LBP, LTP does not divide the pixels into 0 and 1, but into three values. Instead of using a histogram with a larger number of bins, the ternary pattern is commonly split into two binary codes known as LTP_{low} and LTP_{high} respectively.

- The *Local Salient Patterns* (LSP), proposed by Chai et al. (2013), are also a variation of LBP. In this case, instead of computing each pixel considering neighbors, only the most relevant local comparisons are encoded, with the largest positive or negative contrast magnitude in LBP feature representation. As a result, LSP is expected to be more robust than the conventional LBP approach. Moreover, high order cases which explore more local relationships among multiple pixels can be defined. Those described in the original paper are considered in our analysis: LSP_0, LSP_1, LSP_2, LSP_{01} and LSP_{012}.

- The *Weber Local Descriptor* (WLD), proposed by Chen et al. (2010), has been considered for image texture classification. It is based on the fact that human perception depends on both changes in stimulus (e.g. sound, light, etc.) and the intensity of the stimulus. Specifically, WLD consists of two components: differential excitation and orientation. The differential excitation component is a function of the ratio between the relative intensity differences of a pixel and its neighbors and the pixel intensity. The orientation component is the pixel gradient orientation.

- The *Local Phase Quantization* (LPQ), proposed by Ojansivu and Heikkilâ (2008), has also been considered for image texture classification. It is computed based on quantizing the phase information of the local Fourier transform, and shows a good performance in the case of blurred images.

- The *Intensity based Local Binary Patterns* (NILBP) proposed by Liu et al. (2012). Unlike LBP, NILBP compares regional image medians rather than raw image intensities. The descriptor is computed by comparing image medians over a novel sampling scheme, which can capture both microstructure and macrostructure texture information.

- The *Local Oriented Statistics Information Booster* (LOSIB), proposed by García-Olalla et al. (2014), is also based on LBP, but mean values are computed for the pixels in each cell.

Recent literature on facial analysis together with our own conclusions on gender classification, see Castrillón-Santana et al. (2016); Castrillón-Santana et al. (2016c), leads us to tackle the age classification problem not by just using different descriptors, but also by combining them in order to achieve a more accurate approach. The evaluation of descriptors and ROIs is later compared by means of a standard classification approach based on Support Vector Machines (SVM) with RBF kernel, see Vapnik (1995).

In our studies related to gender classification, we have considered a fusion strategy as being either at feature, matching score, or decision level. *Feature Level* (FL) fusion, where features are concatenated in a single feature vector, will likely keep more information, but increasing the problem of dimensionality, and reducing parallelization possibilities. Moreover, decision level fusion ignores the use of substantial valuable information. These disadvantages suggest the use of *Score Level* (SL) fusion, trying to balance between speed and performance. In this sense, and similar to Heisele et al. (2007), we adopt a two-layered architecture. Each first layer classifier would be trained for a specific descriptor and ROI. Each first stage SVM-based classifier output is a score that indicates the sample's proximity to the border between both classes, i.e. adult/non-adult. SL fusion is applied in the second stage where information from different descriptors and ROIs may be combined. Thus, the second layer fuses those scores feeding them into a single and fast SVM classifier. In brief, SL fusion is selected as it provides a good compromise between speed and performance, avoiding an unnecessary increase in the feature vector dimension. Moreover, first stage classifiers may be launched taking advantage of parallel architectures. In addition, our own experience suggests that this approach improves accuracy while reducing the number of difficult cases, i.e. ambiguities as recently evidenced in Castrillón-Santana et al. (2016).

In short, the local descriptor fusion proposal makes use of a two-layered architecture. Each first layer output score indicates the proximity to the decision boundary between both classes for each classifier. The second stage contains a single classifier that combines these first layer scores.

In Section 3, we compare a single descriptor and a single pattern classification with SL fusion of patterns and descriptors considering a 5-fold evaluation protocol.

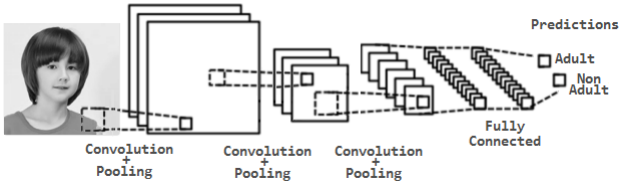

Fig. 3. CNN architecture.

2.2. Convolutional Neural Networks

CNNs have reported excellent results in soft biometric related classification, see Liu et al. (2015); Levi and Hassner (2015), with recent proposals of its combination with hand crafted features particularly for gender classification as in van de Wolfshaar et al. (2015); Mansanet et al. (2016); Castrillón-Santana et al. (2016c).

In this work, we present results considering CNNs architecture proposed by Levi and Hassner (2015) that comprises three convolutional layers and two fully connected layers (see Fig. 3). Convolutional layers follow the Conv/ReLU/Pool scheme. Thus, each convolution preserves the spatial relationship between pixels by learning feature maps in a receptive field. These receptive fields for the second and third layers are fixed at 5x5 and 3x3, and 256 and 384 feature maps, respectively. The receptive field of the first layer was selected by exploring several configurations, as described in the experimental section below.

