Under Material Skin Lie the Bones of Identity

Padraig Boulton and Peter Hall
Department of Computer Science
University of Bath

Abstract

This paper explores the automated recognition of objects and materials and their relation to depictions in images of all kinds: photographs, artwork, doodles by children, and any other visual representation. The way artists of all cultures, ages and skill levels depict objects and materials furnishes a gamut of "depictions" so wide as to present a severe challenge to current algorithms – none of them perform satisfactorily across any but a few types of depiction. Indeed, most algorithms exhibit a significant performance loss when the images used are non photographic in nature. This loss can be explained using the tacit assumptions that underlay nearly every algorithm for recognition. Appeal to the Art History literature provides an alternative set of assumptions, that are more robust to variations in depiction and which offer new ways forward for automated image analysis. This is important, not just to advance Computer Vision, but because of the new understanding and applications that it opens.

1 Introduction

Humans possess a remarkable ability: the ability to understand the world visually. We see things – objects – and recognise them as belonging to an object class: dogs, trees, vases and so on. We see what these things are made of – their material – and again ascribe a class: glass, fur, metal, etc. We can infer the presence of things even when we cannot see them directly, hot air causes a shimmering for example. We can see what things are doing (people walking) and in some cases infer intent (walking to greet a friend). More than that, we have another equally remarkable ability: to communicate our visual understanding in pictures. We can recognise dogs in photographs, dogs painted by Reynolds, dogs drawn by children, dogs that wear clothes and walk on two legs (as in animation), dogs in road-signs; we recognise dogs made of porcelain and see their fur.

Object recognition by computer is an important and well researched problem. The technical problem is to ascribe the correct (agree with humans) label to (part of) an image showing an instance of the class. To date, computers are able to recognise objects from hundreds of different classes at rates around 98%. Yet these impressive figures mask a deep problem that highlights the extent to which machines are limited compared to humans: the figures in the literature derive from experiments using data sets comprised almost exclusively of photographs (also called "natural images"). When algorithms receive images outside their training depiction (e.g. photo, line drawing)

there is a fall in performance to below 60%. This is witnessed by Ginosar et al. (2014) , Cai, Wu & Hall (2015), Cai, Wu, Corradi & Hall (2015) , Collomosse et al. (2017), Geirhos et al. (2018*a*), Kubilius et al. (2018) and others. Such lower performance is evident too in Crowley & Zisserman (2014), Li et al. (2017), Li et al. (2018), Jenicek & Chum (2019). The only algorithm we know that generalises in a stable way from one depiction to another is Wu et al. (2014).

Research into automated recognition of materials is less common than object recognition. The literature is dominated by approaches to recover reflectance or transmissive properties of materials (Matusik et al. 2002, 2003, Deschaintre et al. 2018, Kim et al. 2010, Guarnera et al. 2016). Such literature will typically seek to elicit numbers for a mathematical model such as the Bidirectional Reflectance Distribution Function (BRDF) or one of its variants. A BRDF is defined at a tiny patch of some object: it is the ratio of light energy output in some direction to the light energy input from some possibly different direction. This is actually a 4-dimensional surface (two dimensions for input angle, two more for output angle) and different materials have different BRDFs (ie surfaces of different shapes). This could be used to identify materials, but only in photographs. The question of automatically recognising material type is a question not of estimation (of a BRDF) but of classification where the output is some class such as "metal, or "cotton or "brick. It is this problem that is of greater interest here, and it is one that has more recently come into view (Liu et al. 2010, Hu et al. 2011, Bell et al. 2013, 2015). All of this work assumes photographic like images for input. So far as we are aware, there is no work at all on automated recognition of materials under the many variations exhibited by artwork. For example, artists have depicted human hair in many different ways using many different media, all of which result in recognisable hair. But the artists don't know and likely don't care about the BRDF of hair.

We posit that both object recognition and material recognition are intertwined with the recognition of depiction style. This means that recognising a picture implies being able to recognise the content (the things), the materials (what things are made of), and depiction (photograph, drawing, painting etc) all at the same time. At the moment this is beyond any algorithm, but humans are able to perceive all of these (and much more) with no conscious effort.

The remainder of this paper discusses the above problems in greater detail to understand how the visual world can be effectively represented by depictions that do not adhere to physical reality. We first discuss how to recognise objects as *things*, Section 2, followed by the problems computers have when things are depicted in artwork in Section 2.1. We then consider the *materials* things are made of and argue that recognising things has a role in recognising material type. Finally, in Section 4, we investigate recognising the *depiction* style of a picture, where further new results are presented in relation to classifying depiction type. The single most important conclusion we reach is that artwork provides a significant challenge for Computer Vision research. This leads us to challenge the assumptions that are commonly made; analysis of pictures from Art History point to new assumptions that yield more robust algorithms, which in turn could be of greater use in the analysis of art.

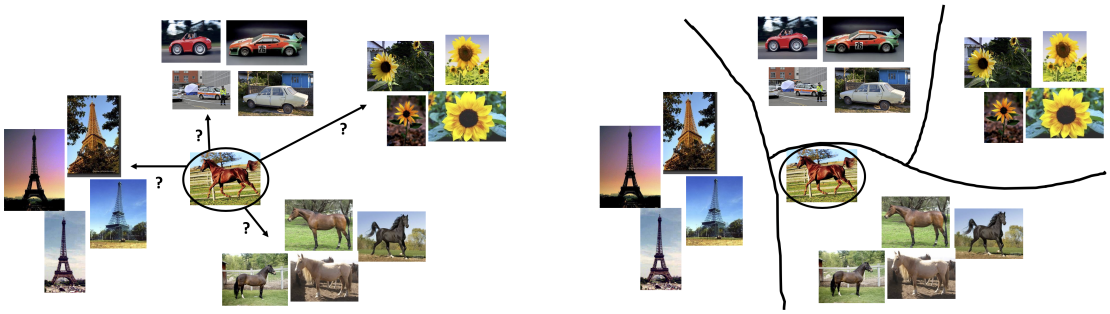

Figure 1: Left: Object recognition is the problem of placing a picture of a "visual object" into the correct "visual class". The classes comprise collections of examples of pictures showing the same kind of thing. Right: Algorithms learn boundaries between classes inside a high dimensional space (training) , then place new inputs into the class using the boundaries. Schematic shown.

