Neural Computing and Applications (2021) 33:11905–11922
https://doi.org/10.1007/s00521-021-05863-5

ORIGINAL ARTICLE

On the correlation between human fixations, handcrafted and CNN features

Marinella Cadoni[1] · Andrea Lagorio[1] · Souad Khellat-Kihel[2] · Enrico Grosso[1]

Received: 31 July 2020 / Accepted: 19 February 2021 / Published online: 19 March 2021

Abstract
Traditional local image descriptors such as SIFT and SURF are based on processings similar to those that take place in the early visual cortex. Nowadays, convolutional neural networks still draw inspiration from the human vision system, integrating computational elements typical of higher visual cortical areas. Deep CNN's architectures are intrinsically hard to interpret, so much effort has been made to dissect them in order to understand which type of features they learn. However, considering the resemblance to the human vision system, no enough attention has been devoted to understand if the image features learned by deep CNNs and used for classification correlate with features that humans select when viewing images, the so-called human fixations, nor if they correlate with earlier developed handcrafted features such as SIFT and SURF. Exploring these correlations is highly meaningful since what we require from CNNs, and features in general, is to recognize and correctly classify objects or subjects relevant to humans. In this paper, we establish the correlation between three families of image interest points: human fixations, handcrafted and CNN features. We extract features from the feature maps of selected layers of several deep CNN's architectures, from the shallowest to the deepest. All features and fixations are then compared with two types of measures, global and local, which unveil the degree of similarity of the areas of interest of the three families. From the experiments carried out on ETD human fixations database, it turns out that human fixations are positively correlated with handcrafted features and even more with deep layers of CNNs and that handcrafted features highly correlate between themselves as some CNNs do.

Keywords CNN features · Human fixations · Local image descriptors

1 Introduction

Computer vision researchers have long tried to emulate the biology of primate vision. Visual recognition methods based on features such as the widely adopted Scale Invariant Feature Transform (SIFT) [28], are inspired by computations that take place in the early visual cortex, while early convolutional neural networks such as HMAX of [30] mimic the simple and complex cell hierarchy first described in the seminal work of Hubel and Wiesel [17]. Convolution neural networks (CNNs), adding further steps that likely occur in human vision such as nonlinearity, and being trained on millions of images, are currently employed by neuroscientists to produce plausible computational models not only of lower but also of higher visual cortical areas [39].

In order to understand the functioning of CNNs, much effort is being made to dissect them, mainly by visualizing or labelling the features learned at the hidden layers. This has been achieved by looking for input patterns that maximize the activation of hidden units [12, 42], or by trying to identify salient image features through back-propagation [25, 32]. As a result it has emerged that salient regions extracted from the top layers of CNNs tend to have semantic meaning, i.e. they correspond to objects or subjects relevant to humans [13, 42, 44] and similarly, from a neuroscience perspective, that top layers of hierarchical neural networks are highly predictive of neural responses in the higher visual cortex [40]. Recently, some authors

✉ Marinella Cadoni
maricadoni@uniss.it

[1] Computer Vision Laboratory, University of Sassari, Sassari, Italy

[2] Department of Informatics, University of Sciences and Technologies Mohamed-Boudiaf, Oran, Algeria

have been looking at ways to improve attention maps generated by CNNs: by direct guidance on the attention maps generated by a weakly supervised learning deep neural network [23] or by attribute based textual explanation [38] and, inspired by the human visual system, by task-specific top-down signals together with visual stimuli [10].

Increasing knowledge on the primate visual cortex system has, on the other side, led to several saliency models that try to predict where humans look in a scene [4, 15, 18] and, more recently, the application of CNNs

to the definition of saliency models improved prediction performance [22]. It seems that the gap between computational visual recognition methods and the primate visual system is narrowing. In order to determine the extent of their similarity we think some of the fundamental questions we need to answer are whether deep convolutional neural networks actually "look" where humans look in an image and if they are getting any closer to where humans look with respect to earlier biologically inspired models. In this paper, we address these questions by proposing a methodology to establish the similarity between human fixations and the features used (or learned) by biologically inspired computational visual recognition methods. These questions have been partially addressed in [8], where human fixations have been compared to some relevant points of two CNNs, and earlier in [11] where SIFT, SURF [1] and the Harris Corner Detector (HCD) [33] have been compared to human fixations. Our proposed method draws inspiration from both works and it improves them in several ways: with a reliable definition of the regions of interest starting from human fixations or interest points, with the definition of two comparison protocols, one global and one local, implemented to establish regions similarity. Furthermore, the approach is applied to a large variety of handcrafted features and CNNs. To evaluate feature similarity, we run several experiments on the MIT eye tracking dataset (ETD) [20] comparing both human fixations to a variety of CNNs and handcrafted interest points, and also comparing intraclass and interclass points from handcrafted interest points and CNNs. A thorough statistical analysis shows that there is a positive correlation between human fixations and handcrafted features, overturning results in [11]. The results show that highest correlations occur between intraclass features and that human attention regions tend to be contained in the regions of interest defined by most features.

2 Related work

Prior to the introduction of CNNs, in [11] the authors investigated the correlation between human fixations and interest points extracted with SIFT, SURF and HCD, concluding that the similarity is not much different to that obtained with randomly generated interest points, with the exclusion of SURF points, although no tests to determine whether the difference in similarity was statistically significant were conducted.

After the development of high-performance CNNs, much effort has been dedicated to visualize their behaviour at individual unit levels. Erhan et al. [12] were the first to look for input image patterns that maximize the activation of hidden units. Zeiler et al. in [42] introduced a visualization technique that reveals which image input stimuli excite individual feature maps. They do it by mapping back feature maps activity from hidden layers back to the original input image simply by reversing the process from the considered layer to the original image (deconvnet). In the proposed method, we also reverse the process to extract our features, but since we are interested in determining the points coordinates on the input image rather than passing the feature maps through a deconvnet layer to visualize the features, what we do is simply to back-propagate the maximum response points coordinates to the original image. Yosinski et al. [41] released an interactive software to visualize the activations produced in each layer of a DNN when an image is processed. Both studies by Zeiler et al. [42] and Yosinski et al. [41] reveal the hierarchical nature of the features: shallow layers respond to corners, edges or colours or a combination of them, intermediate layers tend to be class specific so, for instance, they respond to animal faces or legs, while deepest layers respond to entire objects, e.g. a dog. A further step to interpret DNNs representation was made in [43], where the alignment between individual units and visual semantic concepts is evaluated. Again, it is confirmed that the deeper the layer, the higher the capacity to represent concepts of high semantic complexity, such as entire objects or scene parts.

Attention modules in vision were also applied to enhance the interpretability of neural networks [37]. For example, Chen et al. [10] utilised a human saliency dataset to boost their network's performance, while the recent Transformer [36] architecture relies on attention to achieve the state-of-the-art results. While providing good performance, attention allows the network to dynamically assign relative importance to the features, allowing greater transparency. However, none of these works involving attention conduct qualitative nor quantitative evaluations against human fixations, which is the main focus of this work.

Mopuri et al. [26] proposed a method for conducting evidence tracing from the prediction layer to the image in order to identify discriminative pixel locations. While their method also produces a set of fixation points similar to our work, they do not evaluate the similarity between the

obtained CNN points and human fixation points. Instead, their work focuses mainly on weakly supervised object localization and caption grounding.

Several other network dissection works focus on the quality and accuracy of visualizations rather than similarity with human saliency. In [29], although there is evaluation against human attention for the task of Visual Question Answering (VQA), the emphasis is again on visualization quality and, importantly, there is not a comparison of different CNN architectures on the basis of similarity with human saliency.

3 Interest points extraction

In this section, we describe the interest points we use to define visual attention regions. While human fixations can be acquired by eye tracking devices and handcrafted feature points are determined by the feature extraction algorithms, the process of extracting interest points from CNNs needs to be defined.

3.1 Human fixations

Human fixations are defined to be the image points where eye gaze is stable, or its speed is below a set threshold (see [27]). The human fixations we use in this paper were collected at the Massachusetts Institute of Technology as part of a project that focused on visual attention [20]. In the dataset, called the MIT eye tracking dataset (ETD), the fixations from 15 users asked to free view 1003 images randomly selected from Flickr were collected. Images were shown in succession to each viewer, each image was shown on screen for 3 s, with a grey screen interval of one second between consecutive images. Among the available human fixation datasets, the MIT-ETD is the best suited for our purposes for the high number of images and the fact that they are randomly picked, have different resolution, orientation and content (779 are landscape and 228 are portrait). For each image I in the dataset we consider the set of cumulative fixations of all 15 users, and we name it $HFix(I)$.

3.2 Handcrafted features

SIFT, SURF and the Harris corner detector (HCD) have been extensively used to perform a variety of tasks, from face recognition [5] to object recognition [7] and object tracking [45]. SIFT and SURF both derive from the Hessian of the Scale Space representation of the image, but, as shown in [6], the SIFT descriptor tends to locate points around edges, while the SURF one around corners, so, since they do not look exactly at the same image points it is interesting analysing both of them to evaluate the extent of their correlation. Interest points for the three descriptors are extracted from all images in the dataset according to the original implementations in [1, 28, 33], yielding, for a given image I the interest points sets SIFT(I), SURF(I) and HCD(I). Although the cardinality of these sets is often higher than that of human fixations, there is no a priori criterion for selecting highly significant interest points among SIFT, SURF and HCD, so all the extracted points have been considered.

3.3 CNN's features

We consider 7 pretrained deep convolutional neural networks from 6 different families: AlexNet [21], VGG-19 [31] and VGG-F [9], InceptionV3 [34], ResNetV2-50 [14], DenseNet-201 [16], and EfficientNet-b7 [35]. All networks were pretrained on more than a million images from the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) 2012 classification dataset with 1,000 object categories. To locate image points that are significant to a deep neural network, we look for the points that have maximum filter response and map them back onto the original image. We do not resize or crop the images we feed into the networks. This allows to maintain sufficient points positions accuracy, which is critical since accuracy is also affected by the reduction of the images as they are processed by the networks, something that intrinsically causes a degree of localization uncertainty when the points coordinates are mapped back to the original image. Since we are interested in selecting the points at convolution stages, before the fully connected layers occur, feeding images of different sizes is not an issue, nor it is in terms of points significance, since the networks are trained on images of objects captured at different scales and from different view points.

As an example, Fig. 1 depicts the general VGG-19 architecture with the feature maps from which the interest points are selected. As the image shows, we extract interest points from the feature maps obtained after the first five convolutions that precede the max pooling steps; in particular, in Fig. 1 the first set of feature maps is obtained at the level relu1_1 after the first bunch of convolutions. There are 64 such maps resulting from the convolution conv1_2 with 64 filters. From each of these maps, we extract the point that has the maximum filter response, obtaining 64 maximum response points. The points coordinates are mapped back through the inverse scaling function, leading to 64 interest points on the original image.