ReLU is an element wise operation and replaces all negative pixel values in the feature map with zero. The purpose of ReLU is to introduce non-linearity in our CNN, since most of the facial data we would want our CNN to learn would be non-linear. Then, the spatial pooling reduces the dimensionality of each

feature map but retains the most important information. In the present work, the same max pool has been used with a receptive field of 3x3 and a stride of 2 pixels. Finally, the last two layers are fully connected ones with each one a size of 512. In these two layers, the dropout regularization technique described by Srivastava et al. (2014) was introduced to avoid overfitting.

Fig. 4. Normalized Boys2Men HS samples from 12 to 21 years old.

3. Results

In this section, we summarize the experimental setup, first presenting the dataset used, and then defining the global 5-fold cross validation experiment. Finally, the results are given, considering the collection of local descriptors, their fusion and the application of CNN.

3.1. Datasets

Before summarizing the set of experiments, we briefly present the data collection used for our particular problem. With the intention of building a large, challenging and more balanced dataset, we have combined samples from different collections already present in the literature. In particular we have selected:

- Boys2Men by Castrillón-Santana et al. (2016a,b). Given the difficulties to find in the research community a dataset specifically designed for the adult/non-adult classification problem, in our preliminary works, we gathered from the web a reduced dataset, containing about 1,000 web samples, with individuals from 12 until 21 years old. Such dataset provided us an initial benchmark to determine the feasibility of using local descriptors for this problem, focusing on individuals in ages close to 18 years old. Samples of each age group are shown in Fig. 4.

- Adience by Eidinger et al. (2014). According to the authors, this collection covers a far wider range of challenging real-world imaging conditions, comprising changes in appearance, noise, pose, lighting and more. Flickr is the data source, containing around 26,000 samples belonging to more than 2,200 individuals. The age annotations considered the following non-overlapping groups: (0-2), (4-6), (8-12), (15-20), (25-32), (38-43), (48-53) and (60-100). They are certainly not identical to other age datasets, see below those used by Gallagher and Chen (2009), but the first four groups are associated with non-adults. A normalized sample is shown in Fig. 5.

Table 1. Dataset statistics.

Dataset	Adult	Non-adult
Adience	6,763	7,480
Boys2Men	399	579
FGNet	362	638
GenderChildren	-	1,411
The Images of Groups	23,034	5,100
AgeMega	30,558	15,208

- The Images of Groups by Gallagher and Chen (2009). Large dataset containing more than 28,000 annotated faces at different resolutions. The collection was built up once more from Flickr selecting images with more than one individual. Each face is described in terms of gender and age, considering the following age groups: (0-2), (3-7), (8-12), (13-19), (20-36), (37-65) and (66+). For our experimental setup, the first four groups are associated with non-adults. A normalized sample is shown in Fig. 5.

- GenderChildren by Satta et al. (2014). Created to evaluate gender classification in children, which has been proven to need diverse features compared to adults. This collection contains images obtained from web sites, once a face detector, see Viola and Jones (2004), has provided a valid detection. A normalized sample is shown in Fig. 5.

- FG-Net, see Panis et al. (2016). Dataset released with aim at understanding the changes in facial appearance related to age. For this purpose, the collection contains a reduced number of identities, but encloses multiple images per individual, with precise information of his/her age in each one. A normalized sample is shown in Fig. 5.

The final dataset includes faces with positive automatic eye detection, see Castrillón et al. (2011). Below, we refer to the resulting dataset as AgeMega. This data collection contains more than 44,000 samples where the number of adult faces doubles the number of non-adults. See Table 1 for the respective statistics, of the individual datasets and AgeMega.

A light normalization is applied to each face using detected eye positions. This normalization process includes a rotation, scale and translation to obtain images similar to the one depicted in Fig. 6, where the inter-eye distance is 26 pixels. This normalized pattern is referred to as Head and Shoulders (HS), as it contains not just the face but also information of the head silhouette and the upper torso. Black pixels are used to fill empty background, which is not present in the original image. See other examples in first and last samples shown in Fig. 5.

3.2. Single descriptor and pattern

Local descriptors are typically used to represent an image on a histogram. The representation of the facial information making use of a single histogram reduces the original information losing spatial location data greatly. This disadvantage may be solved by the proposal due to Ahonen et al. (2006) that makes use of an image grid, concatenating the respective image cell histograms. Briefly, given a normalized pattern to extract features, a cell grid is defined in terms of the number of horizontal

Fig. 5. Normalized samples of Adience, The Image of Groups, GenderChildren and FG-Net respectively.

