Fusion of CNN- and COSFIRE-based features with application to Gender Recognition from Face Images

Azzopardi, George; Simanjuntak, Frans

Published in:
Springer series "Advances in Intelligent Systems and Computing"

DOI:
10.1007/978-3-030-17795-9_33

Fusion of CNN- and COSFIRE-Based Features with Application to Gender Recognition from Face Images

Frans Simanjuntak and George Azzopardi(✉)

University of Groningen, Groningen, The Netherlands
franzteve@gmail.com, g.azzopardi@rug.nl

Abstract. Convolution neural networks (CNNs) have been demonstrated to be very effective in various computer vision tasks. The main strength of such networks is that features are learned from some training data. In cases where training data is not abundant, transfer learning can be used in order to adapt features that are pre-trained from other tasks. Similarly, the COSFIRE approach is also trainable as it configures filters to be selective for features selected from training data. In this study we propose a fusion method of these two approaches and evaluate their performance on the application of gender recognition from face images. In particular, we use the pre-trained VGGFace CNN, which when used as standalone, it achieved 97.45% on the GENDER-FERET data set. With one of the proposed fusion approaches the recognition rate on the same task is improved to 98.9%, that is reducing the error rate by more than 50%. Our experiments demonstrate that COSFIRE filters can provide complementary features to CNNs, which contribute to a better performance.

Keywords: VGGFace · COSFIRE · Fusion · Gender recognition

1 Introduction

Convolutional neural networks (CNNs) are advanced versions of the original networks introduced by Fukushima [1]. They are inspired by the feline visual processing system. The architecture of a CNN is built on its predecessor, the ordinary neural network, which has layers that receive input, has activation functions, contains neurons with learnable weights and biases, and performs forward and backward propagation to adjust the network. The fundamental difference between a CNN and an ordinary neural network is that it contains a set of stacked convolutional-pooling pairs of layers in which the output of the last layer is fed to a set of stacked fully connected layers. Since few years ago, CNNs have become state-of-the-art for object detection and image classification. Their effectiveness is attributable to their ability to learn features from training data, instead of handcrafting them. In applications were training data is abundant, a

K. Arai and S. Kapoor (Eds.): CVC 2019, AISC 943, pp. 444–458, 2020.
https://doi.org/10.1007/978-3-030-17795-9_33

new network can be entirely trained from such data. In other applications, however, where training data is limited, one may apply transfer learning techniques to fine tune only the last layer(s) of the network, for instance.

Another approach that uses trainable features is called Combination of Shifted Response (COSFIRE) filter. Unlike CNNs, COSFIRE configures nonlinear filters that achieve tolerance to rotation, scale and reflection. The complexity of the preferred pattern can vary from a simple edge or a line to corners, curvatures, bifurcations and shapes of whole objects like traffic signs. So far, COSFIRE filters have been configured by presenting single training examples. In principle, however, learning algorithms can be applied in order to determine the selectivity of the filters from multiple training examples.

One major challenge for CNNs is dealing with adversarial attacks. The linear nature of the convolutional layers makes them vulnerable to such attacks [2]. On the other hand, COSFIRE filters rely on nonlinear connections between the output of low-level filters and, in principal, they are more robust to adversarial attacks.

Considering the fact that both the CNN and COSFIRE approaches are based on features determined from training data, it is intriguing to investigate a fusion approach that maximizes their strengths. In this study, we fuse the trainable features from CNNs and COSFIRE filters by applying two types of fusion, namely feature fusion and decision fusion. In the former strategy, we concatenate CNN and COSFIRE-based features and use the resulting feature vector as input to another classification model, and in the latter, we learn a stacked classification model without merging the CNN- and COSFIRE-based features. In general, the fusion approaches that we investigate are applicable to any classification task, however, for the sake of demonstration, we use the application of gender recognition from face images to quantify their effectiveness.

The rest of the paper is organized as follows. In Sect. 2 we give an account of related works. In Sect. 3 we describe the proposed approach followed by Sect. 4 where we explain the experiments and report the results. In Sect. 5 we provide a discussion and finally we draw conclusions in Sect. 6.

2 Related Works

CNNs have been applied in many computer vision tasks, such as face recognition [3,4], scene labelling [5–7], image classification [8–12], action recognition [13–15], human pose estimation [16–18], and document analysis [19–22]. CNNs made a breakthrough in image classification in the ImageNet Large-Scale Visual Recognition Challenge in 2014 (ILSVRC14). In that competition, GoogleNet was able to perform the classification and detection of large scale images by achieving an error rate of 6.67% [23], which is very close to human level performance. Various CNN architectures have been proposed, namely Lenet [24], AlexNet [11], VGGNet [25], and ResNet [26]. Closer to the application at hand, Parkhi et al. [4] also proposed an architecture called VGGFace, which is designed to recognize the identity of a person from a single photograph or a set of faces tracked

in a video. CNNs have also been adapted to natural language processing applications, such as speech recognition [27–29] and text classification [30–33]. In general, only 4% Word Error Rate Reduction (WERR) was obtained when the speech was trained on 1000 h of Kinect distance [34] using deep neural networks (DNNs) proposed in [35].