2 Recognising Things

Object recognition is rightly considered one of the most important fields within Computer Vision. From a technical point of view, the problem reduces to placing a new image (a "visual object") into the correct group ("visual class"), as seen in Figure 1. As a note , strictly speaking we should differentiate between "classification" (what is in this picture), "detection" (where is this given thing in a picture), and "recognition" (what and where to be solved simultaneously). However, the use of these terms is a matter of convention rather than formal definition, and the terms are often used interchangeably. For convenience, that is often the case with recognition and classification, and is the case here.

Figure 1 provides a schematic description of the technical classification process. The visual classes are defined by collections of example images showing the same kind of visual object. During training the algorithms place boundaries between the visual classes. These boundaries partition the space that contains the images; in the figure this is represented by a two-dimensional plane but in practice the space has hundreds or even thousands of dimensions. Some techniques layout pictures into columns, with as many elements as pixels. This means each pixel value sits on a distinct dimension, if there are N pixels there are N dimensions. Each picture can now be considered as a single point in this N dimensional space, which contains every possible picture with N pixels. Once this "image space" is partitioned by the boundaries, the training images are no longer needed. Now, given an image not in the training set the question "which class?" becomes "in which partition?". The fractional number of times that the correct partition / class is decided by the algorithm is a measure of its performance.

The above description is very general. Specific algorithms vary in many different ways. For example, approaches may vary how the image space is defined or how the boundaries are constructed. Many algorithms from around 2000 to about 2013 employed the "Bag of Words" paradigm. This imagines that an image can be constructed from a set of visual symbols (words) — in reality small patches of image. This set is sometimes called a "dictionary" and much research at that time was devoted to crafting these words to have properties that designers thought of as important, eg rotating a word should not change it (compare this with CNN design, discussed below). Different visual

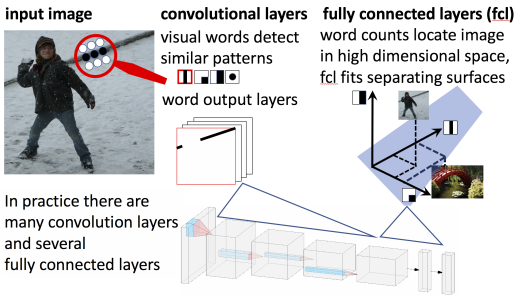

Figure 2: CNNs learn visual words to detect similar patterns; word count place a picture as a point in a space. Fully connected layers separating classes. In practice there are many convolutional and fully connected layers, which complicate but do not alter the basic operation in any fundamental way. Schematic shown.

classes can be differentiated from one another by the frequency with which they contain the visual words in the dictionary (so image space in this case is the set of all possible histograms). Otherwise said, the class model is a histogram of words in the dictionary. A new input image is analysed for the "visual words" it contains, from which a histogram is constructed to compare against the histograms for each of the visual classes. As for visual words, designers expended significant energy on developing different forms of classifier. Regardless of their efforts, this approach typically did not exceed a 70% recognition rate.

Convolutional Neural Networks (CNNs) provide state of the art performance for classification. They rose to prominence in 2012 with the introduction of "AlexNet" (Krizhevsky et al. 2012) – named after its inventor, Alex Krizhevsky. AlexNet raised classification rates from about 67% to above 90%. A schematic of such a network is seen in Figure 2, it contains two main parts. The first part, called the convolutional layers, learn small (typically, 3×3 pixels in size) discriminative image patches, called kernels. Kernels from the early layers, are combined to make larger patches in later layers (in practice the image is made smaller so that a 3×3 patch covers a larger area). These patches are equivalent to visual words, they are used to detect similar patch patterns in an image – we can imagine an image being assembled from the kernel patterns. Now though these words are learned without any human designer being involved. The second part is made of "fully connected layers", which act as the classifier. The aim of these layers is to combine the final patches to output a binary code that identifies an object, more exactly the pixel values in the patches are combined using weight values and then thresholded. Again, no human needs to design the classifier. The purpose of training is to set the kernel values in the convolutional layers and weight values in the fully connected layers; a process formally called "regression". The image space is now the set of all possible values that can be ascribed to all parameters in the network. This can comprise millions values, and the class boundaries twist and turn in this high-dimensional space as they "bend around" the class variations,, which why neural networks need so much data to train. The impressive performance of network compared with earlier work derive from their size and the fact the kernels and classifier weights are simultaneously learned rather than being separately designed by humans - the words and classifier "fit together".

2.1 Recognition in Artwork: the Cross-Depiction Problem

The problem of recognising objects in different depictions, photographs, paintings, sketches and indeed a full gamut of depiction varieties is called the "cross-depiction problem". There is large

and growing body of evidence that no algorithm is able to recognise objects across all depictions with the versatility that humans exhibit. Note that by "artwork" we include not just artefacts in museums and galleries, we include too road-signs, childrens' artwork, clip-art and anything else that expresses understanding in visual form, for which we use the short-hand "visual understanding".

We have conducted extensive tests on the cross-depiction problem (Hall et al. 2015) – the results of which are summarised in Figure 3. We constructed a dataset of 50 different visual classes and differentiated images between "photographs" and "artwork". The algorithms we used covered several variants of Bag of Words (*e.g.* Csurka et al. (2004)) using different features, namely SIFT (Lowe 2004), geomtric blurring (Berg & Malik 2001), self-similarity descriptors (Shechtman & Irani 2007), and two version of histogram of gradients (Dalal & Triggs 2005, Hu et al. 2013). Additionally, we included the Fischler Vector (Fischler & Elschlager 1973) and two constellation models – the well known Deformable Part Model (Felzenszwalb et al. 2010) and the so-called called "multi-graph" (MG). MG is a fully connected graph of multiple labels on nodes (Wu et al. 2014). We also tested using the neural algorithm VGG-19 (Simonyan & Zisserman 2014) that has been used for search of database of Fine Art (Crowley & Zisserman 2014).