The points of highest response to each filter will certainly be included in the subsequent max pooling step and will affect the result of the following convolution steps, in

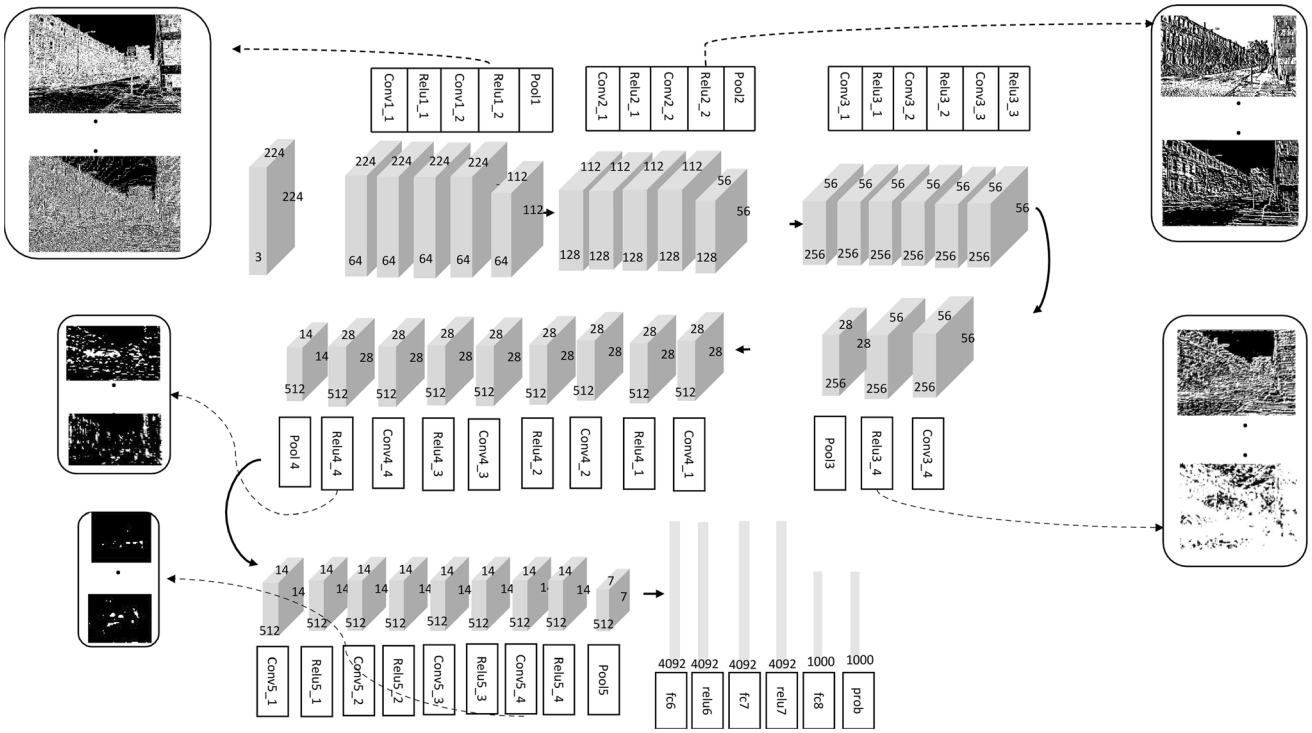

Fig. 1 Scheme of features extraction from VGG-19

other words, they are highly significant for the network. While it would be perfectly reasonable to extract more than one point from each feature map, we choose to select the global maximum since the eye fixations we compare them with were collected from viewers that were shown each image for 3s, a short time in which the human eye can scan only a subset of the areas it might find attractive.

We repeat this process for the stages relu1_2, relu2_2, relu3_4, relu4_4 and relu5_4 in Fig. 1, obtaining 5 sets of 64, 128, 256, 512 and 512 points, respectively, which will be denoted by VGG-19$_{Ci}$, for $i = 1, \ldots, 5$. The set that contains all sets of interest points VGG-19$_{Ci}$, for $i = 1, \ldots, 5$ will be denoted by VGG-19. The same interest points extraction is carried forward for the other networks; in particular, the points are extracted at the end of each block of convolutional layers.

The feature maps of the first layers are quite close to the original input image, and interest points extracted from them can be mapped back to the original image through the inverse of the image scaling function that results from the convolution, relu or max pooling steps. In fact, in [25], by inverting the network representation obtained in the first few layers, an image that is a slightly fuzzier but otherwise a visually faithful representation of the original image is obtained.

Figure 2 contains an image from the ETD with the interest points just defined. Notice that human fixations concentrate on the face, the trophy, the microphone and the hand of the woman, the same areas are mostly targeted by the seven CNNs. SIFT and SURF detect the same areas, but they also capture the pattern of the background, while, as expected, the HCD detects corners and edges. CNNs features in the image correspond to the last but one layer, and, as expected, they are mostly located on areas of semantic meaning, as human fixations are.

4 Interest regions modelling

In this section, we define the region of interest corresponding to a set of interest points. We want to determine which are the areas in the image that highly likely contain human fixations, CNN's and handcrafted features. We use the same methodology adopted in [8], based on a non-parametric estimate of the density of a given set of interest points, which offers more flexibility in modelling our points distributions over parametric methods. In [11], a nonparametric estimate is also adopted, specifically a kernel density estimation (KDE) method with a radial basis function kernel. However, these kind of kernels lack local adaptivity which in our case can lead to spurious bumps; moreover, as discussed in [2], classical bandwidth estimators rely on the "plug-in" method [19] which requires that the data are approximately normal, an assumption that again is not satisfied by our data. To overcome these problems and prevent inaccurate comparisons, to model the

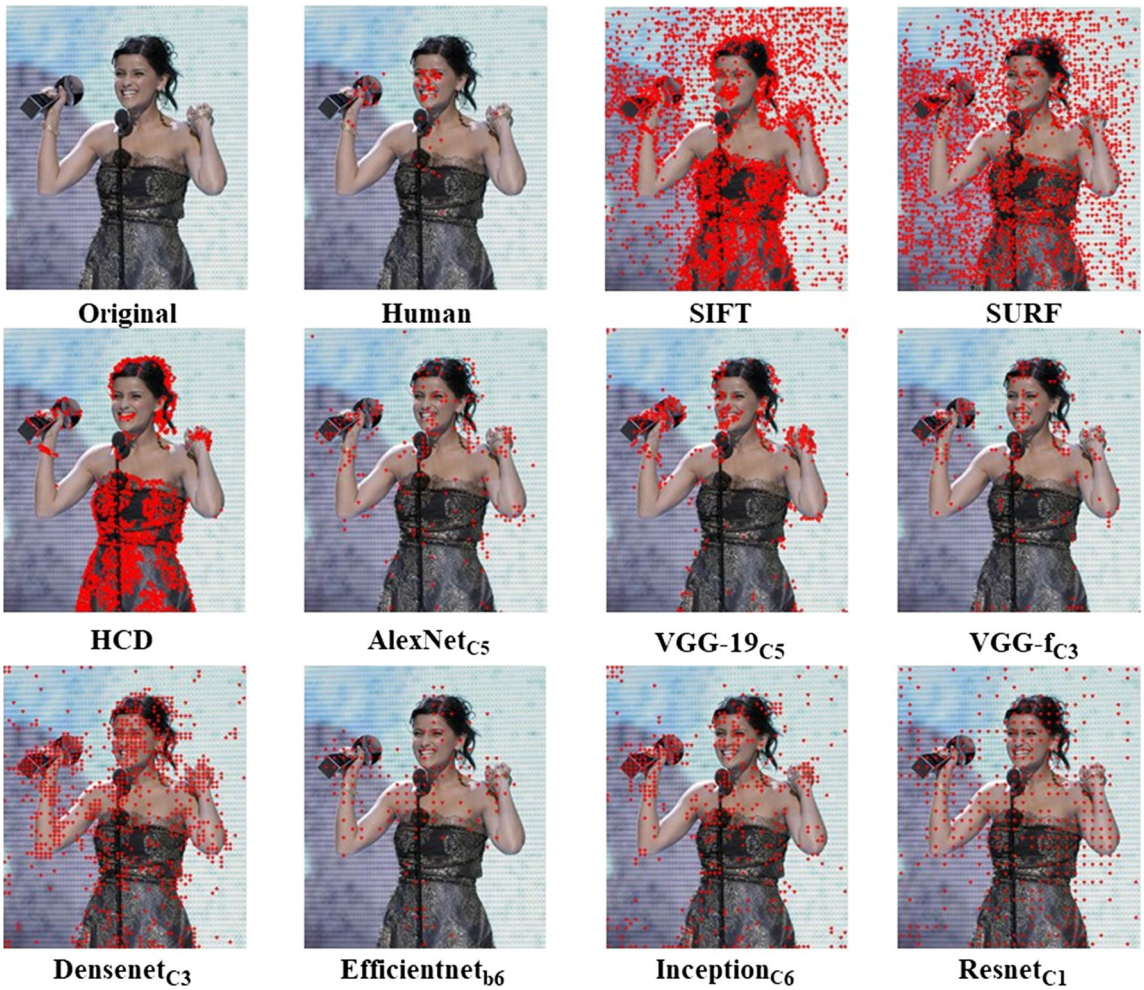

Fig. 2 Interest points. CNN's interest points are from some of the deepest layers: Alex$_{C5}$, VGG-19$_{C5}$, VGG-F$_{C3}$, Densnet$_{C3}$, EfficientNet$_{b6}$, Inception$_{C6}$ Resnet$_{C1}$

density of the sets of points we use the KDE based on a linear diffusion model developed in [2], which has also the advantage of using a bandwidth selector that does not assume data normality. In Fig. 3, we can see some human fixation points (red dots) over the central part of an image (figure on top) with the level sets of the density surface generated via diffusion. On the bottom row, the KDE via diffusion (on the left) and that via a Gaussian kernel (right) of the whole image are shown. As it can be seen, KDE estimation via Gaussian kernel leads to a bumpy surface in the central area, where there is a high density of points, but these bumps are clearly spurious, since they do not correspond to any significant clusters of points (see the red points in the central yellow area in the image on top). This high-density area would be much better modelled by a unique peak, which characterises the surface generated by the KDE via diffusion shown in the top image, represented by level curves, and in the corresponding highest peak in the bottom right image.

Given an image I and a set of features or fixations F_I, the density surface associated to F_I, and evaluated at an image point $\mathbf{x} = (x, y)$ will be called $f_{F_I}(\mathbf{x})$, so, for instance, the density of a set of human fixations for image I will be denoted by $f_{\mathrm{HFix}(I)}(\mathbf{x})$. In Fig. 4, we can see an image from the MIT fixations dataset, with human fixations on the top left and AlexNet interest points on the top right. On the bottom, we can see the estimated densities for the human fixations (left) and for the AlexNet points (right). As it can be seen, the automated bandwidth selection is able to accurately model both densities: one arising from spread points and one from clustered points.