The other approach that we are concerned with, namely, Combination of Shifted Filter Responses (COSFIRE), is also a brain-inspired visual pattern recognition method, which has been effectively applied in various applications, including traffic sign recognition [36], handwritten digit recognition [37], architectural symbol recognition [38], quality visual inspection [39], contour detection [40], delineation of curvilinear structures [41], such as the blood vessels in retinal fundus images and cracks in pavements [42], butterfly recognition [36], person identification from retinal images [43], and gender recognition from face images [44,45]. COSFIRE is a filtering approach whose selectivity is determined in a configuration stage by the automatic analysis of given prototype patterns. In its basic architecture, a COSFIRE filter takes input from a set of low-level filters, such as orientation-selective or filters with center-surround support, and combines them by a nonlinear function [46]. In [47], it has also been demonstrated that hierarchical or multi-layered COSFIRE filters can be configured to be selective for more deformable objects. Similar to CNNs, the number of layers is an architectural design.

In relation to gender recognition from face images, several studies have been conducted using CNNs and COSFIRE. For instance, Liew et al. [48] proposed a CNN architecture which focuses on reducing CNN layers to four and performs cross-correlation to reduce the computation time. The proposed method was evaluated using two public face data sets, namely SUMS and AT&T, and achieved accuracy rates of 98.75% and 99.38%, respectively. Levi et al. [49] introduced a CNN architecture which focuses on age and gender classification. They proposed an architecture, so-called deep convolutional neural networks (DCNN), which has two approaches, namely single crop and over samples and achieved accuracy rates of 85.9% and 86.8%, respectively, on the Audiance benchmark dataset. Dhomne et al. [50] also proposed a deep CNN architecture for gender recognition that focuses mainly on enhancing VGGNet architecture to be more efficient.

Studies have shown that CNNs are vulnerable to adversarial attacks. For instance, Narodytska et al. [51] proposed a simple attack by adding perturbation and applying a small set of constructed pixels using greedy local-search to a random location of the image. That method is able to fool a CNN and increases the misclassification rate. Moosavi-Dezfooli et al. [52] proposed DeepFool based on iterative linearization procedure to generate adversarial attacks. Tang et al. [53] employed a steganographic scheme that aims at hiding a stego message and fooling a CNN at the same time. The experiments showed that it is secure and adequate to cope with powerful CNN-based steganalysis.

Several methods based on the COSFIRE approach have also been proposed for the recognition of gender from face images. The first experiment was performed in 2016 [44] where a set of COSFIRE filters were used and encoded

by a spatial pyramid to form a feature vector that was fed to a classification model. In the experiments, they used two data sets, namely GENDER-FERET and Labeled Faces in the Wild (LFW), and the results show that COSFIRE is able to achieve 93.7% classification rate on the GENDER-FERET and 90% on the LFW. In the following year, they continued the work by conducting another experiment that combines the features from domain-specific and trainable COSFIRE [45]. In that study, the domain specific part uses SURF descriptors from 51 facial landmarks related to the nose, eyes, and mouth. The extracted features from those landmarks were fused with features from COSFIRE filters and achieved accuracy rates of 94.7%, 99.4%, and 91.5% on the GENDER-FERET, LFW, and UNISA-Public data sets, respectively.

3 Methods

In the following, we describe the proposed fusion methods within the context of gender recognition from face images. First, we describe how we perform face detection and if necessary correct the orientation of the face to an upright position. Then, we describe VGGFace as one of the most widely used CNN architectures for face recognition, followed by the COSFIRE filter approach. Finally, we elaborate on our fusion approaches. Figure 1 illustrates the high-level architecture of our pipeline.

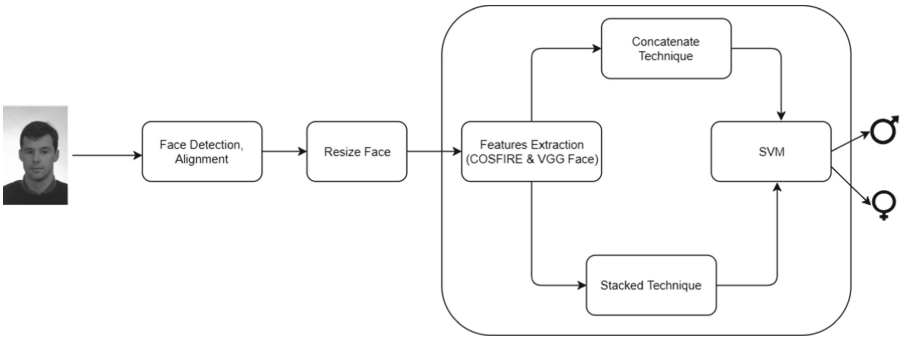


Fig. 1. A high level diagram of the proposed system.

3.1 Face Detection and Alignment

For a given face image, we apply the algorithm proposed by Uricar *et al.* [54] that gives us 68 fiducial landmarks. For the purpose of this work we determine the two centroids of the locations of the two sets of points that characterize the eyes and discard the remaining landmarks. Next, we compute the angle between these two centroids and use it to rotate the face image in such a way that the angle between the eyes becomes zero. Thereby, we ensure that all face images are appropriately aligned.