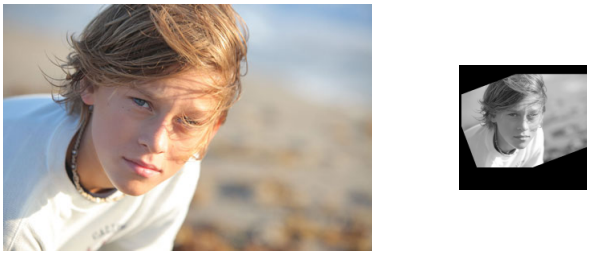

Fig. 6. Original and corresponding normalized image (159×155 **pixels). The normalized image fixes the eye locations to** $(66, 62)$ **and** $(92, 62)$**, with an inter-eye distance of** 26 **pixels.**

and vertical cells, cx and cy respectively. The application of a descriptor to cell i produces a histogram, h_i, where each bin indicates the number of occurrences of a particular descriptor code. The image representation is composed of the concatenation of the $cx \times cy$ histograms, defining the image feature vector $\mathbf{x} = \{h_1, h_2, ..., h_{cx \times cy}\}$ for the particular descriptor.

Initially, we explored different grid configurations for Boys2Men within the range $1 \leq cx \leq 5$ y $1 \leq cy \leq 5$ as reported in Castrillón-Santana et al. (2016a), see an illustration in Fig. 7. The grid exploration results, summarized in Table 2, show the best accuracies obtained for each descriptor and pattern when considering a 5-fold cross-validation experiment for Boys2Men. The total number of descriptors, 15, and grid configurations, 25, were evaluated for four ROIs: head and shoulders (HS), face (F), periocular (P), and mouth areas (M), see Fig. 2. Best single descriptor accuracies reached about 69% for all ROIs except for M. Accuracy differences are, however, not large among the ROIs, but there are certainly differences related to each descriptor, and therefore those performing best for each ROI. For F best values are given by LOSIB, LSP_0 and LSP_{01}; for HS LBP^{u2}; and for P lead HOG, LGP, WLD and LSP_2.

These best performing descriptor configurations are also applied to AgeMega, in a 5-folds cross-validation experiment that comprises about 45,000 samples. The results are summarized in Table 3. A first observation suggests that higher accuracies are achieved compared to Boys2Men. Also, leading descriptors do not generalize. Certainly, AgeMega is a remarkably large dataset that covers a wider age range. Indeed, the sample ages are sparse and are not distributed close to the adult age decision boundary like in Boys2Men. Observing the best accuracies for each ROI, F provides the highest value. In fact, computing

Table 2. 5-folds cross validation results obtained for Boys2Men considering the four ROIs, different grid configurations and local descriptors. For each ROI and descriptor, only the grid setup reaching the highest accuracy (%) is shown.t Extracted from Castrillón-Santana et al. (2016a)

Descriptor	F		HS	
	# (grid)	Acc.	# (grid)	Acc.
HOG	225 (5×5)	68.2	180 (4×5)	64.7
LBP^{u2}	531 (3×3)	68.7	531 (3×3)	**69.7**
LBP	256 (1×1)	61.6	768 (3×1)	68.2
LGP	6400 (5×5)	67.6	5120 (4×5)	64.1
LPQ	768 (1×3)	68.2	768 (3×1)	65.2
WLD	1280 (1×5)	66.7	768 (1×3)	63.6
LOSIB	160 (4×5)	**69.2**	128 (4×4)	59.6
NILBP	885 (3×5)	67.2	885 (3×5)	66.7
LSP_0	684 (4×3)	**69.2**	684 (3×4)	67.7
LSP_1	684 (4×3)	68.2	855 (3×5)	64.7
LSP_2	855 (3×5)	65.2	684 (3×4)	65.2
LSP_{01}	570 (1×5)	**69.2**	1026 (3×3)	65.7
LSP_{012}	855 (3×5)	66.2	518 (3×1)	65.2
LTP_{high}	512 (2×1)	63.1	768 (1×3)	63.6
LTP_{low}	512 (1×2)	63.6	768 (3×1)	63.2

	P		M	
	# (grid)	Acc.	# (grid)	Acc.
HOG	180 (5×4)	**69.7**	225 (5×5)	62.6
LBP^{u2}	885 (5×3)	67.7	590 (5×2)	66.7
LBP	1536 (3×2)	64.1	768 (3×1)	65.6
LGP	5120 (5×4)	**69.2**	3072 (4×3)	63.1
LPQ	512 (2×1)	66.2	512 (1×2)	65.7
WLD	1536 (2×3)	**69.7**	1536 (3×2)	68.2
LOSIB	120 (3×5)	67.7	160 (4×5)	61.1
NILBP	708 (4×3)	67.7	590 (2×5)	68.7
LSP_0	513 (3×3)	67.2	228 (2×2)	67.2
LSP_1	855 (5×3)	65.7	456 (4×2)	64.7
LSP_2	855 (5×3)	**69.2**	456 (4×2)	63.7
LSP_{01}	456 (2×2)	68.7	228 (2×1)	67.7
LSP_{012}	1026 (2×3)	66.2	342 (1×2)	65.2
LTP_{high}	3072 (3×4)	68.7	1024 (1×4)	64.6
LTP_{low}	1024 (1×4)	66.2	1280 (1×5)	63.6

Table 3. 5-folds cross validation results obtained for AgeMega dataset considering the four ROIs, different grid configurations and local descriptors. In addition to the accuracy (%), we provide in brackets the average sample processing time in milliseconds using a Matlab implementation in a quad core i7 3.40 Ghz with 16GB RAM.