5 Distributions comparison

To compare the obtained points densities, we need to take into account several aspects. The first one is the difference between the cardinality of the feature sets we want to

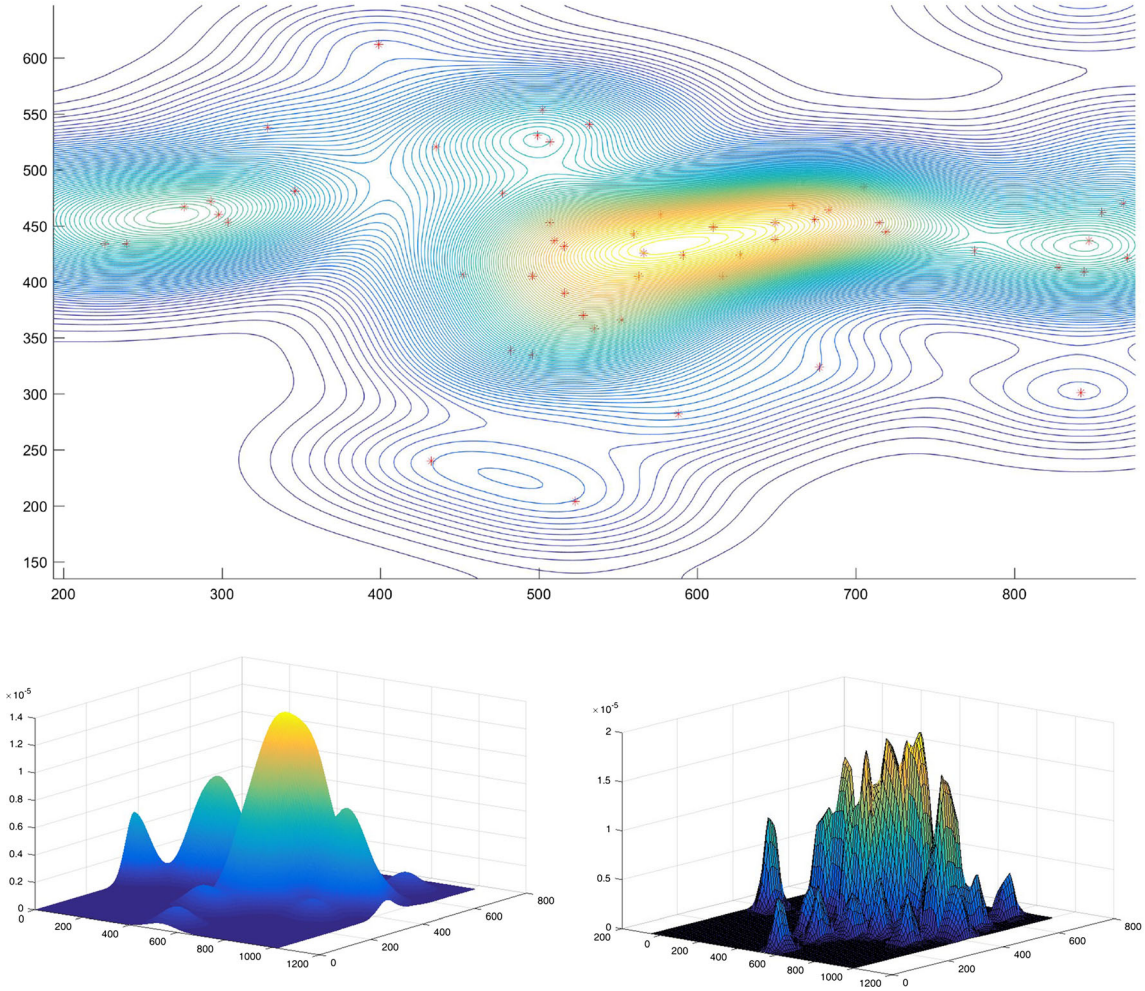

Fig. 3 KDE estimation. On top, the points with the diffusion method surface represented by level curves (warmer colours indicate high density). On the bottom, on the left the KDE via diffusion and on the right the KDE via radial basis functions (Gaussian kernel) (color figure online)

compare, which can vary from an average of 66 in the case of human fixations, to an average of 1696 in case of HCD. The second aspect is the variability of image content, which, for the sake of generality, is not supposed to be restricted to any particular class. The third aspect is that the available human fixations dataset results from exposing each image to the users for 3 s, so users attention tend to concentrate, when present, on the areas that are most significant to humans, such as faces, bodies and texts. Other potentially interesting image areas are not reached by the users because of the time limit, whereas they can be explored by feature extractors or CNNs. On the basis of these premises, to compare points sets we propose three (global) indexes of the difference between any two density distributions and a the local index defined in [8] that can reveal if the image areas targeted by one set of points are a subset of the areas target by the other. The first global index we use is the Bray–Curtis similarity [3], which is widely used in ecology to quantify the similarity between two sample populations and is well suited to assess the global similarity of two points distributions. The same index is also used in [11] which allows us to compare the results we obtain on the similarity of human fixation and handcrafted features with the ones in [11]. The second global index we use is the Jensen–Shannon divergence [24], and the third is the universally known Spearman rank correlation coefficient ρ.

Given an image I and two densities f_1 and f_2, the Bray–Curtis similarity index is defined as $BC_{1,2} = 1 - \frac{\sum_{i=1}^{n} |f_1(\mathbf{x}_i) - f_2(\mathbf{x}_i)|}{\sum_{i=1}^{n} f_1(\mathbf{x}_i) + f_2(\mathbf{x}_i)}$, for each image pixel $\mathbf{x}_i$. The Jensen–Shannon divergence is defined as

$$JSD_{1,2} = \sum_{\mathbf{x}_i} (f_1(\mathbf{x}_i) - f_2(\mathbf{x}_i)) \log \frac{f_1(\mathbf{x}_i)}{f_2(\mathbf{x}_i)}$$

To be able to directly compare the two indexes, we turn the Jensen–Shannon divergence into a similarity measure by defining $JS_{1,2} = 1 - JSD_{1,2}$.

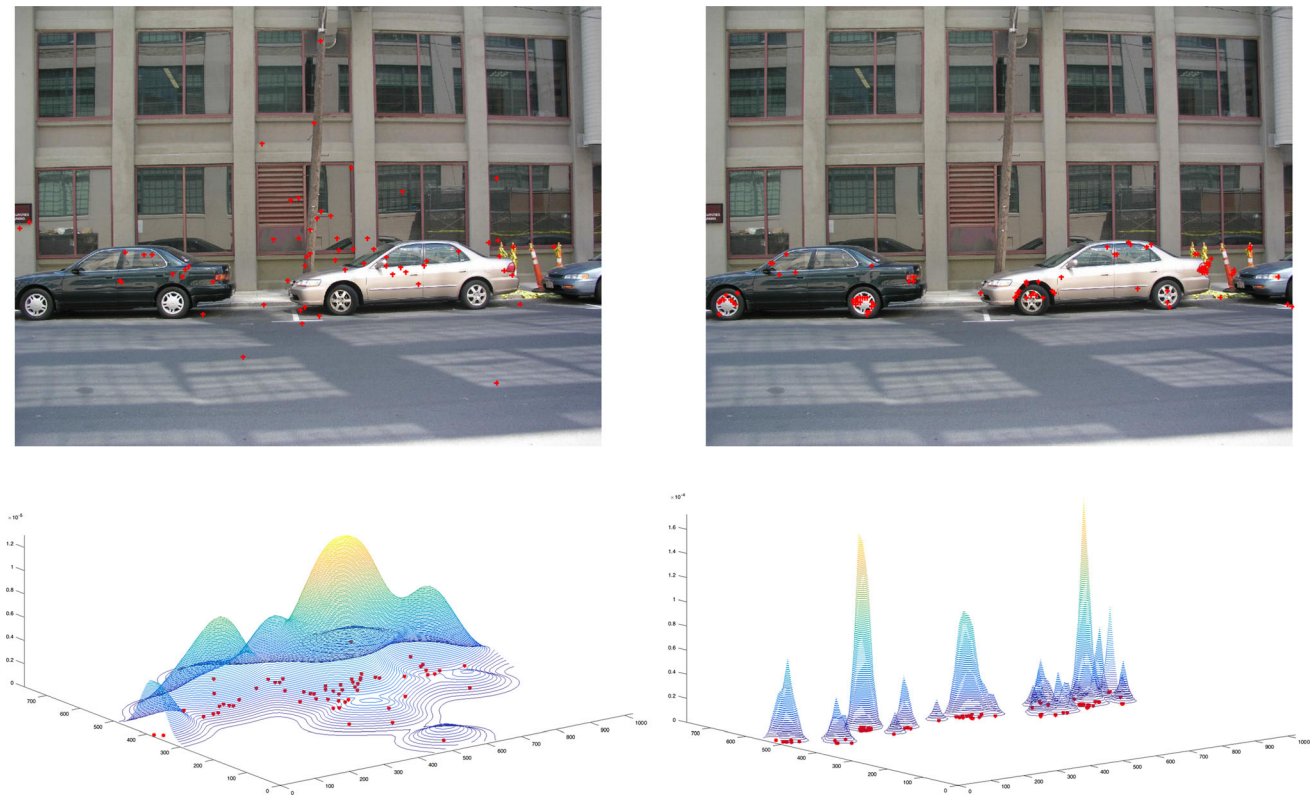

Fig. 4 KDE estimation. On top, the human fixation points (left) and the AlexNet interest points (right). On the bottom, the respective densities estimated via diffusion

While the global indexes will quantify how similar the two densities are globally, they cannot determine whether one of the two sets of points is contained in the density distribution of the other set of points. Establishing this local similarity is particularly meaningful in the cases in which the points concentrate on small regions of the image, which is what happens for human fixations of the ETD dataset, where each image was shown to each person for only 3 s.

To compare the densities at a local level, we adopt the measure defined in [8] in terms of a "two-way" binary classification.