Finally, we apply the Viola-Jones algorithm [45] to crop the close-up of the face. Figure 2 illustrates the preprocessing pipeline that we employ.

(a) (b) (c) (d)

Fig. 2. Input image containing the fiducial landmarks detected by the algorithm proposed in [54]. The blue spots indicate the centroids of the landmarks that describe the eyes. (b) Image rotated appropriately based on the angle found between the blue landmarks in (a). (c) Face detection with the Viola-Jones algorithm and (d) the cropped face image.

3.2 VGGFace

VGGFace is an extended CNN of VGGNet developed by Parkhi *et al.* [4]. They showed that the depth of the network is a critical component for good performance. The goal of VGGFace architecture is to deal with face recognition either from a single photograph or a set of faces tracked in a video [4]. The input to the VGGFace is a face image of size 224×224 pixels and the network consists of 13 convolutional layers, 15 Rectified Linear Units (ReLu), 5 sub sampling (max pooling) layers, 3 fully connected layers, and 1 softmax probability as shown in Fig. 3. For further technical details on VGGFace we refer the reader to [4].

In this study, we use the pre-trained VGGFace to extract features from face images. Following the requirement of VGGFace we resize our pre-processed images to 224×224 pixels. We apply the VGGFace to every given image and take the 4096-element feature vector from the FC7 layer. Finally, we stretch the feature vectors between 0 and 1 such that they share the same range of values of the COSFIRE approach.

3.3 COSFIRE

Combination Of Shifted Filter Responses (COSFIRE) is a trainable filter approach which has been demonstrated to be effective in various computer vision tasks. Here we apply the COSFIRE filters in the same way as proposed in [44] where a spatial pyramid was employed to form a feature vector with COSFIRE features. For completeness sake we briefly describe this method in the following sub-sections.

COSFIRE Filter Configuration. A COSFIRE filter is nonlinear and it is automatically configured to be selective for a given single prototype pattern of interest. The automatic configuration procedure consists of two main steps: *convolution* followed by *keypoint detection* and *description*. In the convolution step a bank of orientation-selective (Gabor) filters is applied with different scales λ and orientations θ, and the resulting feature maps are superimposed on top of each other. In the second step, a set of concentric circles with given radii ρ is considered and the local maximum Gabor responses along those circles are identified as keypoints. Each keypoint i is described with four parameters: $(\lambda_i, \theta_i, \rho_i, \phi_i)$, where λ_i and θ_i are the parameters of the Gabor filter that achieves the maximum response in the location with a distance ρ_i and polar angle ϕ_i with respect to the center of the prototype pattern.

layer	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
type	input	conv	relu	conv	relu	mpool	conv	relu	conv	relu	mpool	conv	relu	conv	relu	conv	relu	mpool	conv
name	–	conv1_1	relu1_1	conv1_2	relu1_2	pool1	conv2_1	relu2_1	conv2_2	relu2_2	pool2	conv3_1	relu3_1	conv3_2	relu3_2	conv3_3	relu3_3	pool3	conv4_1
support	–	3	1	3	1	2	3	1	3	1	2	3	1	3	1	3	1	2	3
filt dim	–	3	–	64	–	–	64	–	128	–	–	128	–	256	–	256	–	–	256
num filts	–	64	–	64	–	–	128	–	128	–	–	256	–	256	–	256	–	–	512
stride	–	1	1	1	1	2	1	1	1	1	2	1	1	1	1	1	1	2	1
pad	–	1	0	1	0	0	1	0	1	0	0	1	0	1	0	1	0	0	1

layer	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37
type	relu	conv	relu	conv	relu	mpool	conv	relu	conv	relu	conv	relu	mpool	conv	relu	conv	relu	conv	softmx
name	relu4_1	conv4_2	relu4_2	conv4_3	relu4_3	pool4	conv5_1	relu5_1	conv5_2	relu5_2	conv5_3	relu5_3	pool5	fc6	relu6	fc7	relu7	fc8	prob
support	1	3	1	3	1	2	3	1	3	1	3	1	2	7	1	1	1	1	1
filt dim	–	512	–	512	–	–	512	–	512	–	512	–	–	512	–	4096	–	4096	–
num filts	–	512	–	512	–	–	512	–	512	–	512	–	–	4096	–	4096	–	2622	–
stride	1	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1	1	1	1
pad	0	1	0	1	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0

Fig. 3. The architecture of the VGGFace as proposed by Parkhi *et al.* [4]. The red bounding box indicates the FC7 layer that we use to extract features from the network.

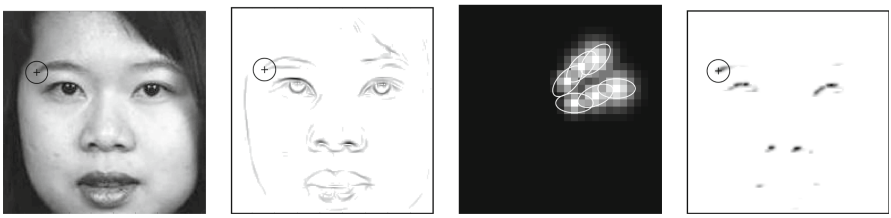
Therefore, a COSFIRE filter is defined as a set of 4-tuples:

$$S_f = \{(\lambda_i, \theta_i, \rho_i, \phi_i) \mid i = 1...k\} \tag{1}$$

where the subscript f represents the prototype pattern and k denotes the number of keypoints.