For each algorithm, we trained on photographs alone (P), artwork alone (A), or a mixture (M) of the two, and then tested on photographs alone (P), on artwork alone (A), and on a mixture (M). In Figure 3, (M-P)) means "train on mix, test on photo", for example. When the training and test set both comprise photographs (P-P), algorithms perform in line with results commonly reported in the the literature: about 67% for BoW and above 90% for neural algorithms. All algorithms, with the exception of MG, suffer a considerable performance loss when generalising away from the training domain (e.g. photographs). The most dramatic fall comes when the training set is photographic and the test set is artistic (P-A): BoW algorithms fall to below 50% and in some cases 30%; even the neural architecture falls to just above 70%. Only the MG algorithm by Wu et al. (2014) remains more-or-less stable across all conditions.

These results are echoed in the data and observations of others, including but not limited to Ginosar et al. (2014), Crowley & Zisserman (2014), Collomosse et al. (2017), Kubilius et al. (2018), Geirhos et al. (2018*a*), Jenicek & Chum (2019). Some researchers have applied emerging techniques such as "meta learning" and "learning to learn" to the cross-depiction problem. In this understanding, the problem is to generalise what has been learned in training domains (depiction styles) to new test domains. For example, use what is known from photographs to influence learning about sketches; these are so-called Domain Generalisation algorithms. MLDG (Li et al. 2018) and MetaReg (Balaji et al. 2018) are two of the most recent neural examples. Both of them adjust the way in which the distance between objects is measured to allow for the difference between depictions. Neither of these are able to achieve a performance above 70% when generalising from photographs to artwork, which is consistent with the performance fall seen in Figure 3.

We equalled the performance of these complex algorithms in a far simpler way; by using fixed random values in the final fully-connected layer (Boulton & Hall 2019). An intuitive understanding may be had by considering Figure 2. The schematic is a useful simplification in that it associates different directions with distinct visual words, but mathematically speaking this is not necessary. If each direction line is rotated about the origin the coordinates (lengths of the dashed lines) of a

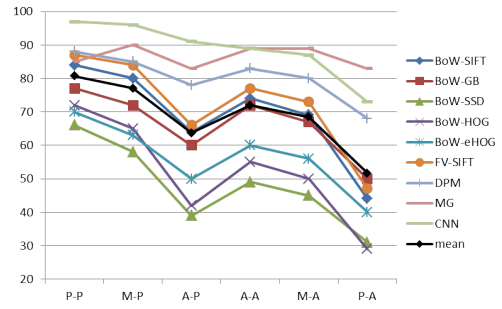

Figure 3: Results from our cross-depiction experiment. Bag of Words (BoW), graph-based models (DPM, MG) and a CNN. All but MG exhibit a significant fall in performance when generalising away from their training domain. The drop from photographs to artwork (P-A) is particularly pronounced.

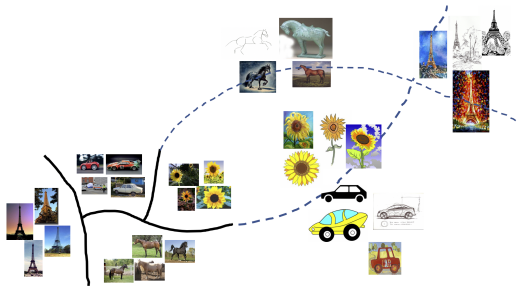

Figure 4: Performance falls because boundaries between classes are highly curved and the distance between artwork and photographs is large. Training on photographs alone may produce high quality boundaries for photographs (solid line), but when extended to reach artwork the boundaries become confused (dashed line). Schematic shown.

picture will change but different pictures will remain separated. An analogy is to tourist locations in cities – hotel maps, guide books, and official maps may all use different coordinate systems, but the relation between the sites remains constant. Now recall CNNs normally learn words and surfaces at the same time, by fixing the fully connected layers we force the (now random) directions to play an equivalent role to a separating surface. (Imagine a book with an arrow attached to its cover, when the book rests flat on a table the arrow points vertically up. If the book is moved in 3D space the arrow moves with it. This works in reverse: knowing where the arrow is tells us where the book is. In our case, the direction of the arrow is fixed.) Our network is forced to learn visual words that place pictures with the same content along the same direction line.

The fact such a simple approach is as effective as far more complicated alternatives suggests that even sophisticated measures in image space are not effective, which suggests that the particular form of image space may not suitable for the cross-depiction problem. In other words, pixels and measures made directly from pixels are poor ways to represent objects. This is not a surprise, because pixels are designed to store, carry and display information agnostic to content.

2.2 Explaining the Failure

The reason for lack of generalisation is "over fitting", Figure 4 provides an illustration of this problem. Evidence from Hall et al. (2015) suggests that photographs of different things (e.g. horse, Eiffel tower) are closer together in image space (more similar) than the same thing in different depictions. Consequently, the high-dimensional boundaries that separate "photo objects" objects must be very closely specified (with a lot of data) to separate classes in the photo-region of image space. When extended to reach artwork, the boundaries are very likely to criss-cross one another, which produces false results and so lowers performance. Simply put: the boundaries do not extrapolate well.

One response to the problem of over-fitting is to increase the quantity of data, in this case to

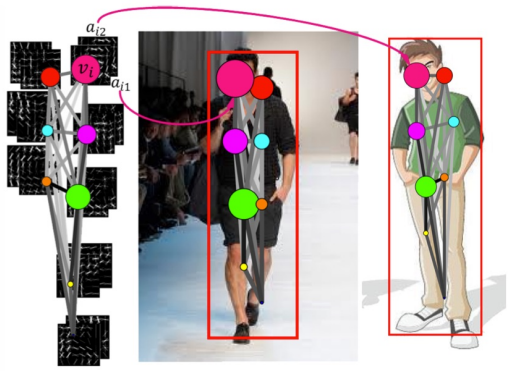

Figure 5: A graph model for the object class "person". Each node corresponds to some part and is connected to every other node by an arc. Projection determines the arrangement of nodes on the image plane. Each node is labelled with more than one visual word, allowing it to describe different denotations).

include images that span more than one depictive type. In principle this is correct because more data allows the between-class boundaries to be more closely defined – provided the machines are large enough to represent the highly curved multi-dimensional surfaces that separate classes. In practice, this would require such a vast number of images from all depictions that it is impracticable. Moreover, the fact that training datasets are heavily biased towards photographs is a major problem. Fitting boundaries between objects that span depictions demands a balanced number of examples in each depiction; any imbalance means the training algorithm are drawn towards the dominant depiction. It is not clear where the additional artwork would come from, and removing photographic data is not an option because neural networks require huge volumes of data to fit the separation surfaces.