Let F_1, F_2 be two sets of interest points for an image I and f_{F_1}, f_{F_2} their respective densities. Let us assume that all points in F_1 are F_2 points (positive) and all remaining image pixels are not (negative). Given a $\mathbf{x} \in F_1$ we can ask what the probability of $\mathbf{x}$ to be a point of type F_2 is by evaluating $f_{F_2}(\mathbf{x})$. By setting a threshold τ, for all $i = 1, \ldots, |F_1|$ we can classify an interest point $\mathbf{x}_i \in F_1$ as an F_2 interest point (a true positive) if $f_{F_2}(\mathbf{x}_i) \geq \tau$. All points in F_1 that do not satisfy the condition will be false negative. All pixels $\mathbf{w} \in I \setminus F_1$ in the image that are not F_1 points for which $f_{F_2}(\mathbf{w}) \geq \tau$ will be false positive, while those for which $f_{F_2}(\mathbf{w}) < \tau$ will be true negative. We can determine the true and false positive rates for a given threshold τ and, by varying τ, build a ROC curve. The area under the ROC curve $AUC(F_1, f_{F_2})$ is the probability that the classifier will rank a (randomly selected) F_1 point higher than a (randomly selected) $I \setminus F_1$ pixel, so the ability of the classifier to correctly classify the F_1 points tells to which extent the set F_1 is contained in the density f_{F_2}, or equivalently, to which extent the interest points F_1 are a "subset" of F_2. By switching the two sets of points, we can evaluate the probability $f_{F_1}(\mathbf{x})$ a point $\mathbf{x} \in F_2$ has to be a point of type F_1, and so we can evaluate if the points in F_2 are contained in f_{F_1} with the index $AUC(f_{F_1}, F_2)$. The two indexes describe how the two densities intersect. Notice that $AUC(F_1, f_{F_2}) = AUC(f_{F_1}, F_2) = 1$ can never be achieved since even in high density f_1 areas there will always be some pixels that are not F_2 points. Nevertheless, high values of $AUC(F_1, f_{F_2})$ mean that a high number of F_1 points are contained in f_{F_2}, and to see if the reverse is true, we need to look at $AUC(f_{F_1}, F_2)$: if it is smaller (bigger), it means that $F_1 \setminus F_2$ has a smaller (bigger) area than $F_2 \setminus F_1$, if they have similar values the areas covered by $F_1 \setminus F_2$ and $F_2 \setminus F_1$ are similar. The meaning of the two indexes is schematised in terms of density areas intersection in Fig. 5.

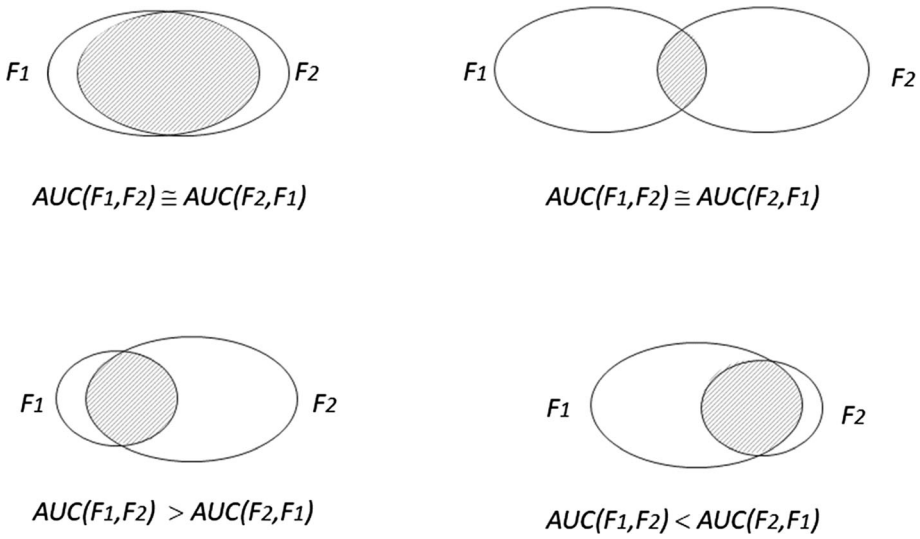

Fig. 5 Similarity indexes relationship in terms of areas of interest intersection

6 Experiments and results

We conducted a series of experiments to compare human fixations with CNN's and handcrafted features and, more generally, to evaluate the intraclass and interclass similarity of the handcrafted and CNN's families. We use the 1003 images from the ETD, the Eye tracking database at MIT [20]. As outlined in Sect. 3.1, for each image $I \in$ ETD, the database provides a set of point coordinates that result from the union of the fixations of 15 users, which for image I, according to the notation in 3.1 we call $HFix(I)$. As illustrated in Sect. 3, for each image $I \in$ ETD, we extract the features from each CNN: for instance, VGG-19_{Ci} is referred to the interest points of VGG-19 extracted at the layer Ci. We then extract the interest points $SIFT(I)$, $SURF(I)$ and $HCD(I)$. For each of the previous sets of interest points, we estimate a probability density function as described in Sect. 4, which we will use to establish points similarity.

6.1 Correlation between human fixations and interest points

We first compare human fixations to interest points using the global similarity indexes. Let us denote by $F_j(I)$ any of the feature set F over image I, and by $f_{F_j}(I)$ its density. For each image $I \in$ ETD we compare f_{F_j} with $f_{HFix}(I)$ using the global indexes defined in Sect. 5. This results in three similarity scores: the Bray–Curtis $BC_{f_{F_j} f_{HFix}}(I)$ and Jensen–Shannon $JS_{1,2}$similarities and the Spearman rank correlation coefficient ρ between the set $F_j(I)$ and $HFix(I)$ over image I. If we average the similarities scores as I ranges in the dataset ETD, we obtain the similarity between the set of features F and the human fixations $HFix$. The average scores are reported in Table 1.

As a baseline experiment, human fixations densities of each of the 1003 images in the ETD database are compared to 100 random densities and the obtained similarity indexes are averaged. Average random scores for the similarity indexes BC and JS are reported in the first line of Table 1. To see if the differences between the average similarity indexes scores and the random scores are statistically significant, a two-tailed Wilcoxon rank-sum test was run for each comparison in the table rows. The resulting p values were well under the set threshold of 0.05 for all experiments but the ones reported in bold in the table.

The three similarity indexes are coherent most of the times, with the Spearman rank correlation coefficients in between the Bray–Curtis and the Jensen–Shannon indexes. We can see, in particular, that among handcrafted features, SURF has the highest similarity with human fixations. Among CNNs, we can see how the most shallow layers poorly correlate with human fixations, while the middle to deep layers show the highest correlation. Deep layers of EfficientNet have the highest correlation of all interest points, followed by Resnet, VGG-19, DenseNet and AlexNet. Inception has a somehow peculiar behaviour, with low correlation at the shallow layers which further decreases as the layers deepen(and even becomes negative). This might be due to the different architecture of the network which is based on convoluting the image with filters of multiple sizes at the same layer. The extracted features thus follow a different pattern than that of the other networks which start from low level features to follow with high level ones.

It is interesting to compare the results we obtained relatively to the handcrafted points with the ones in [11], where no significant differences between features/fixations

Table 1 Similarity scores between interest points and human fixations

Feature type	Layer	$BC_{f_{F_j} f_H}$ (%) $R = 21.26\%$	$JS_{f_{F_j} f_H}$ (%) $R = 34.11\%$	$\rho_{f_{F_j} f_H}$ (%)
SIFT		32.78	50.32	33.89
SURF		34.92	53.23	24.25
HCD		30.58	46.21	33.88
AlexNet	*C*1	25.51	38.95	31.36
	*C*2	31.61	48.59	29.66
	*C*3	34.45	53.05	37.19
	*C*4	34.47	52.94	36.38
	*C*5	34.09	52.49	35.39
VGG-19	*C*1	14.03	22.99	14.47
	*C*2	20.51	32.12	21.94
	*C*3	20.24	47.91	20.34
	*C*4	33.46	50.05	33.91
	*C*5	35.18	53.44	37.16
VGG-f	*C*1	34.11	51.77	41.02
	*C*2	31.25	48.30	28.40
	*C*3	34.31	52.42	33.94
ResnetV2-50	*C*1	22.18	**34.27**	25.20
	*b*1	32.51	49.42	30.24
	*b*2	35.33	54.24	41.12
	*b*3	29.42	47.09	36.81
	*b*4	29.09	46.71	31.75
InceptionV3	*c*1	27.03	41.27	32.96
	*c*2	26.51	40.40	28.68
	*c*3	**21.87**	**33.38**	21.34
	*c*4	**21.87**	**33.53**	19.49
	*c*5	**21.88**	**34.93**	12.99
	*c*6	27.54	44.63	5.53
	*c*7	25.62	42.35	-6.14
DenseNet-201	*C*1	**21.66**	**33.53**	24.25
	*C*2	29.82	45.24	27.28
	*C*3	34.65	52.82	32.24
	*C*4	32.36	50.92	30.16
	*C*5	29.02	46.60	32.20
EfficientNet-b7	*b*1	28.27	42.95	36.69
	*b*2	34.85	52.64	40.12
	*b*3	32.42	50.17	32.59
	*b*4	31.21	49.01	28.63
	*b*5	33.37	51.42	31.15
	*b*6	37.71	56.98	52.33
	*b*7	37.37	56.36	44.92

Third column: BC similarity scores, fourth: Jensen–Shannon similarity scores, fifth: Spearman correlation coefficients. Bold percentages refer to scores for which the two-sided Wilcoxon test returned a p value > 0.05, so the difference between similarity with human fixations and random is not statistically significant

similarities and random points/fixations similarities were found. In particular, limiting the comparison to the BC index (which is the only similarity measure used the work), we can see higher similarity scores between the distributions of random points and human fixations, coupled with lower scores between the distributions of SIFT, SURF and HCD and distributions of human fixations. This might be due to the different kernel density estimation techniques: an accurate bandwidth selection of the kernel is essential for robust density estimation, and possibly to a different methodology, which is not detailed, to estimate random densities. Contrarily to the conclusions in [11], we can affirm that global similarities between human fixations and SIFT, SURF and HCD, although not high, are statistically significant.

6.2 Local similarity assessment

To further explore the correlation between interest points and human fixations, we compare them on a local level using the local similarity AUC indexes defined in Sect. 5. Given an image I from the ETD dataset, the sets $F_i(I)$ and $HFix(I)$ of interest points and human fixations and the two densities $f_{F_i(I)}$ and $f_{HFix(I)}$, we determine the ROC curve deriving from the classification of the $F_i(I)$ points with the human fixations density $f_{Hfix}(I)$ and calculate the indicator $AUC_{F_i f_{Hfix}}(I)$ and conversely, the ROC curve and the index $AUC_{F_{Hfix} f_{F_i}}(I)$ deriving from the classification of the human fixations $HFix(I)$ with the density estimated from the interest points $F_i(I)$. We repeat the procedure for each image $I \in$ ETD, and we average over the indexes $AUC_{F_i f_{Hfix}}(I)$ and $AUC_{F_{Hfix} f_{F_i}}(I)$ to get $AUC_{F_i f_{Hfix}}$ and $AUC_{F_{Hfix} f_{F_i}}$, which reveal to which extent the areas covered by the interest points of type F_i are contained in the areas covered by human fixations and vice versa.