In our experiments we configure multiple COSFIRE filters with equal number of randomly selected local patterns from male and female face training images. Figure 4 illustrates the configuration of a COSFIRE filter with a local pattern selected from a face image, as well as its application to the same image.

COSFIRE Filter Response. The response of a COSFIRE filter is computed by combining with a geometric mean the intermediate response maps generated from the tuples describing the filter. There is a pipeline of four operations applied for each tuple in a given COSFIRE filter. It consists of *convolution*, *ReLU*, *blurring* and *shifting*. The pipelines of the tuples can be run in parallel as they are independent of each other. In the convolution step, the given image is filtered



(a) (b) (c) (d)

Fig. 4. Configuration example of a COSFIRE filter using a training female face image of size 128×128 pixels. The encircled region in (a) shows a randomly selected pattern of interest that is used to configure a COSFIRE filter. (b) Superposition of a bank of antisymmetric Gabor filters with 16 orientations ($\theta = \{0, \pi/8, ...15\pi/8\}$ and a single scale ($\lambda = 4$). (c) The structure of the COSFIRE filter that is selective for the encircled pattern in (a). (d) The inverted response map of the COSFIRE filter to the input image in (a).

with the Gabor filter with scale λ_i and orientation θ_i as specified in tuple i. Next, similar to CNNs, the Gabor response map is rectified with the ReLU function. Unlike the pooling layer of the CNNs, COSFIRE applies a max blurring function to the rectified Gabor responses in order to allow for some tolerance with respect to the preferred position of the concerned keypoint followed by shifting by ρ_i pixels in the direction opposite to ϕ_i. The blurring operation uses a sliding window technique on the Gabor response maps, where the Gabor responses in every window are weighted with a Gaussian function whose standard deviation σ grows linearly with the distance ρ_i: $\sigma = \sigma_0 + \alpha\rho_i$ where σ_0 and α are determined empirically. The output of the blurring operation is the weighted maximum. Finally, all blurred and shifted Gabor responses $s_{\lambda_i,\sigma_i,\rho_i,\phi_i}(x, y)$ are combined using geometric mean and the result is denoted by r_{S_f}:

$$r_{S_f}(x, y) = \left(\prod_{i=1}^{n} s_{\lambda_i,\sigma_i,\rho_i,\phi_i}(x, y) \right)^{\frac{1}{n}} \tag{2}$$

Face Descriptor. In contrast to CNNs, instead of downsizing the feature maps and eventually flattening the last map into a feature vector, we treat the COSFIRE response maps by a spatial pyramid of three levels. In level zero we consider one tile, which is the same size of the COSFIRE response map and take the maximum value. In the next two levels we consider 2×2 and 4×4 tiles, respectively, and take the maximum COSFIRE response in each tile. For n COSFIRE filters and $(1 + 4 + 16 =)$ 21 tiles, we describe a face image with a $21n$-element feature vector. Moreover, the set of n COSFIRE filter responses per tile is normalized to unit length [45]. An example of the COSFIRE face descriptor using a single filter is shown in Fig. 5.