We argue that the problem is in the way image space is represented. Pixels are designed for engineering problems such as the transmission, storage, and display of any image - regardless of content. Pixels by design, are agnostic to image content. People are very different. The fact they tend to draw what they know rather than what they see - and the fact they are so adept at recognition - and the fact they can do things like make a picture from a text description (and vice versa) is evidence that people use a very different way to represent image content. Image space for humans is not the same image space as used by machines: one has dimensions defined by low-level values, the other is probably bound up with natural language see, e.g. Pylyshyn (1973), Kosslyn & Pomerantz (1977) for philosophical discussion.

The artwork of children is particularly useful when considering an alternative way to represent things in images. Art Theorist Willats (1997) notes that provided the connectivity and spatial relations between parts is maintained, the object in childrens' art can be recognised. This means that rather than using pixel based visual words to represent objects, humans may use a graph based representation [1]. Consistent with this, we found representations based on graphs and spatial relations generalised well (Wu et al. 2014) – the MG-graph in Figure‘3 is reasonable stable. Figure 5 shows a typical MG-graph model – graph nodes correspond to object parts such as "head", "body", "hand" etc. In the model, the underlying skeleton for the object class is the same regardless of how any object in the class is depicted. The spatial location of these parts can change, upto a limit, and still be the same thing. The specific appearance of the object in an image is dictated by the

[1] A graph is mathematical way to encode relationships between pairs of things, Figure 5 shows a graph with "nodes" corresponding to things like heads and hands. The graph "arcs" connect node pairs possibly with a number to indicate the strength of the relationship.

choice of visual words / kernels/ pixel-patterns used to depict the individual parts. Words taken from the "photographic" region of pixel-space yield an overall photographic appearance, whereas visual words taken from the "line drawing" region yield a line-drawing like image. This model is not only robust to variations in depiction, but a variant of it has been used to synthesise child-like art, cave paintings, and output inspired by Miró (Hall & Song 2013).

Our argument – that networks are not good ways to represent objects in images – is not without controversy. Kubilius et al. (2018) have conducted careful laboratory controlled experiments to compare human and algorithm performance. They claim that their model is "sufficiently rich to support suprahuman-level performance across multiple visual domains". They trained on ImageNet (Geirhos et al. 2018*b*) and tested using processed versions of photographs: blurring, swirling, and block-shuffling image squares to make new pictures from old; they also used depictions they name paintings, sketches, cartoons, and line-drawings. They limited the number of test object classes to 10, and used 12 images per class in the various forms. They found that the network clearly outperformed humans when presented with block-shuffled photographs, and barely outperformed humans for swirled photographs. Humans out performed algorithms for some blurred photographs and all of the artwork. We argue that the correct conclusion is that the algorithms are learning low-level features that are robust to image edits such as block-shuffling, whereas humans are using representations that are sensitive to arbitrary changes in the spatial location of parts. We note too that the network outperforms humans only in cases that resemble training conditions – their networks is trained on randomly selected image sub-blocks that have been resized (blurred), for example. Kubilius et al. (2018) also found that a network trained on all images outperformed humans in all cases; this is not a test for generalisation and given the small number of classes and test images along with the large size of the network means the network has the capacity to encode all the data, by analogy it "remembers" everything. Furthermore, Geirhos et al. (2018*a*) have also directly compared human and algorithm performance using processed photographs, in their case adding noise of various kinds and to varying degrees in a way that did not change the spatial configuration of image parts. They use state-of-the art neural algorithms: ResNet-152 (He et al. 2016) , GoogLeNet, (Bengio et al. 1994), and VGG-19 (Simonyan & Zisserman 2014). They found humans consistently outperformed all of the algorithms, which is consistent with all of the evidence we have.

An important point lies within the details of Kubilius et al. (2018) experiment. They gave people a fleeting 0.1 seconds to recognise image content. The results show that people were better at recognising objects in artwork (close to 100%) than they are at recognising objects in photographs (about 80%). This is consistent with artwork being an abstraction of photographic content; abstract in the sense of discarding information that is redundant for recognition. Yet the artwork they use includes geometric and spatial distortions, a face has too-close eyes and lacks a nose, for example – but unlike block-shuffling, these distortions don't change the underlying structural relations between parts. All of which is evidence that the way humans represent things in images is very different from the way computers currently do.

3 Recognising Materials

In the introduction we distinguished between algorithms that attempt to recover reflective properties of things and algorithms that try to classify what a thing is made of. As noted, these are not the same problem – we are interested in the classification problem. In general, it is impossible to estimate reflective properties of materials from artwork, even in cases where the material can be recognised. Indeed, it is that fact that makes recognising materials in art such an interesting open question: what image properties do people use to correctly infer material type, even when the rendering contain no information about reflective properties. Whatever the answer is, it is clear that the problem of recognising materials in artwork is a more general and difficult problem than recognising materials in photographs.

Materials have been convincingly depicted in art since the time of van Eyck: gold, fabrics, milk, the skin of fruits, and the skin of people are but a few of the many that have been believably rendered. The translucent nature of oil paint possibly provides an advantage over media such as tempra and fresco, because many materials are themselves translucent. Just as humans can recognise objects robust to a wide variation in depiction, so too they can name materials also across a broad range of depictions, not limited to oil paints. So it is that hair, water, fire and even the breath of God have been depicted in fresco, pastel, and indeed many other media that lack transparency. Ideally, an algorithm would be able to capture such things, and maybe infer physical properties such as the feel of a woven garment, the weight of a concrete slab, or the coldness of ice. We are a long way from such sophistication.