Desc.	F	HS	P	M
HOG	84.79 (8)	79.10 (9)	82.32 (8)	**82.29** (9)
LBP^{u2}	83.61 (31)	80.84 (34)	81.18 (57)	80.47 (55)
LBP	74.61 (20)	75.16 (66)	79.13 (78)	77.90 (52)
LGP	83.21 (191)	78.68 (178)	79.07 (146)	77.75 (105)
LPQ	**85.16** (45)	79.26 (52)	82.11 (34)	80.64 (46)
WLD	82.70 (316)	76.22 (322)	82.50 (111)	80.56 (146)
LOSIB	80.99 (7)	75.92 (7)	78.34 (6)	77.76 (7)
NILBP	84.62 (54)	81.83 (63)	80.80 (44)	79.84 (46)
LSP_0	83.20 (111)	80.02 (129)	79.13 (44)	77.58 (39)
LSP_1	83.62 (100)	81.15 (119)	81.50 (54)	80.23 (48)
LSP_2	82.75 (113)	78.97 (122)	79.29 (52)	77.15 (47)
LSP_{01}	82.65 (106)	**82.37** (144)	80.37 (42)	78.46 (38)
LSP_{012}	83.93 (121)	76.73 (119)	82.88 (74)	79.38 (44)
LTP_{high}	79.50 (37)	77.88 (59)	**82.08** (125)	79.16 (71)
LTP_{low}	80.88 (35)	76.58 (61)	78.60 (48)	79.34 (87)

HOG on the facial pattern, F_{HOG} reports an accuracy over 84%. This is a considerably larger value compared to those achieved for Boys2Men, while covering in the experiment a significantly larger population (45,000 vs. 1,000) and acquisition variability. The differences across descriptors for each pattern are within the range 0-5%. These results suggest that some of them may be extracting quite similar information, but some differences are evident. Each descriptor's behavior is affected by the pattern analyzed. The significant differences in processing time should certainly be considered for the selection of descriptors to be applied.

3.3. Fusion

After obtaining the accuracies for single descriptor and ROI, summarized in the previous subsection, we have evaluated the performance when fusing their scores. SL fusion is adopted due to the reduced feature vector dimension compared to FL fusion, and the additional advantage of offering the possibility of parallel computation. The latter is of particular interest when a real-time scenario is considered. The image processing distributed approach has already proven to be reliable for different purposes, see Cattaneo et al. (2014); Liu et al. (2017).

A first exploration is made fusing descriptors for a single ROI, their respective results are presented in the upper part of Table 4. Observing the high cost of evaluating any possible combination, i.e. 2^{15} per ROI, the maximum number of descriptors to be combined has been limited to three. For each pattern or ROI, we summarize the top-5.

The facial area, F, is the ROI reporting the best rates. Indeed, the SL fusion of three descriptors for F reaches values close to 87%, suggesting it would be interesting to make use of complementary information contained by different descriptors.

Another evaluated possibility is the fusion of descriptors extracted from different ROIs. Certainly, the exploration space is significantly larger. In this sense, we have just explored combining up to two descriptors per ROI. As the increase in accuracy when using more than one has been rather quite insignificant, while requiring higher processing costs, only the top-5 combinations using up to one descriptor per ROI are presented in the lower part of Table 4. The reported results reached an accuracy over 86.87%. This mean value was obtained computing LGP for F, HOG for HS, LPQ for P and NILBP for M. These results are again significantly higher than those achieved for Boys2Men. However, they are practically identical to the rate achieved using three descriptors computed on F. A closer look at their respective *Receiver Operating Characteristic* (ROC) curves, see Fig. 8, suggests again quite similar performances for both leading approaches based on SL fusion of local descriptors. However, the fusion of more than one descriptor per ROI does not provide much improvement in the classification rate, and it is therefore not included.

Table 4. Summary of score fusion level results (%) combining descriptors and/or ROIs for AgeMega.

Pattern	Acc.	Descriptors
F	86.91	HOG + LPQ + NILBP
	86.86	LPQ + NILBP + LSP_0
	86.81	LPQ + NILBP + LSP_1
	86.80	LPQ + NILBP + LSP_{012}
	86.79	HOG + LPQ + LSP_1
HS	83.49	LPQ + NILBP + LSP_{01}
	83.38	NILBP + LSP_1 + LSP_{01}
	83.34	NILBP + LSP_2 + LSP_{01}
	83.33	LPQ + LSP_2 + LSP_{01}
	83.31	LPQ + LSP_1 + LSP_{01}
P	84.25	HOG + LPQ + WLD
	84.20	LPQ + WLD + LSP_{01}
	84.14	LPQ + WLD + LSP_1
	84.07	LBP^{u2} + LSP_{012} + LTP_{high}
	84.06	LPQ + WLD + LSP_2
M	83.50	HOG + LSP_1 + LSP_{012}
	83.49	HOG + NILBP + LSP_1
	83.38	HOG + LSP_1 + LSP_2
	83.37	HOG + LPQ + LSP_1
	83.37	HOG + NILBP + LSP_{012}
F+HS+P+M	86.87	F_{LPQ} + HS_{LPQ} + P_{HOG} + M_{LSP_1}
	86.83	F_{LPQ} + HS_{NILBP} + P_{HOG}+ M_{LSP_1}
	86.81	F_{LPQ} + HS_{NILBP} + P_{WLD} + M_{LSP_1}
	86.80	F_{LPQ} + HS_{LPQ} + + P_{WLD} + M_{LSP_1}
	86.79	F_{NILBP} + $HS_{LSP_{01}}$ + P_{HOG} + M_{LSP_1}