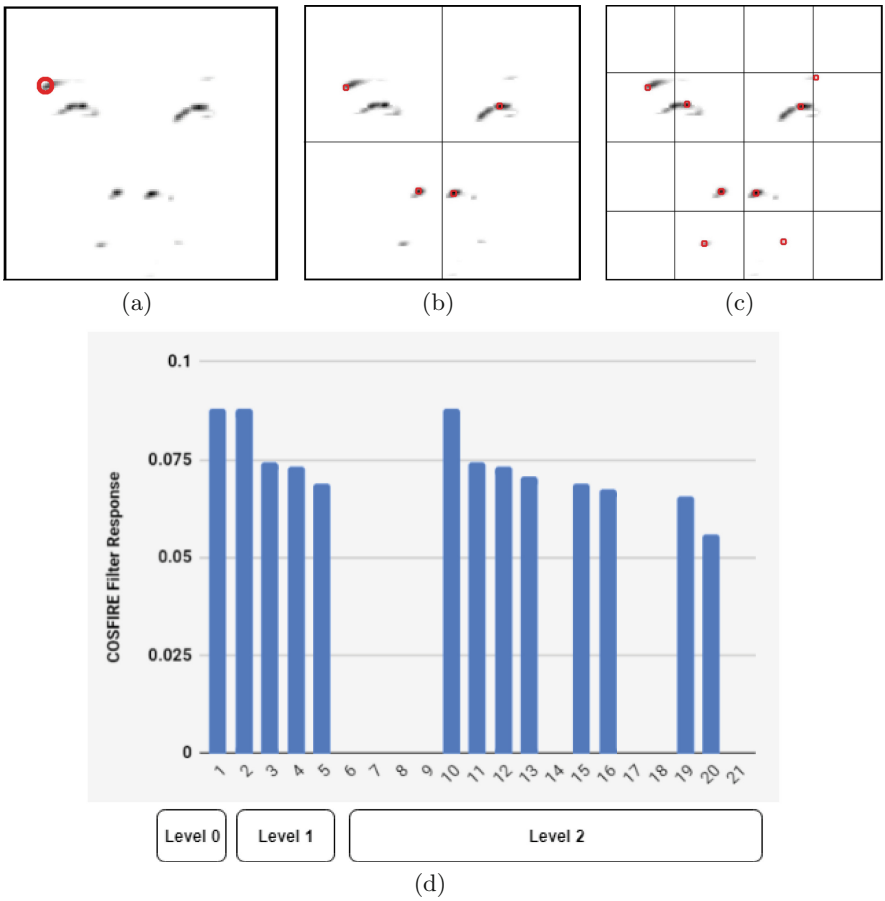


Fig. 5. Application of a COSFIRE filter to a face image using a spatial pyramid of three layers, with square grids of 1, 4 and 16 tiles. The red circles in (a–c) indicate the maximum values within the tiles which are shown in the bar plot in (d).

3.4 Fusion Methods

We investigate two fusion strategies that combine CNN- and COSFIRE-based features, one which combines features and the other which combines the decisions of two separate classifiers.

Feature Fusion. In the first approach we concatenate the 4096-element feature vector of VGGFace that is extracted from the FC7 layer with the $21n$-element feature vector obtained by the spatial pyramid approach employed to COSFIRE feature maps. The value of n represents the number of COSFIRE filters. Here, we set $n = 240$, a value that was determined empirically. Combining the two sets of features results in a fused feature vector of $(4096 + 21 \times 240 =)$ 9136

elements. Finally, we used the resulting fused feature vectors from the training data to learn an SVM classification model with linear kernel.

Decision Fusion. The other approach that we investigate is called stacked classification. In this approach we keep the CNN and COSFIRE feature vectors separately and learn an SVM (with linear kernel) classification model for each set of features. Then we apply the SVMs to the training data and combine the returned values of both SVMs in a feature vector which we use to learn another classification SVM model with linear kernel. The application of SVMs return as many values as the number of classes. In our case we have two classes (male and female), so each SVM in the lower layer returns two values that are related to the probabilities of having a certain gender. Subsequently, the SVM in the top layer is fed with a vector of four values.

4 Experiments and Results

Here, we describe the experimental design along with the data sets that we used and the results obtained for both standalone methods and fusion strategies.

4.1 Data Sets

We used two benchmark data sets, namely GENDER-FERET [55] and Labeled Face in the Wild (LFW) [56]. The GENDER-FERET data set consists of 946 face images of people with different expression, age, race, and pose in controlled environment. The data set, which is publicly available[1], is already divided equally into training and test sets. The LFW data set gives us the opportunity to validate the proposed methods in unconstrained environment. It consists of 13,000 images of 5,749 celebrity, athlete, and politician faces collected from websites when the subjects were doing their daily activities, such as playing sports, doing a fashion show, giving a speech, doing an interview, among others. Looking at the facts that the photographs of the subjects were taken in their natural environment, multiple faces may appear in the same image. Also, the data set shows variability in illumination, pose, background, occlusions, facial expression, gender, age, race, and image quality.

Following the recommendations in [44] and [45], 9,763 grayscale images were chosen for the experiment in which 2,293 are females while the rest are males. We labeled the gender manually as it is not provided with the data set. The images were aligned to an upright position using facial landmark tracking [54] as explained in the previous section and whenever we were in dilemma we discarded the faces whose gender was not easy to establish. Since the number of images between male and female is not balanced, we applied 5-fold cross-validation by partitioning images into five subsets of similar size and keeping the same ratio between male and female [57]. Then, the accuracy was computed by taking the average of all folds.

[1] https://www.nist.gov/programs-projects/face-recognition-technology-feret.

Table 1. Results of the COSFIRE and VGGFace-based standalone approaches on the GENDER-FERET and LFW data sets using SVM classifier.

Method	Data set	Accuracy (%)
COSFIRE-only	GF	93.85
	LFW	99.19
CNN-only	GF	97.45
	LFW	99.71
Feature fusion	GF	98.30
	LFW	99.28
Decision fusion	GF	98.94
	LFW	99.38

4.2 Experiments

Below, we report the evaluation of the standalone approaches followed by the evaluation of the two above mentioned fusion strategies.

Following similar procedures as explained in [44] and [45], we conducted several experiments with the COSFIRE-based method on the GENDER-FERET and LFW data sets. Instead of employing 180 filters as used in the prior works, we configured 240 COSFIRE filters in order to have more variability. Of the 240 COSFIRE filters, 120 are configured from randomly selected local patterns (of size 19×19 pixels) of male face training images and the other half from randomly selected local patterns of female training faces. If a randomly selected local pattern was sufficiently salient and resulted in a COSFIRE filter that consisted of at least five keypoints (tuples), then we considered it as a valid prototype, otherwise we discarded it and chose another random pattern. As suggested in [44], we set the parameters of the COSFIRE filters as follows: $t_1 = 0.1$, $t_2 = 0.75$, $\sigma_0 = 0.67$, $\alpha = 0.1$ and selected keypoints from a set of concentric circles with radii $\rho = \{0, 3, 6, 9\}$. For the CNN-based approach we used the 4096-element features vectors along with SVM with linear kernel.