The results can be seen in Table 2. For all feature sets, both AUC indicators are well above the 50% value corresponding to a random classifier, showing that there is a significant intersection with human fixations. Furthermore, in most cases, the first indicator (second column of the table) is greater than the second (third column of the table). This means that, on average, the image areas target by humans tend to be contained in those occupied by the other features. Only some shallow layers of some the CNNs have featured that are contained in the attention areas of humans (AlexNet, VGG-19, DenseNet). The highest value of 79.25% is given by EfficientNet$_{b6}$, while among handcrafted features, SURF has the highest score of 73.28%. The higher values of the first indicators could be explained by the way the human fixations were collected. Indeed, the time limit of 3 s for each image exposure means that fixations concentrate on the image areas that most capture

Table 2 AUC indexes of the comparison between interest points and human fixations

Feature type	Layer	$AUC_{F_i,f_{Hfix}}$ (%)	$AUC_{F_{Hfix},f_{F_i}}$ (%)
SIFT		69.77	58.41
SURF		73.28	58.78
HCD		67.54	61.48
AlexNet	*C*1	63.49	63.96
	*C*2	69.73	58.29
	*C*3	73.89	58.69
	*C*4	73.84	58.99
	*C*5	73.35	58.71
VGG-19	*C*1	58.04	63.70
	*C*2	59.66	60.82
	*C*3	71.51	62.69
	*C*4	72.92	61.54
	*C*5	74.19	59.68
VGG-f	*C*1	71.55	61.65
	*C*2	69.49	57.67
	*C*3	73.60	59.07
ResnetV2-50	*C*1	70.37	59.65
	*b*1	76.15	58.52
	*b*2	69.81	53.57
	*b*3	69.37	53.22
	*b*4	60.04	62.70
InceptionV3	*c*1	64.44	62.93
	*c*2	63.49	61.50
	*c*3	59.01	56.85
	*c*4	58.85	56.07
	*c*5	63.66	51.08
	*c*6	66.08	52.59
	*c*7	53.25	50.35
DenseNet-201	*C*1	59.21	62.52
	*C*2	67.06	60.64
	*C*3	74.10	59.03
	*C*4	75.45	55.14
	*C*5	69.75	53.16
EfficientNet-b7	*b*1	64.79	64.07
	*b*2	71.55	62.24
	*b*3	69.55	58.25
	*b*4	68.17	56.15
	*b*5	71.12	58.00
	*b*6	79.25	61.87
	*b*7	78.20	61.94

human attention, other possible areas of interest targeted by local descriptors or CNNs might not be observed by humans simply because of lack of time.

In Figs. 6 and 7 are shown the images of the ETD that produce, respectively, the highest and lowest *AUC* scores for SIFT points (the red crosses in the first images of both figures) and human fixations (the blue crosses). For the image in Fig. 6, we have the values $AUC_{SIFT,f_{Hfix}} = 85.21\%$ and $AUC_{HFix,f_{SIFT}} = 97.19\%$, indicating that the human fixations are almost completely contained in the density of the SIFT points, as confirmed by the image on the bottom right. On the other hand, the smaller value of the $AUC_{SIFT,f_{Hfix}}$ indicates that there are areas that contain SIFT points that are not looked at by humans, as the image on the bottom left shows. Notice how human fixations tend to concentrate on the writings. In this case, the value of the global index was 60.65%, which hints that the two sets are correlated but does not shed light on how they intersect.

For the image in Fig. 7, the values of the indexes are $AUC_{SIFT,f_{Hfix}} = 37.01\%$ and $AUC_{HFix,f_{SIFT}} = 43.37\%$, which are both low values that indicate the two sets of points hardly intersect, as shown in the two images at the bottom of the figure. In this case the global BC similarity has a value of 23.45%, below the random value of about 26%.

In Figs. 8 and 9, we show the images for which the local correlation between AlexNet points and human fixations is, respectively, maximum and minimum. Local correlation indexes values for the image in 8 are $AUC_{AlexNet_{C3},f_{Hfix}} = 88.05\%$ and $AUC_{HFix,f_{AlexNet_{C3}}} = 56.81\%$ while the global BC similarity index is 38.98%. As it can be seen from the density images, the proportion of AlexNet_{C3} points that is contained in the human fixation density (left bottom image) is higher than the proportion of human fixations contained in the AlexNe_{C3} density, which is correctly reflected by the *AUC* scores.

Local correlation indexes for the image in 9 are $AUC_{AlexNet_{C3},f_{Hfix}} = 36.38\%$ and $AUC_{HFix,f_{AlexNet_{C3}}} = 34.29\%$, the minimum local correlation between $\text{AlexNe}t_{C3}$ point and human fixations in the ETD dataset. The BC similarity index has a very low value of 9.08%. The two sets of points hardly intersect, in accordance with the low *AUC* values.

6.3 Similarity assessment with CNN fixations extracted with other methods

The correlation between humans and CNNs might depend on the way CNN's interest regions are selected. While our main intent is to analyse the correlation between human fixations and CNNs at all layers, by unveiling what the filters learn stage by stage, it is also interesting to see how human fixations compare with CNN's interest regions extracted with other methods that take into account also the fully connected layers (see last block in Fig. 1). To do this, we extracted the VGG-16 fixations from all images in the ETD dataset using the technique described in [26], where discriminative pixel locations that guide the network prediction are obtained by considering feature dependencies

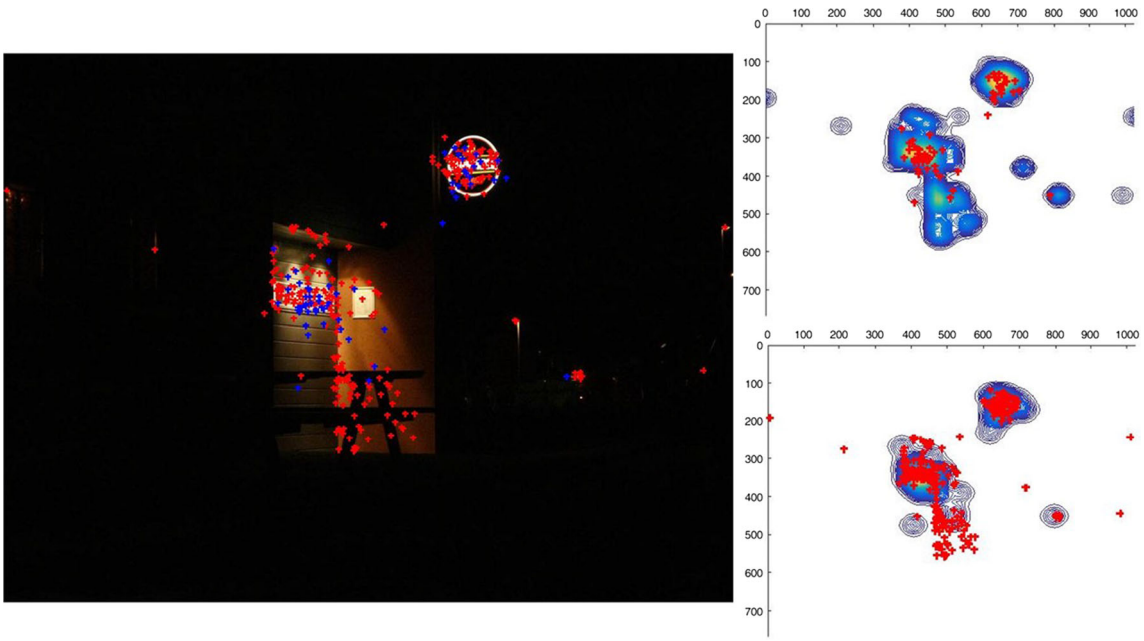

Fig. 6 SIFT-human fixations local correlation maximum. Top image: red crosses are SIFT points, blue crosses are human fixations. Bottom images: (left) planar projection of human density and sift points (red crosses), (right) SIFT density and human fixations (red crosses) (color figure online)

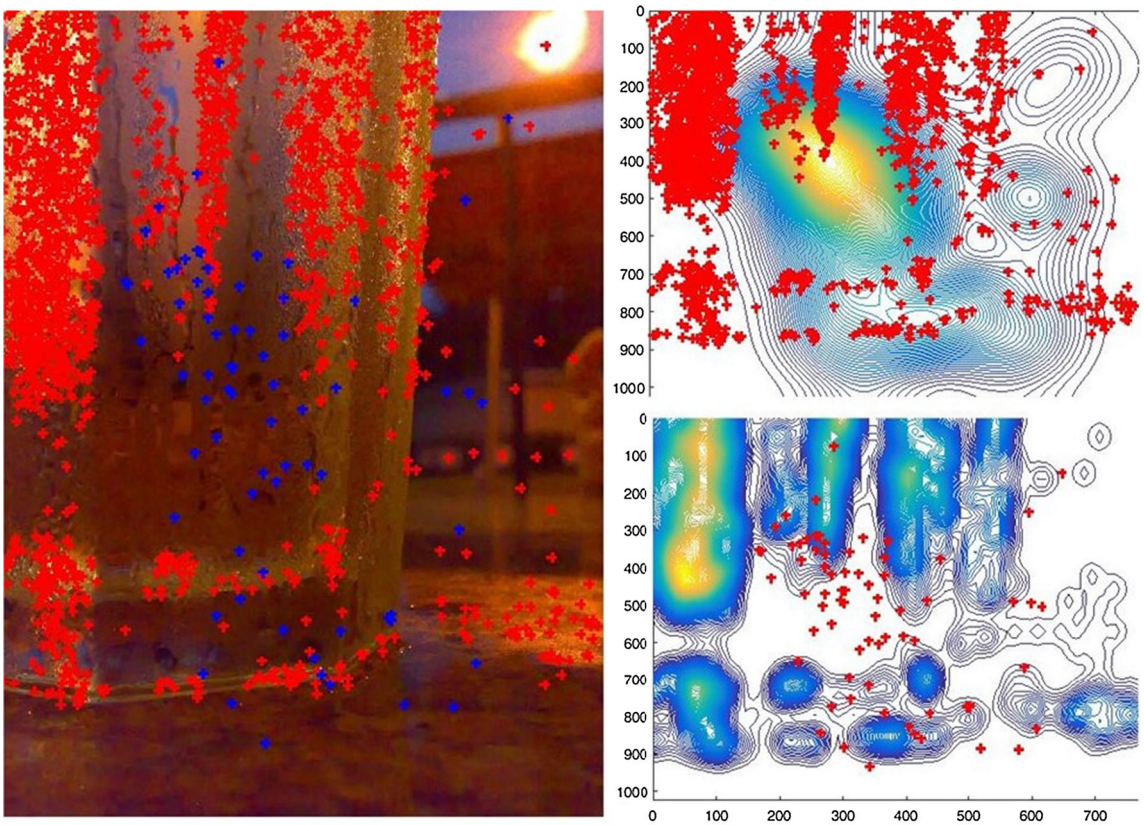

Fig. 7 SIFT-human fixations local correlation minimum. Top image: red crosses are SIFT points, blue crosses are human fixations. Bottom images: (left) planar projection of human density and sift points (red crosses), (right) SIFT density and human fixations (red crosses) (color figure online)