The majority of the computational methods that exist are aimed at recovering physical reflectance properties, usually the bi-directional reflectance distribution function (BRDF), or similar alternatives (Torrance & Sparrow 1966, Deschaintre et al. 2018, Ergun et al. 2016). Recognition of material type, such as "glass", "metal", "brick"and so on does not require reflection properties to be inferred, instead algorithms us low-level statistical measures taken directly from an image. Recent authors have constructed databases and employed deep learning to identify materials (Liu et al. 2010, Hu et al. 2011, Schwartz & Nishino 2013, Bell et al. 2013, 2015, Caesar et al. 2018). It seems that *material* is harder to recognise than *things* – accuracy rates reaching about 71% are reported. When we take into account the wide range of appearances that any one thing may have, even in natural images, this figure is not altogether surprising. For example, water is coloured by the things it reflects, any particles suspended in it, and at sufficient density the water will become mud; material appearance is a complex psycho-physical phenomenon. Furthermore, the images in databases are all photographic.

Convincing explanations of human perception of materials are taking shape, e.g. Fleming (2014). In particular, the hypothesis is that humans form statistical models of appearance, and that such models are feasible because the appearance of objects changes systematically when light conditions change. This means that a set of measures can be taken directly from a compact distribution in measure space. Interestingly, Matusik et al. (2002) proves that physical realisable reflectance models form a convex set: any image measures used by the human visual system must be correlated to these, and a linear mapping would helpfully preserve convexity. Wiebel et al. (2015) use the mean,

standard deviation, skewness, and kurtosis along with minimum and maximum values to argue that skew correlates with gloss perception in computer graphic images, but standard deviation is a better correlate with gloss in photographs.

The fact that people can attribute the same material label to objects across many different depictions is, we argue, an under-researched phenomenon. Evidence from object recognition (eg Figure 3) and from depiction recognition (eg Figure 7) shows that the underlying statistics for some thing can be very different – yet humans continue to recognise that thing as being made of the same material. This strongly suggests that an explanation for material recognition cannot appeal to low-level statistics alone, but that knowing what a thing is will influence what it is seen to be made of. As with objects, it is unlikely that pixels are the optimal way to represent materials, rather some kind of class-conditional representation may be needed. Material recognition in artwork opens many interesting questions.

4 Recognising Depiction

There is a seemingly endless variety of depictions, photographs, oil paints, line drawings, tapestry, stained glass, doodles, and many more. All of them are capable of communicating things and materials. There is a small quantity of work on automatically recognising depiction style. The authors often use their own definitions of style mixed with more widely accepted terms. Karayev et al. (2013) use terms such as "HDR", "vintage", and "Noir" to label images in photographic dataset, while "Baroque", "Cubism", and "Impressionism" are example labels used for an artistic database. The authors utilize several different kinds of features to describe images and then classify the images into their style terms. They obtain reasonable results, with an error rate that depend on style: ranging from about 61% for "Romantic" photographs to 94% for Ukiyo-e and other artwork. Bar et al. (2014) adopt a very similar approach; they test a collection of low-level image descriptors and conclude that a particular design of visual words called "Local Binary Patterns" (Ojala et al. 1996) are to be preferred. Likewise, Falomir et al. (2018) use colour descriptors to categorise artwork into "Baroque", "Impressionism", and "post-Impressionism" to about 65% accuracy. Gultepe et al. (2018) learn rather than pre-define features. The above represent the general approach which is to take measures from local image regions and then build a classifier using a collection of such measures. We note that this methodology echoes the low-level approaches used to recognise things and material.

We have found a different, simpler approach to be effective for classification of depiction type. By depiction type we refer to the rendering media and the way it is applied as opposed to genre or school, etc. Recall our observations in Section 2.1 that (a) object classification does not generalise well over variations in depiction, and (b) that this can be explained (in part) by the wide variation in low-level statistic differences between depictions. We turned this to our advantage by conjecturing that depictions will respond differently to convolutional kernels. This is justified because a given kernel tends to produce a maximum response at image patches that resemble the kernel, and minimal values where the pattern is the kernel's negative. As example, a kernel that looks like a dark line on a white background will tend to pick up "ridges", where as a kernel dark one side

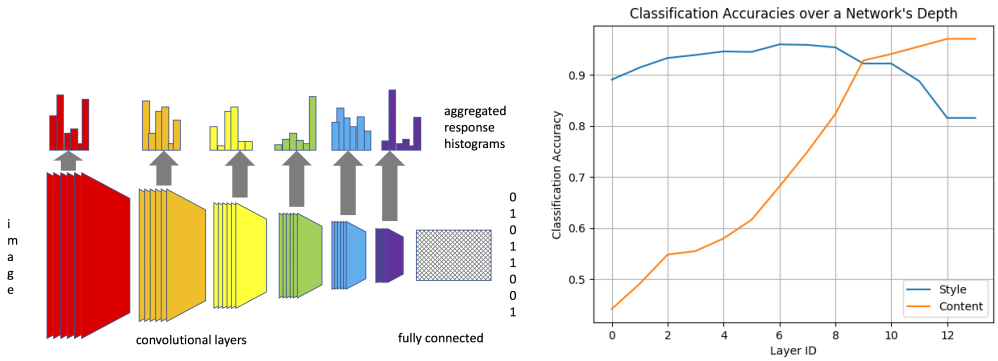

Figure 6: Left: a schematic of our approach to depiction classification. Right: results of depiction and object classification on a per-layer basis

and light the other will detect contrast boundaries (commonly called "edges"). Photographs tend to contain many contrast boundaries, especially at the small scales used by kernels, whereas (e.g.) line drawings will contain a greater number of ridges. We can therefore expect that the relative abundance of the kernels in an image (as measured aggregating the signal from each kernel over whole image) is informative with respect to depiction.