Moreover, we conducted other experiments where we evaluated the two fusion strategies that combine both approaches. The first is referred to as feature fusion where we concatenated the VGG-Face and COSFIRE feature vectors into longer ones, and the other is decision fusion where we used a classification stacking approach as explained in Sect. 3.4.

Table 1 reports the results obtained by the two standalone and the two fusion approaches to the GF and LFW data sets. For the GF data set, the standalone CNN-based approach performs significantly better than the standalone COSFIRE approach, and for the other data set the accuracies of both standalone methods are roughly the same, with the marginal difference not being statistical significant. As to the fusion, we observe that both strategies improve the accuracy rate with high statistical significance on the GF data set, while there is no statistical difference between the results of all methods for the LFW data set.

Comparison with Other Methods. We also compare the results of our approaches with those already published in the literature, Table 2. For the GF data set both fusion strategies that we propose outperform existing works with high statistical significance. For the LFW data set, we do not observe statistical difference between any of the methods.

Table 2. Comparison of the results between the proposed approaches and existing ones on both the GF and LFW data sets.

	Method	Description	Accuracy (%)
GF data set	Azzopardi *et al.* [58]	RAW LBP HOG	92.60
	Azzopardi *et al.* [44]	COSFIRE	93.70
	Azzopardi *et al.* [45]	COSFIRE SURF	94.70
	Proposed 1 (Feature fusion)	COSFIRE VGGFACE	98.30
	Proposed 2 (Decision fusion)	COSFIRE VGGFACE	98.90
LFW data set	Tapia *et al.* [59]	LBP	92.60
	Dago-Casa *et al.* [60]	Gabor	94.00
	Shan *et al.* [57]	Boosted LBP	94.81
	Azzopardi *et al.* [45]	COSFIRE SURF	99.40
	Proposal 1 (Feature fusion)	COSFIRE VGGFACE	99.28
	Proposal 2 (Decision fusion)	COSFIRE VGGFACE	99.38

5 Discussion

The most important contribution of this study is that COSFIRE and CNN features from a pre-trained CNN can indeed complement each other for gender recognition from face images. The experiments on the GENDER-FERET data set demonstrate this complementarity where the decision fusion approach reduced the error rate by more than 50%. For the other data set, the fact that both standalone (CNN and COSFIRE) methods achieved very high recognition rates (above 99%), left very little room for further improvement when fused together. We are eager to find out whether the same or similar improvement can be observed in other challenging recognition applications, and aim to investigate this matter in our future works.

Both COSFIRE and CNNs are approaches that learn features directly from the given training data. CNNs with high number of layers, such as the VGG-Face that we use here, rely on deep learning with gradient descent to determine the best features, while COSFIRE is conceptually simpler as it configures the selectivity of a filter from a single prototype pattern by establishing the mutual spatial arrangement of keypoints within the given local pattern. So far, COSFIRE filters have been configured with single examples with empirically-determined setting of hyper parameters, such as the standard deviations of the blurring functions. In future, we will investigate a learning mechanism that can determine better filters by analyzing multiple prototypes.

COSFIRE filters share important steps with CNNs, including the convolution and the ReLU layers along with the possibility of designing architectures with multiple layers. The fundamental difference between the two approaches lies in

the fact that COSFIRE is a filter, while a CNN is a fully embedded classification technique. The input to the COSFIRE approach can be any complex scene, while for a CNN to detect an object of interest the given input image must contain the concerned object roughly in the center and must take the majority of the space. The latter requirement is due to the downsizing decision that CNNs implement, a step that is not present in the COSFIRE approach. Instead, COSFIRE applies a blurring function that allows some tolerance with respect to the mutual spatial arrangement of the defining features of an object of interest.

6 Conclusions

The proposed fusion strategies prove to be very effective in combining the COSFIRE and CNN approaches. We used a case study of gender recognition to evaluate our methods and it turned out that with the fusion approaches the error rate drops by more than 50% on the GENDER-FERET data set. Considering the simplicity of the COSFIRE filters, the achieved results are very promising. The proposed fusion approaches are independent of the application at hand and thus they can be adapted to any image classification task.