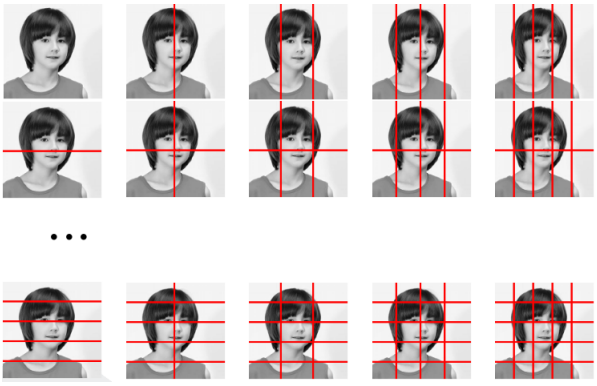

Fig. 7. Grid configurations evaluated.

Table 5. Accuracy (%) per fold and mean (%) using CNN. The average sample processing time using Caffee model running in a NVIDIA GTX 960 with 4GB RAM is 37 msecs.

Mean	fold 1	fold 2	fold 3	fold 4	fold 5
87.95	88.18	88.34	88.43	87.43	87.39

3.4. CNNs

The second adopted focus investigates the use of CNNs. As mentioned in Section 2.2, the configuration of the first convolutional layer was selected using two different sizes for the receptive field: 5×5 and 7×7. Also, we analyzed the performance of two different numbers of features (96 and 128), and three different strides: 5, 3 and 2 pixels. Among the different configurations, the results shown in Table 5 were obtained using 96 features computed with a receptive field of 5×5 and a stride of 2 pixels for the first convolutional layer. Given this setup the output of each level is described as follows:

- Input layer: Three channel images of 232×232 pixels that correspond to the HS ROI.
- First Conv/RelU/Pool layer: 96 features maps of 57×57 pixels.
- Second Conv/RelU/Pool layer: 256 features maps of 28×28 pixels.
- Second Conv/RelU/Pool layer: 384 features maps of 14×14 pixels.
- First fully connected layer: 512 features.
- Second fully connected layer: 512 features.
- Output layer: 2 classes.

The resulting rates, see Table 5, indicate an homogeneous accuracy for the five folds, with a limited overall increase reaching almost 88%. This slight improvement, compared to SL fusion of local descriptors, is also confirmed by observing the corresponding ROC curve, see Fig. 8.

3.5. Discussion

The previous subsections have shown that the facial area and its local context provide useful information for the problem of adult/non-adult classification.

A first focus computed local descriptors, extracting features from different ROIs, reaching promising adult/non-adult classification rates.

A second focus adopted the use of CNNs for this purpose. Feeding in HS images, classification rates achieved reported a slight improvement reducing the system designer workload devoted to local descriptor selection and setup, requiring, however, specific hardware.

As a final test, we were interested in evaluating the fusion of both focuses, similar to recently published results that combine local descriptors and CNNs, see van de Wolfshaar et al. (2015); Mansanet et al. (2016); Castrillón-Santana et al. (2016c). To do this, we have included CNN outputs, in terms of likelihood

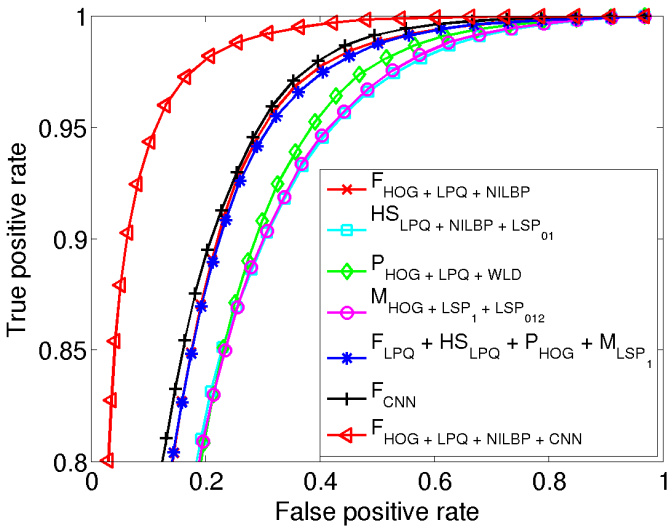

Fig. 8. ROCs.

Table 6. Accuracy (%) per fold and mean (%) using SL fusion of local descriptor scores and CNN outputs.

Mean	fold 1	fold 2	fold 3	fold 4	fold 5
93,06	93.20	92.96	93.83	92.74	92.58

scores, as an additional input in the second level SL fusion architecture, together with those descriptors that reported the best classification rates for F, i.e. F_{HOG}, F_{LPQ}, and F_{NILBP}. The reported accuracies per fold and mean, see Table 6, comfortably beat any previous approach, reaching a mean value of over 93%. This overall rate improvement is also clearly visible in its ROC curve, see Fig. 8.