We tested this intuition using the Inception-v3 neural network (Szegedy et al. 2016); a very large neural network designed for object classification. The network is normally trained using ImageNet which contains thousands of object classes (Russakovsky et al. 2015); we trained for object class recogniton using the PACS dataset (Li et al. 2017) that comprises seven object classes each depicted in four styles, named "Photo", "Art", "Cartoon","Sketch". Following our intuition, we would expect the kernels of the trained Inception network to respond differently to different depictions: for any input image, we aggregate kernel responses over the whole image into a histogram. If our intuition is correct, then these histograms of kernel responses will characterise depiction style. We used these histograms to cluster and classify the images using a KNN classifier. We extract histograms after every layer to gain insight into how the Inception network operates in a multi-depiction setting.

Our procedure and results are shown in Figure 6. We see that early layers of the network can be used to classify depiction but not objects, where as later layers classify objects but not depiction. The high performance of object classification in later layers is explained because: (a) Inception is a large net designed for 1000s of classes, whereas PACS has 7 object types and 4 depiction styles; and (b) the net was trained on all data and no attempt was made to generalise to new depictions. As we will explain, the net is large enough that it has the capacity to internally learn to recognise different depictions of the same visual object, rather than learn a single depiction-agnostic method for identifying a visual object.

Figure 7 provides a more detailed examination of our results. It shows image-space representations from three locations in the network layers; an early layer, a mid layer close to where the classification curves in Figure 6 cross, and the final convolutional layer. We visualise these clusters twice – once in which datapoints are coloured by depiction class, and a second time coloured by object class. We used the t-SNE visualisation method to project data in a high dimensional space

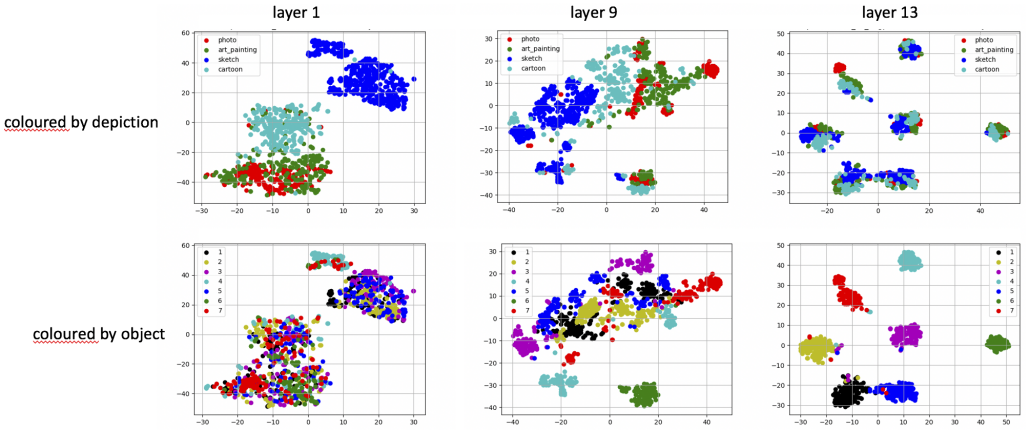

Figure 7: Visualisations of how images in PACS cluster at different layers of the Inception network. Each dot is an image, each column shows clusters from a given layer of the network, with rows colouring to dots by depiction (top) and object (bottom). In early layers the clusters are separated by depiction type, in later layers the cluster form into object classes. This explains the data seen in Figure 6; layer 9 is the crossing point classification curves and clusters illustrates the confusion.

in two dimensions (Maaten & Hinton 2008). The figure illustrates that in the early layers of the network the images implicitly cluster by depiction class, whereas later they cluster by object class. "Photo" and "Art" images seem to cluster together in this visualisation, but are separable in the original high-dimensional space, as evidenced in the classification results of Figure 6.

We did not generalise the trained Inception net to recognise objects in new depiction styles, rather the training data contained all styles. The results show clusters made by the test set only. Note that high object classification rate is consistent with Kubilius et al. (2018) who argued networks can exhibit "suprahuman-level performance ", but we found that the characteristic fall in object classification performance still occurs when generalising to unseen depiction classes. These results are evidence that the network does not learn a single generalising template for an object class; instead, the network is large enough to learn to aggregate distinct depiction-specific kernel responses into single object classes. This behaviour is problematic because it hinders generalisation to new depictions. Even using pre-trained networks that claim to employ a generic feature detector, the image space still seems to organise itself as shown in Figure 6 – see Wilber et al. (2017) for visualisations of ResNet embeddings of the BAM dataset, for example.

Before leaving this section we should make an important note. The first is that computer recognition of objects, materials, and depiction is all largely premised on visual words, whether learned or designed. We have already commented on the paucity of such representations for both object and material recognition, and here we are compelled to again rise above these low-level forms to understand art. Styles such as Byzantine art that uses inverse perspective, or traditional Chinese art that uses orthogonal projection cannot be distinguished from images made using the same media and application method, but constructed using linear perspective. As Willats Willats (1997) points out spatial organisation plays a significant role in style identification, just as it does in object and material identification.

5 Concluding Remarks

Object recognition is not solved. In particular, there is no algorithm capable of recognising objects regardless of the style in which they are depicted. Equally, the recognition of materials and of style (or depiction) remains open. Yet humans are able to disentangle all three; humans can make statements such as "man on horse, wearing armour, depicted using hatching".

The difference between humans and machine performance is fascinating, and a driver for investigation. It seems all but unarguable that humans make pictures to express their understanding of the world in visual terms. Exactly how humans represent images is at best unclear, and we are not willing to speculate. The Art History literature offers a way forward for Computer Vision. We have been heavily influenced by theorist John Willats (Willats 1997), who pointed out that childrens' art preserves connectivity between parts, that spatial relations between parts is variable within limits, and that shape and other geometric factors are important. Willats goes on to differentiate "projection" from "denotation", the first of which says *where* a part is, the second says *what* relates to the way marks are made. In Computer Vision terms these equate to spatial mappings and texture mappings. As described above, nearly all algorithms are based on texture and ignore the spatial component. MG algorithm (Wu et al. 2014) is one of the few exceptions – it was heavily influenced by Willats, and is reasonably robust to variations in depiction.