References

1. Fukushima, K.: Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol. Cybern. **36**(4), 193–202 (1980)
2. Goodfellow, I., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: International Conference on Learning Representations (2015)
3. Lawrence, S., Giles, C.L., Tsoi, A.C., Back, A.D.: Face recognition: a convolutional neural-network approach. IEEE Trans. Neural Netw. **8**(1), 98–113 (1997)
4. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: British Machine Vision Conference (2015)
5. Farabet, C., Couprie, C., Najman, L., LeCun, Y.: Learning hierarchical features for scene labeling. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1915–1929 (2013)
6. Pinheiro, P., Collobert, R.: Recurrent convolutional neural networks for scene labeling. In: Xing, E.P., Jebara, T. (eds.) Proceedings of the 31st International Conference on Machine Learning Research, PMLR, 22–24 June 2014, Beijing, China, vol. 32, pp. 82–90 (2014)
7. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 640–651 (2017)
8. Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., Chen, T.: Recent advances in convolutional neural networks. Pattern Recogn. **77**(C), 354–377 (2018)
9. Strigl, D., Kofler, K., Podlipnig, S.: Performance and scalability of GPU-based convolutional neural networks. In: 2010 18th Euromicro Conference on Parallel, Distributed and Network-based Processing, pp. 317–324, February 2010
10. Uetz, R., Behnke, S.: Large-scale object recognition with CUDA-accelerated hierarchical neural networks. In: 2009 IEEE International Conference on Intelligent Computing and Intelligent Systems, vol. 1, pp. 536–541, November 2009

11. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems, vol. 25, pp. 1097–1105. Curran Associates, Inc. (2012)
12. Yan, Z., Jagadeesh, V., DeCoste, D., Di, W., Piramuthu, R.: HD-CNN: hierarchical deep convolutional neural network for image classification. CoRR, abs/1410.0736 (2014)
13. Kim, H.-J., Lee, J.S., Yang, H.-S.: Human action recognition using a modified convolutional neural network. In: Proceedings of the 4th International Symposium on Neural Networks: Part II–Advances in Neural Networks, ISNN 2007, pp. 715–723. Springer, Heidelberg (2007)
14. Le, Q.V., Zou, W.Y., Yeung, S.Y., Ng, A.Y.: Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, pp. 3361–3368. IEEE Computer Society, Washington (2011)
15. Wang, P., Cao, Y., Shen, C., Liu, L., Shen, H.T.: Temporal pyramid pooling based convolutional neural networks for action recognition. CoRR, abs/1503.01224 (2015)
16. Ferrari, V., Marin-Jimenez, M., Zisserman, A.: Progressive search space reduction for human pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (2008)
17. Weiss, D.J., Sapp, B., Taskar, B.: Sidestepping intractable inference with structured ensemble cascades. In: NIPS, pp. 2415–2423. Curran Associates, Inc. (2010)
18. Toshev, A., Szegedy, C.: Deeppose: human pose estimation via deep neural networks. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, pp. 1653–1660. IEEE Computer Society, Washington (2014)
19. Guyon, I., Albrecht, P., Le Cun, Y., Denker, J., Hubbard, W.: Design of a neural network character recognizer for a touch terminal. Pattern Recogn. **24**(2), 105–119 (1991)
20. Zhu, R., Mao, X., Zhu, Q., Li, N., Yang, Y.: Text detection based on convolutional neural networks with spatial pyramid pooling. In: 2016 IEEE International Conference on Image Processing (ICIP), pp. 1032–1036, September 2016
21. Bengio, Y., LeCun, Y., Henderson, D.: Globally trained handwritten word recognizer using spatial representation, convolutional neural networks, and hidden Markov models. In: Cowan, J.D., Tesauro, G., Alspector, J. (eds.) Advances in Neural Information Processing Systems, vol. 6, pp. 937–944. Morgan-Kaufmann (1994)
22. Yin, X., Yin, X., Huang, K., Hao, H.: Robust text detection in natural scene images. IEEE Trans. Pattern Anal. Mach. Intell. **36**(5), 970–983 (2014)
23. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Computer Vision and Pattern Recognition (CVPR) (2015)
24. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. In: Proceedings of the IEEE, pp. 2278–2324 (1998)
25. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR (2014)
26. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR, abs/1512.03385 (2015)
27. Abdel-Hamid, O., Mohamed, A.-R., Jiang, H., Deng, L., Penn, G., Yu, D.: Convolutional neural networks for speech recognition. IEEE/ACM Trans. Audio Speech Lang. Proc. **22**(10), 1533–1545 (2014)