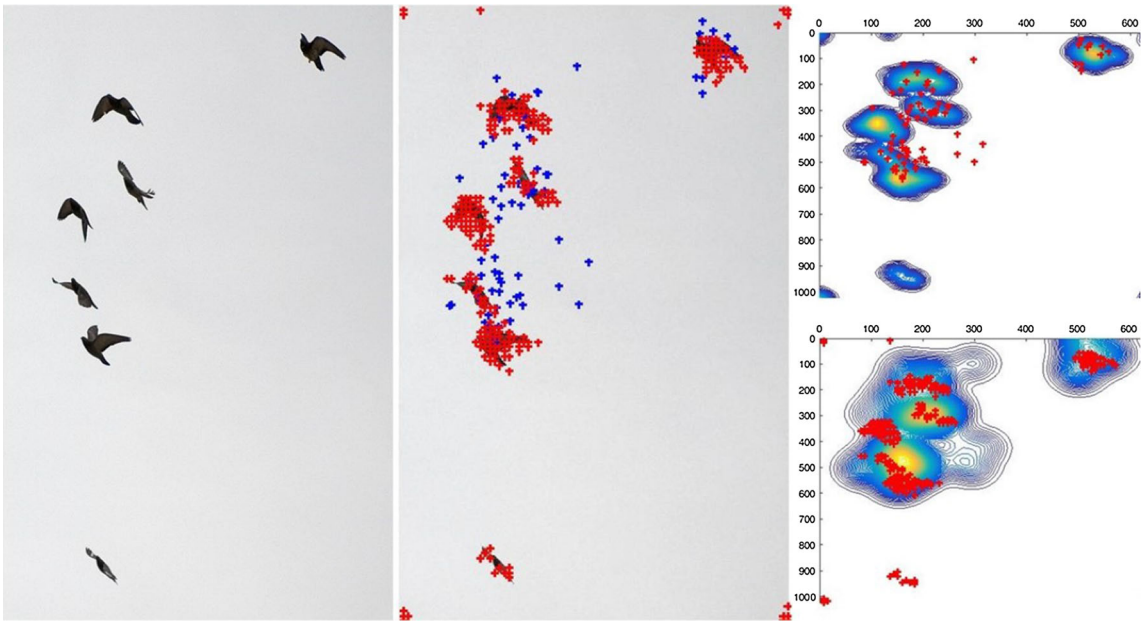

Fig. 8 AlexNet$_{C3}$-human fixations local correlation maximum. Left image: original. Middle: AlexNet$_{C3}$ points (red), human fixations (blue). Right images: (top) planar projection of AlexNet$_{C3}$ density and human fixations (red crosses), (bottom) human density and AlexNet$_{C3}$ points (red crosses) (color figure online)

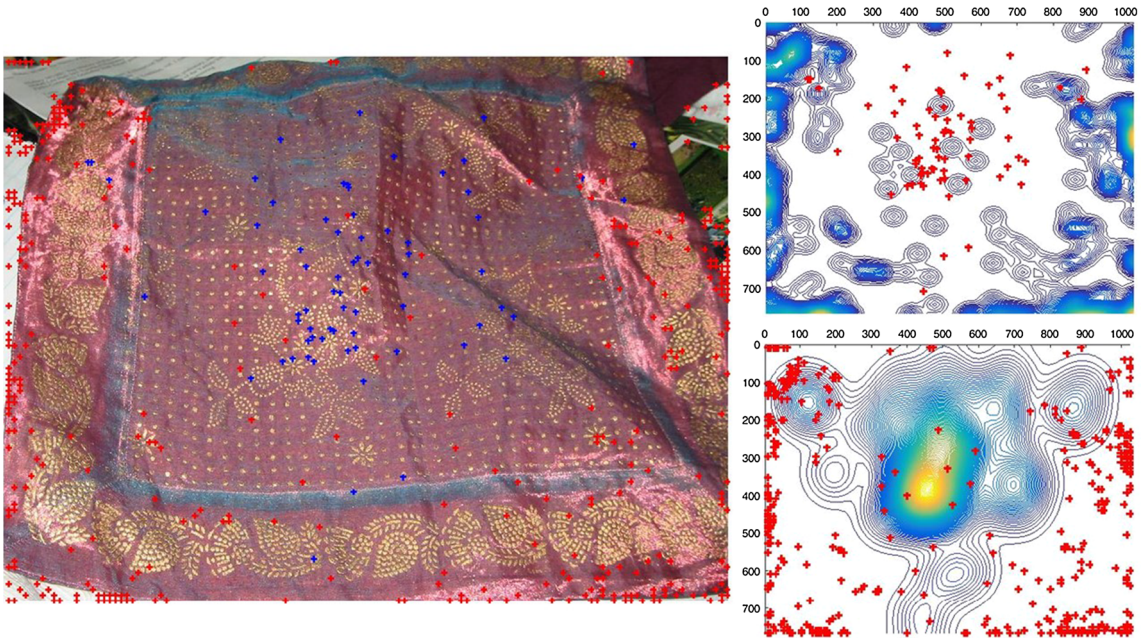

Fig. 9 AlexNet$_{C3}$-human fixations local correlation minimum. Left image: AlexNet$_{C3}$ points (red) and human fixations (blue). Right images: (top) planar projection of AlexNet$_{C3}$ density and human fixations (red crosses), (bottom) human density and AlexNet$_{C3}$ points (red crosses) (color figure online)

between pairs of consecutive layers. The resulting pixel locations, or interest image points, are fed into our KDE module to produce the regions of interest that are compared to the regions of interest generated by human fixations. In Tables 3 and 4, we can see the similarity scores (global and local, respectively) between human fixations and VGG-16 interest points extracted with our method from each layer before the max pooling step (lines 2–6), and the similarity scores between human fixations and VGG-16 fixations extracted according to [26] (line 7).

The correlation between VGG-16 fixations of [26] and human fixations is similar to the correlation between VGG-

Table 3 Similarity scores between human fixations and VGG-16 interest points extracted with our method from layers $C1, \ldots, C5$ and with CNN fixations as in [26]

Feature type	Layer	$BC_{f_{F_j} f_H}$ (%)	$JS_{f_{F_j} f_H}$ (%)	$\rho_{f_{F_j} f_H}$ (%)
VGG-16 interest points	$C1$	14.95	24.56	16.55
	$C2$	20.06	31.57	21.38
	$C3$	27.67	41.80	30.66
	$C4$	35.53	52.90	36.27
	$C5$	35.58	53.93	37.48
VGG-16 fixations [26]	–	38.71	56.68	50.12

Third column: BC similarity scores, fourth: Jensen–Shannon similarity scores, fifth: Spearman correlation coefficients

Table 4 AUC indexes of the comparison between human fixations and VGG-16 with intersect points extracted with our method from layers $C1, \ldots, C5$ and with CNN fixations as in [26]

Feature type	Layer	$AUC_{F_i f_{Hfix}}$ (%)	$AUC_{F_{Hfix} f_{F_i}}$ (%)
VGG-16 interest points	$C1$	53.94	63.50
	$C2$	57.70	64.38
	$C3$	66.28	63.96
	$C4$	74.82	62.07
	$C5$	74.13	59.69
VGG-16 fixations [26]	–	75.06	64.90

16 interest points of the last two layers and human fixations. This was somehow predictable, since the features extracted from the last layers tend to have semantic meaning. The slightly superior correlation humans fixations have with VGG-16 fixations of [26] is probably due to outlier removal.

6.4 Similarity across features types

Having established the similarity between features and human fixations, we investigate the intraclass and interclass features correlation. This question has hardly been explored in the past, even for handcrafted features. In [5, 6], there is evidence that SIFT and SURF points only partially overlap, but the experiments are limited to images of human faces. To have an idea of how the various interest points correlate, we select a representative of each CNN family, namely the set of features relative extracted from the deepest layer. For each pair of feature sets, we calculate the similarity indexes across all images of the ETD database. In Table 5, the three global similarity scores are reported for all pairs of features sets compared, arranged in decreasing order from top to bottom. Similarity scores of human fixations with the last layer of CNNs and handcrafted features are reported again for ease of comparison.

On a global level, interest points appear to be more correlated among themselves than with human fixations. Greatest correlations occur between DenseNet and Resnet ($BC = 93.91\%$, $JS = 99.47\%$, $\rho = 75.06\%$), Inception and Resnet $BC = 85.45\%, JS = 97.33\%, \rho = 41.97\%$, DenseNet Inception ($BC = 85.51\%, JS = 97.38\%, \rho = 40.86\%$), VGG-f and AlexNet ($BC = 77.28\%, JS = 92.47\%, \rho = 78.70\%$), SIFT and SURF ($BC = 77.16\%, JS = 92.72\%, \rho = 77.66\%$). We can generally see that interest points from the CNN's family highly correlate, as they do sets from the handcrafted points family, while interfamilies correlations are weaker (humans included).

Global scores can be better interpreted together with local scores. To this end, the AUC values for all pairs of feature types of Table 5 are calculated, and the results are shown in the plot, where, for two feature sets (F_i, F_j), the two local indexes are displayed as a point of coordinates $(AUC_{F_i f_{F_j}}, AUC_{F_j f_{F_i}})$. As the plot shows, human fixations scores (see Table 2) are all above the bisector, indicating that a notably higher fraction of humans fixations are contained in all other features, as discussed in Sect. 6.2. This also holds for Harris Corner points, which tend to be contained in other sets but SIFT and SURF. Looking at CNNs, it is surprising to discover that the pairs that have a high global similarity, such as DenseNet and Resnet (58.19%,58.04%), Inception and Resnet (56.14%,56.84%), DenseNet and Inception (56.56%, 55.85%) have not so high local similarity indexes, while all being near the bisector. This is due to the fact that the interest points of the deepest layers of the mentioned CNNs are generated by points that are quite spread (and not numerous) over the images, so they will generate densities that share a similar support and shape. The global indexes, based on the shape of the densities, will score high values, while the local similarity is more subtle, since the low probability values of the densities cause more false negatives and the many pixels sitting in the densities support but not belonging to the interest points sets will be false positive, thus leading to a low local similarity scores. How the features are spread on the last layers of the networks Densnet-201 and Resnet-V2-50 can be seen in Fig. 10. The small size of the feature

Table 5 Similarity scores between all pairs of interest points, in decreasing order