The proposed local descriptors and CNN outputs SL fusion approach certainly involves higher computational cost, however, a system designer may decide which elements to integrate according to the application needs, as hardware requirements and average sample processing times are known. In any case, the CNN-based reported error was over 12%, hough it was reduced to less than 7% after combining with local descriptors. A remarkable error reduction.

The resulting confusion matrix, see Fig. 9, suggests quite a balanced distribution of errors, but certainly the percentage of uncorrected classified samples is slightly larger for non-adults taking into account the total number of class samples. It can be observed that the number of non-adult faces in the training sample is halved for this class. This unbalanced nature of the dataset can bias performance measures such as accuracy, F-measure or ROC. In this sense, we also provide the set of measures based

Table 7. Performance measures for unbalanced datasets.

Measure	Value
Younden's Index (γ)	0.84
Positive Likelihoood (ρ_+)	8.55
Negative Likelihoood (ρ_-)	0.05
Discriminative Power (DP)	2.80

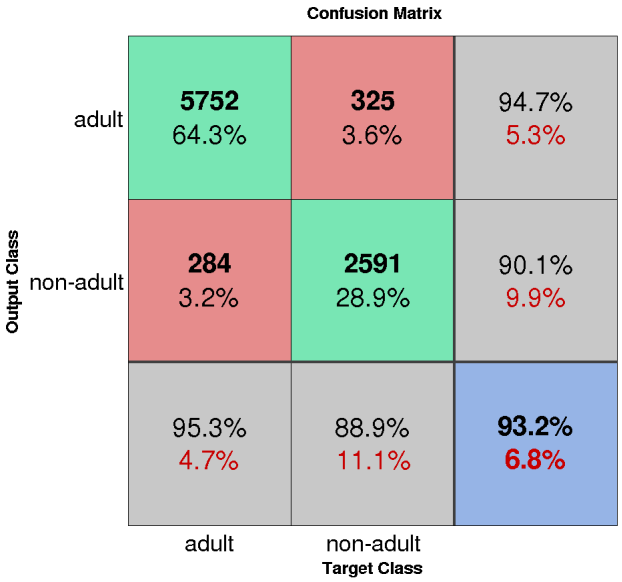

Fig. 9. Confusion matrix obtained for fold 1 using SL fusion of local descriptors and CNN outputs.

on specificity and sensitivity for unbalanced problems proposed in (Sokolova et al., 2006). Table 7 presents Younden's index, likelihood and Discriminative Power of the proposed adult/non-adult classifier. The high value for Younden's index means that the proposal is good at avoiding failures. This fact together with the value for Discriminative Power, close to 3, confirms that the proposed method exhibits good performance. Finally, both high positive and low negative likelihood values indicate no unbalanced performance on positive or negative samples.

A final comment on the criminal scenario considered, is the possibility that the given images might be manipulated, affecting the system's performance. In our proposal, we have not addressed this issue, but image forgery detection should be taken into account in a real-life scenario Farid (2009); Qazi et al. (2013).

4. Conclusions

This study has tackled the adult/non-adult classification from facial cues. The aim is to assist police forces in the automatic detection of CAM, in order to speed up the process of CAM localization.

To do this, we have evaluated the existing age literature datasets, arguing that they are not well suited for our purpose. In this sense, we have combined a collection of datasets in order to create a larger dataset of unrestricted images.

After defining the experimental setup, the evaluation has first covered a wide range of local descriptors and ROIs to later evaluate their SL fusion. The best rates achieved an accuracy close to 87%. A second focus based on CNNs was also analyzed, exploring a range of configurations reaching accuracy values over 88%. Thus, both approaches reported a classification error of about 12-13%.

This error is significantly reduced to less than 7% when local descriptor scores and CNNs outputs are all combined considering SL fusion. The reported accuracy, over 93%, confirms our previous evidence obtained for face based gender classification, where local descriptors and CNNs provide complementary information, which may certainly be used to boost classification performance in automatic facial analysis.

Acknowledgments

This work has been partially supported by the Spanish Ministry of Economy and Competitiveness TIN2015-64395-R (MINECO/FEDER).

References

Ahonen, T., Hadid, A., Pietikäinen, M., 2006. Face description with local binary patterns: Application to face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 28, 2037–204.

Almeida, V., Travieso, C., Alonso, J., Dutta, M., Singh, A., 2016. Automatic age detection based on facial images, in: Proceedings of The 2nd IEEE International Conference on Communication Control and Intelligent System (CCIS), p. Accepted and in press.

Castrillón, M., Déniz, O., Hernández, D., Lorenzo, J., 2011. A comparison of face and facial feature detectors based on the violajones general object detection framework. Machine Vision and Applications 22, 481–494.