28. Mao, Q., Dong, M., Huang, Z., Zhan, Y.: Learning salient features for speech emotion recognition using convolutional neural networks. IEEE Trans. Multimedia **16**, 2203–2213 (2014)
29. Santos, R.M., Matos, L.N., Macedo, H.T., Montalvão, J.: Speech recognition in noisy environments with convolutional neural networks. In: 2015 Brazilian Conference on Intelligent Systems (BRACIS), pp. 175–179, November 2015
30. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, 25–29 October 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, pp. 1746–1751 (2014)
31. Kalchbrenner, N., Grefenstette, E., Blunsom, P.: A convolutional neural network for modelling sentences. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), June 2014, Baltimore, Maryland, pp. 655–665. Association for Computational Linguistics (2014)
32. Wang, P., Xu, J., Xu, B., Liu, C., Zhang, H., Wang, F., Hao, H.: Semantic clustering and convolutional neural network for short text categorization. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pp. 352–357. Association for Computational Linguistics (2015)
33. Johnson, R., Zhang, T.: Effective use of word order for text categorization with convolutional neural networks. In: The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2015, 31 May–5 June 2015, Denver, Colorado, USA, pp. 103–112 (2015)
34. Wasenmüller, O., Stricker, D.: Comparison of kinect v1 and v2 depth images in terms of accuracy and precision, November 2016
35. Huang, J., Li, J., Gong, Y.: An analysis of convolutional neural networks for speech recognition. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4989–4993, April 2015
36. Gecer, B., Azzopardi, G., Petkov, N.: Color-blob-based COSFIRE filters for object recognition. Image Vis. Comput. **57**(C), 165–174 (2017)
37. Azzopardi, G., Petkov, N.: A shape descriptor based on trainable COSFIRE filters for the recognition of handwritten digits. In: Wilson, R., Hancock, E., Bors, A., Smith, W. (eds.) Computer Analysis of Images and Patterns, pp. 9–16. Springer, Heidelberg (2013)
38. Guo, J., Shi, C., Azzopardi, G., Petkov, N.: Recognition of architectural and electrical symbols by COSFIRE filters with inhibition. In: CAIP (2015)
39. Fernández-Robles, L., Azzopardi, G., Alegre, E., Petkov, N., Castejón-Limas, M.: Identification of milling inserts in situ based on a versatile machine vision system. J. Manuf. Syst. **45**, 48–57 (2017)
40. Azzopardi, G., Rodríguez-Sánchez, A., Piater, J., Petkov, N.: A push-pull CORF model of a simple cell with antiphase inhibition improves SNR and contour detection. PLOS One **9**(7), 1–13 (2014)
41. Strisciuglio, N., Petkov, N.: Delineation of line patterns in images using B-COSFIRE filters. In: 2017 International Conference and Workshop on Bioinspired Intelligence (IWOBI), pp. 1–6, July 2017
42. Strisciuglio, N., Azzopardi, G., Petkov, N.: Detection of curved lines with B-COSFIRE filters: a case study on crack delineation. CoRR, abs/1707.07747 (2017)
43. Azzopardi, G., Strisciuglio, N., Vento, M., Petkov, N.: Trainable COSFIRE filters for vessel delineation with application to retinal images. Med. Image Anal. **19**(1), 46–57 (2015)

44. Azzopardi, G., Greco, A., Vento, M.: Gender recognition from face images with trainable COSFIRE filters. In: 2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 235–241, August 2016
45. Azzopardi, G., Greco, A., Saggese, A., Vento, M.: Fusion of domain-specific and trainable features for gender recognition from face images. IEEE Access **6**, 24171–24183 (2018)
46. Azzopardi, G., Petkov, N.: Trainable COSFIRE filters for keypoint detection and pattern recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**(2), 490–503 (2013)
47. Azzopardi, G., Petkov, N.: Ventral-stream-like shape representation: from pixel intensity values to trainable object-selective COSFIRE models. Front. Comput. Neurosci. **8**, 80 (2014)
48. Liew, S.S., Khalil-Hani, M., Radzi, F., Bakhteri, R.: Gender classification: a convolutional neural network approach. Turkish J. Electr. Eng. Comput. Sci. **24**, 1248–1264 (2016)
49. Levi, G., Hassncer, T.: Age and gender classification using convolutional neural networks. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 34–42, June 2015
50. Dhomne, A., Kumar, R., Bhan, V.: Gender recognition through face using deep learning. Procedia Comput. Sci. **132**, 2–10 (2018). International Conference on Computational Intelligence and Data Science
51. Narodytska, N., Kasiviswanathan, S.P.: Simple black-box adversarial perturbations for deep networks. CoRR, abs/1612.06299 (2016)
52. Moosavi-Dezfooli, S.-M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. CoRR, abs/1511.04599 (2015)
53. Tang, W., Li, B., Tan, S., Barni, M., Huang, J.: CNN based adversarial embedding with minimum alteration for image steganography. CoRR, abs/1803.09043 (2018)
54. Uricar, M., Franc, V., Hlavac, V.: Facial landmark tracking by tree-based deformable part model based detector. In: 2015 IEEE International Conference on Computer Vision Workshop (ICCVW), pp. 963–970, December 2016
55. Gender recognition dataset. http://mivia.unisa.it/datasets/video-analysis-datasets/gender-recognition-dataset/. Accessed 28 May 2018
56. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: a database for studying face recognition in unconstrained environments. Technical report 07-49, University of Massachusetts, Amherst, October 2007
57. Shan, C.: Learning local binary patterns for gender classification on real-world face images. Pattern Recogn. Lett. **33**(4), 431–437 (2012). Intelligent Multimedia Interactivity
58. Azzopardi, G., Greco, A., Vento, M.: Gender recognition from face images using a fusion of SVM classifiers. In: Campilho, A., Karray, F. (eds.) Image Analysis and Recognition, pp. 533–538. Springer, Cham (2016)
59. Tapia, J.E., Perez, C.A.: Gender classification based on fusion of different spatial scale features selected by mutual information from histogram of LBP, intensity, and shape. IEEE Trans. Inf. Forensics Secur. **8**(3), 488–499 (2013)
60. Dago-Casas, P., González-Jiménez, D., Yu, L.L., Alba-Castro, J.L.: Single- and cross- database benchmarks for gender classification under unconstrained settings. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 2152–2159, November 2011