Feature types	$BC_{f_{F_j}f_H}$ (%)	Feature types	$JS_{f_{F_j}f_H}$ (%)	Feature types	$\rho_{f_{F_j}f_H}$ (%)
DenseNet-201 Resnet-v2-50	93.91	DenseNet-201 Resnet-v2-50	99.47	VGG-f AlexNet	78.70
DenseNet-201 InceptionV3	85.51	DenseNet-201 InceptionV3	97.38	SIFT SURF	77.66
InceptionV3 Resnet-v2-50	85.45	InceptionV3 Resnet-v2-50	97.33	DenseNet-201 Resnet-v2-50	75.06
VGG-f AlexNet	77.28	SIFT SURF	92.72	VGG-f VGG-19	71.10
SIFT SURF	77.16	VGG-f AlexNet	92.47	AlexNet VGG-19	70.55
Efficientnet-b7 Resnet-v2-50	73.73	Efficientnet-b7 Resnet-v2-50	90.76	SIFT HCD	68.27
Efficientnet-b7 DenseNet-201	73.62	Efficientnet-b7 DenseNet-201	90.66	VGG-19 Efficientnet-b7	66.22
AlexNet VGG-19	72.33	AlexNet VGG-19	89.62	SURF HCD	65.35
VGG-f VGG-19	71.52	VGG-f VGG-19	88.92	Efficientnet-b7 DenseNet-201	64.88
AlexNet Efficientnet-b7	70.76	AlexNet Efficientnet-b7	88.59	Efficientnet-b7 Resnet-v2-50	63.15
VGG-19 Efficientnet-b7	70.28	VGG-19 Efficientnet-b7	88.46	AlexNet Efficientnet-b7	62.78
Efficientnet-b7 InceptionV3	69.00	Efficientnet-b7 InceptionV3	87.79	VGG-f Efficientnet-b7	62.24
VGG-f Efficientnet-b7	67.46	VGG-f Efficientnet-b7	86.13	AlexNet DenseNet-201	60.62
AlexNet Resnet-v2-50	66.10	AlexNet Resnet-v2-50	85.15	VGG-f DenseNet-201	60.30
AlexNet DenseNet-201	65.99	AlexNet DenseNet-201	85.04	VGG-19 DenseNet-201	59.69
SURF DenseNet-201	65.90	SURF DenseNet-201	83.95	VGG-19 Resnet-v2-50	59.09
SURF Resnet-v2-50	65.32	SURF Resnet-v2-50	83.54	AlexNet Resnet-v2-50	58.47
SURF Efficientnet-b7	64.52	SURF Efficientnet-b7	83.42	VGG-f Resnet-v2-50	57.52
SURF AlexNet	63.73	SURF AlexNet	83.15	Human Efficientnet-b7	50.63
AlexNet InceptionV3	63.47	AlexNet InceptionV3	83.11	SURF AlexNet	45.72
VGG-19 Resnet-v2-50	62.73	VGG-19 Resnet-v2-50	83.09	SURF VGG-19	44.70
VGG-19 DenseNet-201	62.44	VGG-19 DenseNet-201	82.84	SIFT AlexNet	44.41
SIFT HCD	62.21	VGG-f Resnet-v2-50	81.34	Human SURF	43.46
SURF InceptionV3	61.65	VGG-f DenseNet-201	81.24	SURF Efficientnet-b7	43.24
VGG-f Resnet-v2-50	61.50	VGG-19 InceptionV3	80.86	SIFT VGG-19	43.23
VGG-f DenseNet-201	61.41	SURF VGG-19	80.74	SURF VGG-f	43.05
SURF VGG-19	60.79	SURF InceptionV3	80.70	SIFT VGG-f	43.04
SURF VGG-f	60.39	SIFT HCD	80.68	InceptionV3 Resnet-v2-50	41.97
VGG-19 InceptionV3	60.16	SURF VGG-f	80.05	DenseNet-201 InceptionV3	40.86
VGG-f InceptionV3	59.29	VGG-f InceptionV3	79.49	HCD VGG-19	40.35
SIFT AlexNet	58.69	SIFT AlexNet	78.83	HCD AlexNet	39.91
SIFT DenseNet-201	58.57	SIFT DenseNet-201	77.85	HCD VGG-f	38.17
SIFT Resnet-v2-50	57.97	SIFT Resnet-v2-50	77.35	SIFT Efficientnet-b7	37.89
SIFT Efficientnet-b7	57.33	SIFT Efficientnet-b7	77.27	Human VGG-19	37.16
SURF HCD	56.73	SIFT VGG-19	76.46	HCD Efficientnet-b7	36.39
SIFT VGG-f	56.21	SIFT VGG-f	76.38	SIFT DenseNet-201	35.44
SIFT VGG-19	56.12	SURF HCD	76.03	Human AlexNet	35.39
SIFT InceptionV3	55.06	SIFT InceptionV3	74.83	SURF DenseNet-201	35.33
HCD AlexNet	44.13	HCD AlexNet	64.08	AlexNet InceptionV3	34.13
HCD VGG-19	43.43	HCD VGG-19	63.12	Human VGG-f	33.94
HCD VGG-f	42.77	HCD VGG-f	62.33	Human SIFT	33.89
HCD Efficientnet-b7	42.49	HCD Efficientnet-b7	62.03	Human HCD	33.88
HCD DenseNet-201	40.87	HCD DenseNet-201	60.15	VGG-f InceptionV3	33.86
HCD Resnet-v2-50	40.42	HCD Resnet-v2-50	59.68	VGG-19 InceptionV3	32.52
HCD InceptionV3	38.63	HCD InceptionV3	57.59	Human DenseNet-201	32.30
Human Efficientnet-b7	37.05	Human Efficientnet-b7	56.13	Human Resnet-v2-50	31.75
Human VGG-19	35.18	Human VGG-19	53.44	SURF Resnet-v2-50	31.66
Human SURF	34.93	Human SURF	53.23	SIFT Resnet-v2-50	30.49

Table 5 (continued)

Feature types	$BC_{f_{F_j} f_H}$ (%)	Feature types	$JS_{f_{F_j} f_H}$ (%)	Feature types	$\rho_{f_{F_j} f_H}$ (%)
Human VGG-f	34.31	Human AlexNet	52.49	HCD DenseNet-201	29.43
Human AlexNet	34.10	Human VGG-f	52.42	Efficientnet-b7 InceptionV3	28.43
Human SIFT	32.82	Human SIFT	50.32	HCD Resnet-v2-50	24.91
Human HCD	30.58	Human Resnet-v2-50	46.71	SIFT InceptionV3	12.03
Human Resnet-v2-50	29.09	Human DenseNet-201	46.61	SURF InceptionV3	12.00
Human DenseNet-201	29.02	Human HCD	46.21	HCD InceptionV3	9.10
Human InceptionV3	25.62	Human InceptionV3	42.35	Human InceptionV3	-6.14

First column: BC similarity scores, second: Jensen–Shannon similarity, third: Spearman correlation coefficients

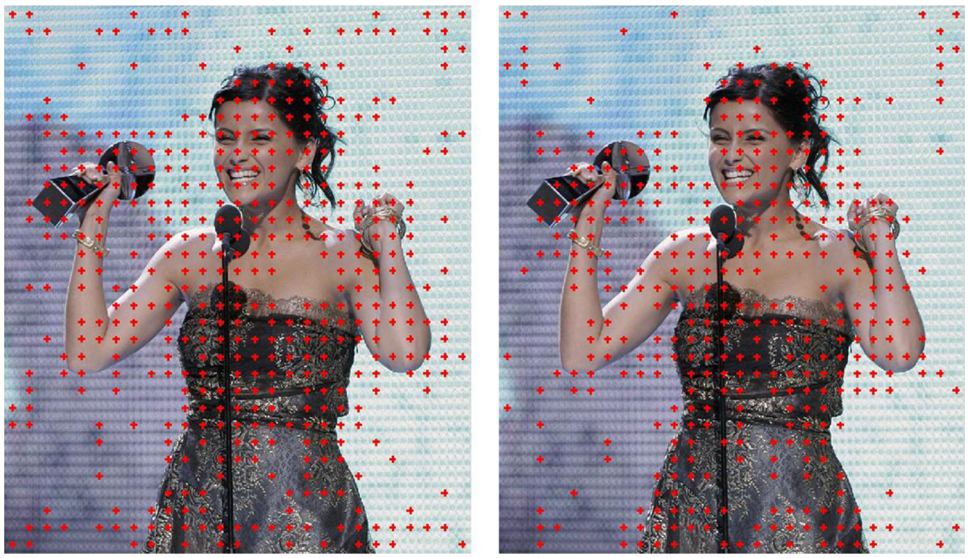

Fig. 10 (Left) DenseNet-201 layer *c*5 interest points and (right) Resnet-V2-50 layer *b*4 interest points. Notice how the points are evenly spread at these two deepest layers

maps at these two deepest layers contributes to the "grid" effect.

On the contrary, as Fig. 11 and Table 2 show, the local measure tells us that human fixations highly correlate with EfficientNet, VGG, AlexNet, SIFT and SURF, albeit being subsets of them.

Human fixations tend to be located in areas that are targeted by local descriptors and CNNs and, at the same time, there are areas targeted by local descriptors and CNNs that humans do not seem to look at.

7 Conclusion

The more we learn about the mechanisms of the primate visual system, the more these mechanisms can be embedded in computational models to improve their performance. While earlier local image descriptors were inspired by the mechanisms of the early visual cortex, today's CNNs embed processes that take place on the higher cortical areas of the primate visual system. It is then natural to ask if these computational models rely on the same image areas humans focus their attention on when performing recognition or classification tasks. In this paper, we measured the similarity between attention areas of humans, handcrafted features (we used the three local image descriptors SIFT, SURF and HCD) and seven deep CNNs from the families AlexNet, VGG, Resnet, Inception, DenseNet, EfficientNet. To do so, we used three global similarities and a local one to compare the area of interest of the three classes humans, local image descriptors and CNNs. Extensive experiments were carried out on the ETD dataset, to establish intraclass and interclass similarities. The obtained results indicate that human fixations positively correlate with SIFT, SURF and HCD. Slightly higher correlations can be seen with the deepest layers of some of the networks, notably EfficientNet, Resnet, DenseNet, VGG and AlexNet, while there is weak or no correlation with shallow layers. Only Inception learns features through the layers that do not follow this pattern: correlation with humans is always weak or negative, especially at the intermediate layers. Local comparisons highlight that humans attention areas tend to