Castrillón-Santana, M., De Marsico, M., Nappi, M., Riccio, D., 2016. MEG: Texture operators for Multi-Expert Gender classification. Computer Vision and Image Understanding (in press). URL: http://dx.doi.org/10.1016/j.cviu.2016.09.004, doi:10.1016/j.cviu.2016.09.004.

Castrillón-Santana, M., Lorenzo-Navarro, J., Freire-Obregón, D., 2016a. Análisis de descriptores locales para la detección de menores en imágenes, in: XXXI Simposium Nacional de la Unión Científica Internacional de Radio.

Castrillón-Santana, M., Lorenzo-Navarro, J., Freire-Obregón, D., 2016b. Boys2men, an age estimation dataset with applications to detect enfants in pornography content, in: First International Workshop on Biometrics and Image Forensics.

Castrillón-Santana, M., Lorenzo-Navarro, J., Ramón-Balmaseda, E., 2016c. Descriptors and regions of interest fusion for in- and cross-database gender classification in the wild. Image and Vision Computing (in press). doi:10.1016/j.imavis.2016.10.004.

Castrillón-Santana, M., Lorenzo-Navarro, J., Ramón-Balmaseda, E., 2016. Multi-scale score level fusion of local descriptors for gender classification in the wild. Multimedia Tools and Applications (in press). URL: http://dx.doi.org/10.1007/s11042-016-3653-2, doi:http://dx.doi.org/10.1007/s11042-016-3653-2.

Cattaneo, G., Roscigno, G., Petrillo, U.F., 2014. A scalable approach to source camera identification over hadoop, in: IEEE 28th International Conference on Advanced Information Networking and Applications (AINA).

Chai, Z., Sun, Z., Tan, T., Mendez-Vazquez, H., 2013. Local salient patterns - a novel local descriptor for face recognition, in: International Conference on Biometrics (ICB).

Chen, J., Shan, S., He, C., Zhao, G., Pietikainen, M., Chen, X., Gao, W., 2010. WLD: A robust local image descriptor. Pattern Analysis and Machine Intelligence, IEEE Transactions on 32, 1705–1720. doi:10.1109/TPAMI.2009.155.

Choi, S., Lee, Y., Lee, S., Park, K., Kim, J., 2011. Age estimation using a hierarchical classifier based on global and local facial features. Pattern Recognition 44, 1262–1281.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection, in: Schmid, C., Soatto, S., Tomasi, C. (Eds.), International Conference on Computer Vision & Pattern Recognition (CVPR), pp. 886–893.

Dantcheva, A., Elia, P., Ross, A., 2016. What else does your biometrics data reveal? A survey on soft biometrics. IEEE Transactions on Information Forensics And Security 11, 441–467.

Detector, A., 2016. Online age detector. URL: http://agedetector.tequnique.com/.

Eidinger, E., Enbar, R., Hassner, T., 2014. Age and gender estimation of unfiltered faces. Transactions on Information Forensics and Security, Special issue on Facial Biometrics in the Wild 9, 2170–2179.

Facebook, 2016. Face.com. URL: https://es-es.facebook.com/Face.com/.

Farid, H., 2009. Image forgery detection 26, 16–25.

Gallagher, A., Chen, T., 2009. Understanding images of groups of people, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 256–263.

Gao, F., Ai, H., 2009. Face age classification on consumer images with gabor feature and fuzzy lda method. Lecture Notes In Computer Science, in: Proceedings of the Third International Conference on Advances in Biometrics 5558, 132–141.

García-Olalla, O., Alegre, E., Fernández-Roble, L., González-Castro, V., 2014. Local oriented statistics information booster (LOSIB) for texture classification, in: International Conference on Pattern Recognition (ICPR).

Geng, X., Zhou, Z.H., Smith-Miles, K., 2008. Correction to "automatic age estimation based on facial aging patterns". IEEE Transactions on Pattern Analysis and Machine Intelligence 30, 368–368.

Guo, G., Fu, Y., Dyer, C., Huang, T., 2008. Image-based human age estimation by manifold learning and locally adjusted robust regression. IEEE Transactions on Image Processing 17, 115–128.

Guo, G., Mu, G., 2011. Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

He, Z., Li, X., Zhang, Z., Zhang, Y., Xiao, J., Zhou, X., 2016. Structure-aware slow feature analysis for age estimation. IEEE Signal Processing Letters 23, 1702–1706.

Heisele, B., Serre, T., Poggio, T., 2007. A component-based framework for face detection and identification. International Journal of Computer Vision Research 74.

Jenkins, P., 2009. Failure to launch: Why do some social issues fail to detonate moral panics? British Journal of Criminology 49, 35–47.

Jones, M.J., Rehg, J.M., 1998. Statistical Color Models with Application to Skin Detection. Technical Report Series CRL 98/11. Cambridge Research Laboratory.

Jun, B., Kim., D., 2012. Robust face detection using local gradient patterns and evidence accumulation. Pattern Recognition 45, 3304–3316.

KAIROS AR, I., 2016. Determining how old you are from photos and video. URL: https://www.kairos.com/blog/determining-how-old-you-are-from-photos-and-video.

Kawale, N., Patil, S., 2014. An approach to maintain the stroage of contentious image in the form of descriptor, in: IEEE International Conference on Computational Intelligence and Computing Research (ICCIC).