Decomposing style into projection and denotation is a useful idea, it moves us away from low-level image statistics and towards global semantic descriptions of depiction style. Thus it offers the possibility of disentangling things from their depiction. Once the class of a thing is known, the material it is made of tends to be limited – we know dogs tend to be covered in fur. We also know that dog figurines can be made of porcelain, as mentioned in the Introduction. The extent to which such background knowledge interacts with an intuition regarding porcelain reflectivity is not known to us, but is an interesting open question.

Artwork extends the problem of recognition in other ways too. Optical illusions such as the elephant with an indeterminate number of legs imply some information (ears, tusk) is more important to identity than others (legs). The vase/two-faces illusion, along with rabbit/duck and others, suggest that algorithms should not necessarily aim to produce a single answer but allow the possibility to "flip" between solutions. A similar conclusion can be drawn from Fine Art: when Magritte renders an apple over a face, when Archimboldo constructs faces from fruit, and countless other examples further challenge accepted assumptions of recognition. In such cases, a rendered thing serves two roles simultaneously: "pear" and "nose", for example. Moreover, recognition of mythical beasts, such as Minotaurs, and metaphorical representations of death and the devil hold yet further questions – recognition requires knowledge not in the image itself. Likewise, authors such as Berger (2008) who rightly point to the wider, societal meaning of symbols open similar questions for automated recognition.

Art and Art History has the potential to propel Computer Vision in fundamental ways. And in doing so may well benefit from a deeper, more robust tools to assist human experts. Clearly, there is a lot of work to be done.

References

Balaji, Y., Sankaranarayanan, S. & Chellappa, R. (2018), Metareg: Towards domain generalization using meta-regularization, *in* 'proceedings Advances in Neural Information Processing Systems 31', pp. 1006–1016.

Bar, Y., Levy, N. & Wolf, L. (2014), Classification of artistic styles using binarized features derived from a deep neural network, *in* 'Workshop at the European Conference on Computer Vision', Springer, pp. 71–84.

Bell, S., Upchurch, P., Snavely, N. & Bala, K. (2013), 'Opensurfaces: A richly annotated catalog of surface appearance', *ACM Transactions on Graphics* **32**(4), 111.

Bell, S., Upchurch, P., Snavely, N. & Bala, K. (2015), Material recognition in the wild with the materials in context database, *in* 'proceedings Computer Vision and Pattern Recognition', pp. 3479–3487.

Bengio, Y., LeCun, Y. & Henderson, D. (1994), 'Globally trained handwritten word recognizer using spatial representation, convolutional neural networks, and hidden markov models', *proceedings Advances in Neural Information Processing Systems* pp. 937–937.

Berg, A. C. & Malik, J. (2001), Geometric blur for template matching, *in* 'proceedings International Conference on Computer Vision and Pattern Recognition'.

Berger, J. (2008), *Ways of seeing*, Vol. 1, Penguin UK.

Boulton, P. & Hall, P. (2019), 'Artistic domain generalisation methods are limited by their deep representations', *arXiv preprint:1907.12622* .

Caesar, H., Uijlings, J. & Ferrari, V. (2018), Coco-stuff: Thing and stuff classes in context, *in* 'proceedings Conference on Computer Vision and Pattern Recognition', pp. 1209–1218.

Cai, H., Wu, Q., Corradi, T. & Hall, P. (2015), 'The cross-depiction problem: Computer vision algorithms for recognising objects in artwork and in photographs', *arXiv preprint arXiv:1505.00110* .

Cai, H., Wu, Q. & Hall, P. (2015), Beyond photo-domain object recognition: Benchmarks for the cross-depiction problem, *in* 'proceedings International Conference on Computer Vision Workshops', pp. 1–6.

Collomosse, J., Bui, T., Wilber, M., Fang, C. & Jin, H. (2017), Sketching with style: Visual search with sketches and aesthetic context, *in* 'proceedings Conference on Computer Vision and Pattern Recognition', pp. 2660–2668.

Crowley, E. J. & Zisserman, A. (2014), In search of art., *in* 'ECCV VisArt Workshops (1)', pp. 54–70.

Csurka, G., Dance, C., Fan, L., Willamowski, J. & Bray, C. (2004), Visual categorization with bags of keypoints, *in* 'ECCV Workshop on statistical learning in computer vision'.

Dalal, N. & Triggs, B. (2005), Histograms of oriented gradients for human detection, *in* 'proceedings Computer Vision and Pattern Recognition', Vol. 2, pp. 886–893.

Deschaintre, V., Aittala, M., Durand, F., Drettakis, G. & Bousseau, A. (2018), 'Single-image svbrdf capture with a rendering-aware deep network', *ACM Transactions on Graphics (SIGGRAPH Conference Proceedings)* **37**(128), 15.

Ergun, S., Önel, S. & Ozturk, A. (2016), A general micro-flake model for predicting the appearance of car paint, *in* 'Proceedings of the Eurographics Symposium on Rendering: Experimental Ideas & Implementations', pp. 65–71.

Falomir, Z., Museros, L., Sanz, I. & Gonzalez-Abril, L. (2018), 'Categorizing paintings in art styles based on qualitative color descriptors, quantitative global features and machine learning (qart-learn)', *Expert Systems with Applications* **97**, 83–94.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D. & Ramanan, D. (2010), 'Object detection with discriminatively trained part-based models', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **32**(9), 1627–1645.

Fischler, M. A. & Elschlager, R. A. (1973), 'The representation and matching of pictorial structures', *IEEE Transactions on Computers* **22**(1), 67–92.

Fleming, R. W. (2014), 'Visual perception of materials and their properties.', *Vision Research* **94**, 62–75.

Geirhos, R. , Temme, C., Rauber, J., Schütt, H. , Bethge, M. & Wichmann, F. (2018*a*), Generalisation in humans and deep neural networks, *in* 'Advances in Neural Information Processing Systems', pp. 7538–7550.

Geirhos, R., Rubisch, P. , Michaelis C, Bethge, M., Wichmann, F. A & Brendel, W. (2018*b*), 'Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness', *arXiv preprint arXiv:1811.12231* .