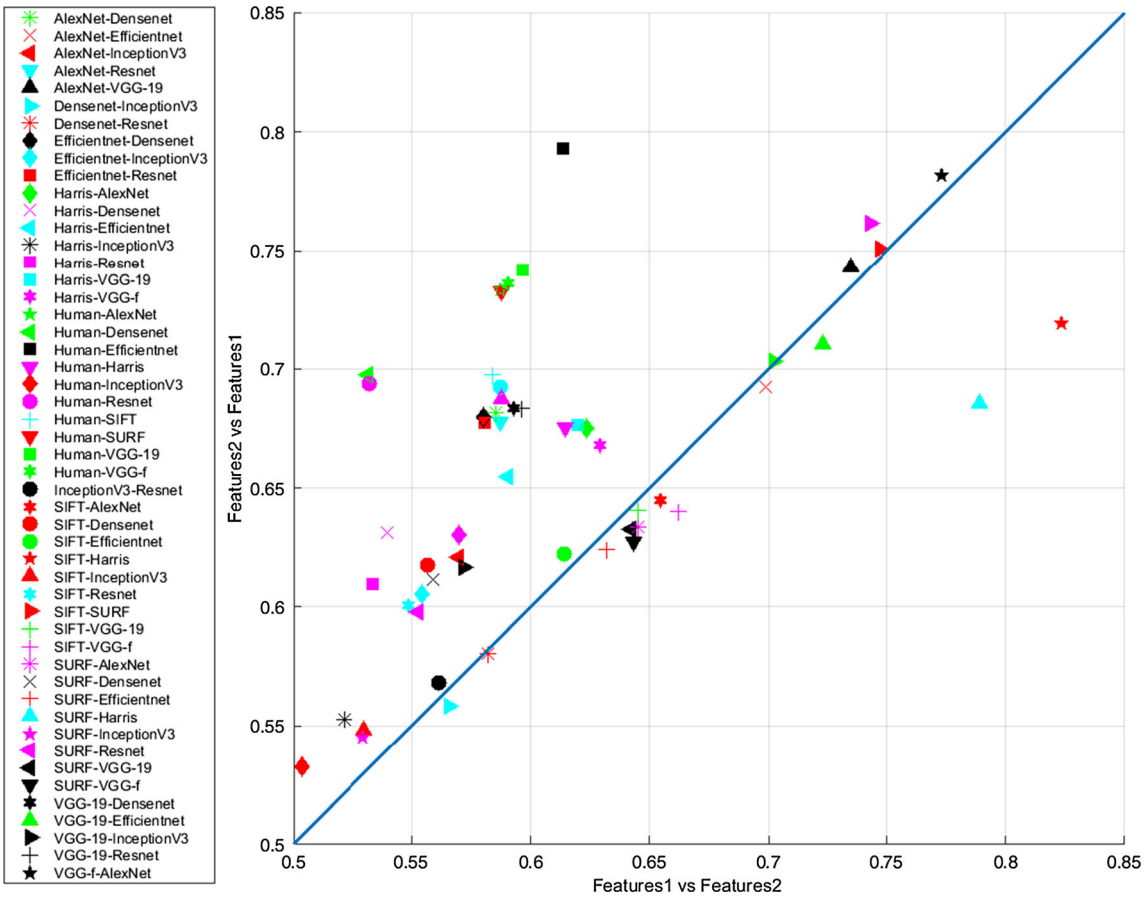

Fig. 11 Local similarity indexes between all interest points. For an ordered pair (F_i = Feature1, F_j = Feature2) of interest point sets in the legend the corresponding point in the graph will have coordinates ($AUC_{F_i f_{F_j}}$, $AUC_{F_j f_{F_i}}$). The origin of the axes corresponds to the scores of a random classifier (0.5, 0.5). Points above the bisector imply Feature1 tends to be contained in Feature2, the converse holds for points below. Points near the bisector are relative to features that have a similar fraction of common surface areas, which will be small for points near the origin and increasingly larger as points move further from it

be contained in areas relevant to local descriptors and most CNN's layers. This might be due to the fact that human fixations were collected by viewing images for only 3 s. Further investigations on how correlation changes when humans can view images for a longer time would shed light on the full extent of the correlations. Moving on to intra-class similarities, we can see how SIFT and SURF highly correlate, as most of the networks do, while interclass similarities are predictably lower, with the most relevant being the ones between humans and CNNs.

Acknowledgements This work has been supported by "Fondo di Ateneo per la ricerca 2020" of the University of Sassari.

Funding Open access funding provided by Università degli Studi di Sassari within the CRUI-CARE Agreement.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Bay H, Ess A, Tuytelaars T, Van Gool L (2008) Speeded-up robust features (surf). Comput Vis Image Underst 110(3):346–359
2. Botev ZI, Grotowski JF, Kroese DP et al (2010) Kernel density estimation via diffusion. Ann Stat 38(5):2916–2957

3. Bray JR, Curtis JT (1957) An ordination of the upland forest communities of Southern Wisconsin. Ecol Monogr 27(4):325–349. https://doi.org/10.2307/1942268
4. Bruce ND, Tsotsos JK (2009) Saliency, attention, and visual search: an information theoretic approach. J Vis 9(3):5
5. Cadoni M, Lagorio A, Grosso E (2014) Iconic methods for multimodal face recognition: a comparative study. In: 2014 22nd international conference on pattern recognition, pp 4612–4617
6. Cadoni M, Lagorio A, Grosso E (2016) Large scale face identification by combined iconic features and 3d joint invariant signatures. Image Vis Comput 52(C):42–55
7. Cadoni M, Lagorio A, Grosso E (2019) Incremental models based on features persistence for object recognition. Pattern Recognit Lett 122:38–44. https://doi.org/10.1016/j.patrec.2019.02.019
8. Cadoni M, Lagorio A, Grosso E (2020) Do cnn's features correlate with human fixations? In: Petkov N, Strisciuglio N, Travieso-González CM (eds) APPIS 2020: 3rd international conference on applications of intelligent systems, APPIS 2020, Las Palmas de Gran Canaria Spain, 7–9 January 2020, ACM, pp 13:1–13:6. https://doi.org/10.1145/3378184.3378197
9. Chatfield K, Simonyan K, Vedaldi A, Zisserman A (2014) Return of the devil in the details: delving deep into convolutional nets. In: Proceedings of the British machine vision conference (BMVC), pp 1–10
10. Chen S, Zhao Q (2018) Boosted attention: leveraging human attention for image captioning. arXiv:1904.00767
11. Dave A, Dubey R, Ghanem B (2012) Do humans fixate on interest points? In: Proceedings of the 21st international conference on pattern recognition (ICPR2012). IEEE, pp 2784–2787
12. Erhan D, Bengio Y, Courville A, Vincent P (2009) Visualizing higher-layer features of a deep network. Technical Report, University of Montreal 1341(3)
13. Gonzalez-Garcia A, Modolo D, Ferrari V (2018) Do semantic parts emerge in convolutional neural networks? Int J Comput Vis 126(5):476–494
14. He K, Zhang X, Ren S, Sun J (2016) Identity mappings in deep residual networks. In: European conference on computer vision (ECCV), pp 630–645
15. Hou X, Zhang L (2007) Saliency detection: a spectral residual approach. In: 2007 IEEE conference on computer vision and pattern recognition. IEEE, pp 1–8
16. Huang G, Liu Z, van der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)
17. Hubel D, Wiesel T (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol 160(1):106–154. https://doi.org/10.1113/jphysiol.1962.sp006837
18. Itti L, Koch C (2000) A saliency-based search mechanism for overt and covert shifts of visual attention. Vis Res 40(10–12):1489–1506
19. Jones C, Marron JS, Sheather SJ (1996) Progress in data-based bandwidth selection for kernel density estimation. Comput Stat 11:337–381
20. Judd T, Ehinger K, Durand F, Torralba A (2009) Learning to predict where humans look. In: 2009 IEEE 12th international conference on computer vision. IEEE, pp 2106–2113
21. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105
22. Kümmerer M, Theis L, Bethge M (2015) Deep gaze i: boosting saliency prediction with feature maps trained on imagenet. In: ICLR workshop, arXiv:1411.1045
23. Li K, Wu Z, Peng K, Ernst J, Fu Y (2018) Tell me where to look: guided attention inference network. 2018 IEEE/CVF conference on computer vision and pattern recognition, pp 9215–9223
24. Lin J (1991) Divergence measures based on the Shannon entropy. IEEE Trans Inf Theory 37:145–151
25. Mahendran A, Vedaldi A (2015) Understanding deep image representations by inverting them. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5188–5196
26. Mopuri KR, Garg U, Venkatesh Babu R (2019) Cnn fixations: an unraveling approach to visualize the discriminative image regions. IEEE Trans Image Process 28(5):2116–2125
27. Mould MS, Foster DH, Amano K, Oakley JP (2012) A simple nonparametric method for classifying eye fixations. Vis Res 57:18–25. https://doi.org/10.1016/j.visres.2011.12.006
28. Nguyen T, Park EA, Han J, Park DC, Min SY (2014) Object detection using scale invariant feature transform. In: Pan JS, Krömer P, Snášel V (eds) Genetic and evolutionary computing. Springer, Cham, pp 65–72
29. Rs R, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: visual explanations from deep networks via gradient-based localization, pp 618–626. https://doi.org/10.1109/ICCV.2017.74
30. Serre T, Wolf L, Bileschi S, Riesenhuber M, Poggio T (2007) Robust object recognition with cortex-like mechanisms. IEEE Trans Pattern Anal Mach Intell 3:411–426
31. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: 3rd international conference on learning representations (ICLR), pp 1–10
32. Simonyan K, Vedaldi A, Zisserman A (2013) Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv preprint arXiv:13126034
33. Stephens M, Harris C (1988) A combined corner and edge detector. In: Alvey vision conference, vol 15
34. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2818–2826
35. Tan M, Le Q (2019) Efficientnet: rethinking model scaling for convolutional neural networks. In: International conference on machine learning, pp 6105–6114
36. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. CoRR, arXiv:1706.03762
37. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y (2015) Show, attend and tell: Neural image caption generation with visual attention. In: Bach F, Blei D (eds) Proceedings of the 32nd international conference on machine learning, PMLR, Lille, France, Proceedings of Machine Learning Research, vol 37, pp 2048–2057
38. Xu W, Wang J, Wang Y, Xu G, Lin D, Dai W, Wu Y (2020) Where is the model looking at? Concentrate and explain the network attention. IEEE J Sel Top Signal Process 14(3):506–516. https://doi.org/10.1109/JSTSP.2020.2987729
39. Yamins DL, DiCarlo JJ (2016) Using goal-driven deep learning models to understand sensory cortex. Nat Neurosci 19(3):356
40. Yamins DL, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc Natl Acad Sci 111(23):8619–8624
41. Yosinski J, Clune J, Nguyen A, Fuchs T, Lipson H (2015) Understanding neural networks through deep visualization. arXiv preprint arXiv:150606579
42. Zeiler MD, Fergus R (2013) Visualizing and understanding convolutional networks. CoRR, arXiv:1311.2901
43. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2015) Object detectors emerge in deep scene cnns. arXiv:1412.6856

44. Zhou B, Bau D, Oliva A, Torralba A (2019) Interpreting deep visual representations via network dissection. IEEE Trans Pattern Anal Mach Intell 41:2131–2145
45. Zhou H, Yuan Y, Shi C (2009) Object tracking using sift features and mean shift. Comput Vis Image Underst 113:345–352. https://doi.org/10.1016/j.cviu.2008.08.006