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Pereira, F., Burges, C., Bottou, L., Weinberger, K. (Eds.), Advances in Neural Information Processing Systems 25. Curran Associates, Inc., pp. 1097–1105.

Lanitis, A., Draganova, C., Christodoulou, C., 2004. Comparing different classifiers for automatic age estimation. IEEE Transactions on Systems, Man and Cybernetics 1, 621–628.

Lecun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition, in: Proceedings of the IEEE, pp. 2278 – 2324.

Levi, G., Hassner, T., 2015. Age and gender classification using convolutional neural networks, in: IEEE Workshop on Analysis and Modeling of Faces and Gestures (AMFG), at the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Boston. pp. 34–42.

Liberatore, M., Erdely, R., Kerle, T., Levine, B., Shields, C., 2010. Forensic investigation of peer-to-peer file sharing networks. Digital Investigation 7, 95–103.

Liu, L., Fieguth, P., Zhao, L., Long, Y., Kuang, G., 2012. Extended local binary patterns for texture classification. Image and Vision Computing 30, 86–99.

Liu, L., Sweeney, C., Arrieta, S., Lawrence, J., Verham, Z., 2017. HIPI hadoop image processing interface. URL: http://hipi.cs.virginia.edu/.

Liu, Z., Luo, P., Wang, X., Tang, X., 2015. Deep learning face attributes in the wild, in: International Conference on Computer Vision.

Mansanet, J., Albiol, A., Paredes, R., 2016. Local deep neural networks for gender recognition. Pattern Recognition Letters 70, 80–86.

Microsoft, 2016. How old do i look? URL: https://how-old.net/.

Nixon, M., Correia, P., Nasrollahi, K., Moeslund, T., Hadid, A., Tistarelli, M., 2015. On soft biometrics. Pattern Recognition Letters 68, Part 2, 218–230.

Ojansivu, V., Heikkilâ, J., 2008. Blur insensitive texture classification using local phase quantization, in: Elmoataz, A., Lezoray, O., Nouboud, F., Mammass, D. (Eds.), Image and Signal Processing, LNCS 5099. Springer, pp. 236–243.

Panis, G., Lanitis, A., Tsapatsoulis, N., Cootes, T.F., 2016. Overview of research on facial ageing using the fg-net ageing database , 37–46.

Qazi, T., Hayat, K., Khan, S.U., Madani, S.A., Khan, I.A., Ko??odziej, J., Li, H., Lin, W., Yow, K.C., Xu, C.Z., 2013. Survey on blind image forgery detection 7, 660–670.

Ricanek, K.J., Tesafaye, T., 2006. MORPH: A longitudinal image database of normal adult age-progression, in: IEEE 7th International Conference on Automatic Face and Gesture Recognition (FG), Southampton, UK. pp. 341–345.

Satta, R., Galbally, J., Beslay, L., 2014. Children gender recognition under unconstrained conditions based on contextual information, in: 22nd IEEE International Conference on Pattern Recognition (ICPR), Stockholm, Sweden.

Sengamedu, S.H., Sanyal, S., Satish, S., 2011. Detection of pornographic content in internet images, in: roceedings of the 19th ACM international conference on Multimedia.

Sokolova, M., Japkowicz, N., Szpakowicz, S., 2006. Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 1015–1021. URL: http://dx.doi.org/10.1007/11941439_114, doi:10.1007/11941439_114.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research 15, 1929–1958. URL: http://jmlr.org/papers/v15/srivastava14a.html.

Suo, J., T. Wu, S.Z., Shan, S., Chen, X., Gao, W., 2008. Smart sight: A tourist assistant system, in: Proc. of the Eighth IEEE International Conference on Automatic Face and Gesture Recognition (FG08).

Takimoto, H., Mitsukura, Y., Fukumi, M., Akamatsu, N., 2008. Robust gender and age estimation under varying facial pose. Electronics and Communications 7, 32–40.

Tan, X., Triggs, B., 2010. Enhanced local texture feature sets for face recognition under difficult lighting conditions. Image Processing, IEEE Transactions on 19, 1635 – 1650.

Thompson, G., 2009. Automatic detection of child pornography, in: Australian Digital Forensics Conference.

Vapnik, V., 1995. The nature of statistical learning theory. Springer, New York.

Viola, P., Jones, M.J., 2004. Robust real-time face detection. International Journal of Computer Vision 57, 151–173.

Wolak, J., Finkelhor, D., Mitchell, K., 2008. Child pornography possessors: trends in offender and case characteristics. Sexual abuse: a journal of research and treatment 23, 22–42.

Wolak, J., Finkelhor, D., Mitchell, K., Jones, L., 2011. Arrests for child pornography production: Data at two time points from a national sample of u.s. law enforcement agencies. Child Maltreatment 16, 184–195.

van de Wolfshaar, J., Karaaba, M.F., Wiering, M.A., 2015. Deep convolutional neural networks and support vector machines for gender recognition, in: IEEE Symposium Series on Computational Intelligence: Symposium on Computational Intelligence in Biometrics and Identity Management.