Ginosar, S., Haas, D., Brown, T. & Malik, J. (2014), Detecting people in cubist art, *in* 'ECCV VisArt Workshop', pp. 101–116.

Guarnera, D., Guarnera, G. C., Ghosh, A., Denk, C. & Glencross, M. (2016), Brdf representation and acquisition, *in* 'Computer Graphics Forum', Vol. 35, Wiley Online Library, pp. 625–650.

Gultepe, E., Conturo, T. E. & Makrehchi, M. (2018), 'Predicting and grouping digitized paintings by style using unsupervised feature learning', *Journal of cultural heritage* **31**, 13–23.

Hall, P., Cai, H., Wu, Q. & Corradi, T. (2015), 'Cross-depiction problem: Recognition and synthesis of photographs and artwork', *Computational Visual Media* **1**(2), 91–103.

Hall, P. & Song, Y.-Z. (2013), Simple art as abstractions of photographs, *in* 'proceedings of the Symposium on Computational Aesthetics', ACM, pp. 77–85.

He, K., Zhang, X., Ren, S. & Sun, J. (2016), Deep residual learning for image recognition, *in* 'proceedings Computer Vision and Pattern Recognition', pp. 770–778.

Hu, D., Bo, L. & Ren, X. (2011), Toward robust material recognition for everyday objects., *in* 'British Machine Vision Conference', Vol. 2, , p. 6.

Hu, R., James, S., Wang, T. & Collomosse, J. (2013), Markov random fields for sketch based video retrieval, *in* 'Proceedings of the 3rd ACM conference on International conference on multimedia retrieval', pp. 279–286.

Jenicek, T. & Chum, O. (2019), 'Linking art through human poses', *arXiv preprint arXiv:1907.03537* .

Karayev, S., Trentacoste, M., Han, H., Agarwala, A., Darrell, T., Hertzmann, A. & Winnemoeller, H. (2013), 'Recognizing image style', *arXiv preprint arXiv:1311.3715* .

Kim, D. B., Seo, M. K., Kim, K. Y. & Lee, K. H. (2010), 'Acquisition and representation of pearlescent paints using an image-based goniospectrophotometer', *Optical engineering* **49**(4), 043604.

Kosslyn, S. M. & Pomerantz, J. R. (1977), 'Imagery, propositions, and the form of internal representations', *Cognitive psychology* **9**(1), 52–76.

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012), Imagenet classification with deep convolutional neural networks, *in* 'proceedings Advances in Neural Information Processing systems', pp. 1097–1105.

Kubilius, J., Kar, K., Schmidt, K. & DiCarlo, J. J. (2018), Can deep neural networks rival human ability to generalize in core object recognition?, *in* 'proceedings Conference on Cognitive Computational Neuroscience'.

Li, D., Yang, Y., Song, Y.-Z. & Hospedales, T. (2017), Deeper, broader and artier domain generalization, *in* 'proceedings International Conference on Computer Vision'.

Li, D., Yang, Y., Song, Y.-Z. & Hospedales, T. M. (2018), Learning to generalize: Meta-learning for domain generalization, *in* 'proceedings AAAI Conference on Artificial Intelligence'.

Liu, C., Sharan, L., Adelson, E. H. & Rosenholtz, R. (2010), Exploring features in a bayesian framework for material recognition, *in* 'proceedings Computer Vision and Pattern Recognition', pp. 239–246.

Lowe, D. G. (2004), 'Distinctive image features from scale-invariant keypoints', *Intl. Journal of Computer Vision* **60**(2), 91–110.

Maaten, L. v. d. & Hinton, G. (2008), 'Visualizing data using t-sne', *Journal of machine learning research* **9**(Nov), 2579–2605.

Matusik, W., Pfister, H., Brand, M. & McMillan, L. (2003), 'Efficient isotropic BRDF measurement', *in* 14th Eurographics Workshop on Rendering, pp 241–248.

Matusik, W., Pfister, H., Ziegler, R., Ngan, A. & McMillan, L. (2002), 'Acquisition and rendering of transparent and refractive objects', *in* 13th Eurographics Workshop on Rendering, pp 267–278 .

Ojala, T., Pietikäinen, M. & Harwood, D. (1996), 'A comparative study of texture measures with classification based on featured distributions', *Pattern recognition* **29**(1), 51–59.

Pylyshyn, Z. W. (1973), 'What the mind's eye tells the mind's brain: A critique of mental imagery.', *Psychological bulletin* **80**(1), 1.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C. & Fei-Fei, L. (2015), 'ImageNet Large Scale Visual Recognition Challenge', *International Journal of Computer Vision* **115**(3), 211–252.

Schwartz, G. & Nishino, K. (2013), Visual material traits: Recognizing per-pixel material context, *in* 'proceedings International Conference on Computer Vision Workshops', pp. 883–890.

Shechtman, E. & Irani, M. (2007), Matching local self-similarities across images and videos, *in* 'proceedings Computer Vision and Pattern Recognition, pp. 1–8.

Simonyan, K. & Zisserman, A. (2014), 'Very deep convolutional networks for large-scale image recognition', *arXiv preprint arXiv:1409.1556* .

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. (2016), Rethinking the inception architecture for computer vision, *in* proceedings Computer Vision and Pattern Recognition, pp 2818–2826.

Torrance, K. E. & Sparrow, E. M. (1966), 'Off-specular peaks in the directional distribution of reflected thermal radiation', *Journal of Heat Transfer* **88**(2), 223–230.

Wiebel, C. B., Toscani, M. & Gegenfurtner, K. R. (2015), 'Statistical correlates of perceived gloss in natural images', *Vision Research* **115**, 175–187.

Wilber, M. J., Fang, C., Jin, H., Hertzmann, A., Collomosse, J. & Belongie, S. (2017), Bam! the behance artistic media dataset for recognition beyond photography, *in* 'proceedings International Conference on Computer Vision', pp 1202–1211.

Willats, J. (1997), *Art and representation: New principles in the analysis of pictures*, Princeton University Press.

Wu, Q., Cai, H. & Hall, P. (2014), Learning graphs to model visual objects across different depictive styles, *in* 'European Conference on Computer Vision', Springer, pp. 313–328.