Pattern Recognition ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Learning scale-variant and scale-invariant features for deep image classification

Nanne van Noord *, Eric Postma

Tilburg Center for Communication and Cognition, Tilburg University, Warandelaan 2, 5037 AB Tilburg, The Netherlands

ARTICLE INFO

Article history:
Received 31 January 2016
Received in revised form
13 May 2016
Accepted 6 June 2016

Keywords:
Convolutional Neural Networks
Multi-scale
Artist Attribution
Scale-variant Features

ABSTRACT

Convolutional Neural Networks (CNNs) require large image corpora to be trained on classification tasks. The variation in image resolutions, sizes of objects and patterns depicted, and image scales, hampers CNN training and performance, because the task-relevant information varies over spatial scales. Previous work attempting to deal with such scale variations focused on encouraging scale-invariant CNN representations. However, scale-invariant representations are incomplete representations of images, because images contain scale-variant information as well. This paper addresses the combined development of scale-invariant and scale-variant representations. We propose a multi-scale CNN method to encourage the recognition of both types of features and evaluate it on a challenging image classification task involving task-relevant characteristics at multiple scales. The results show that our multi-scale CNN outperforms single-scale CNN. This leads to the conclusion that encouraging the combined development of a scale-invariant and scale-variant representation in CNNs is beneficial to image recognition performance.

1. Introduction

Convolutional Neural Networks (CNNs) have drastically changed the computer vision landscape by considerably improving the performance on most image benchmarks [1,2]. A key characteristic of CNNs is that the deep(-based) representation, used to perform the classification, is generated from the data, rather than being engineered. The deep representation determines the type of visual features that are used for classification. In the initial layers of the CNN, the visual features correspond to oriented edges or colour transitions. In higher layers, the visual features are typically more complex (e.g., conjunctions of edges or shapes). Finding the appropriate representation for the task at hand requires presenting the CNN with many instances of a visual entity (object or pattern) in all its natural variations, so that the deep representation captures most naturally occurring appearances of the entity.

Three main sources of natural variation are the location, the viewpoint, and the size of an object or pattern. Variations in location are dealt with very well by a CNN [3], which follows naturally from the weight sharing employed in the convolution layers [4]. CNNs can also handle variations in viewpoint by creating filters that respond to viewpoint-invariant features [5]. Size variations pose a particular challenge in CNNs [6], especially when dealing with image corpora containing images of varying resolutions and depicting objects and patterns at different sizes and scales, as a result of varying distances from the camera and blurring by optical imperfections, respectively. This leads to variations in image resolution, object size, and image scale, which are two different properties of images. The relations between image resolution, object size, and image scale is formalized in digital image analysis using Fourier theory [7]. Spatial frequencies are a central concept in the Fourier approach to image processing. Spatial frequencies are the two-dimensional analog of frequencies in signal processing. The fine details of an image are captured by high spatial frequencies, whereas the coarse visual structures are captured by low spatial frequencies. In what follows, we provide a brief intuitive discussion of the relation between resolution and scale, without resorting to mathematical formulations.

1.1. Image resolution, object size, and image scale

Given an image its resolution can be expressed in terms of the number of pixels (i.e., the number of samples taken from the visual source); low resolution images have fewer pixels than high resolution images. The scale of an image refers to its spatial frequency content. Fine scale images contain the range from high spatial frequencies (associated with small visual structures) down to low spatial frequencies (with large visual structures). Coarse scale images contain low spatial frequencies only. The operation of spatial smoothing (or blurring) of an image corresponds to the operation of a low-pass filter: high spatial frequencies are

* Corresponding author.
E-mail addresses: n.j.e.vannoord@tilburguniversity.edu (N. van Noord), e.o.postma@tilburguniversity.edu (E. Postma).

http://dx.doi.org/10.1016/j.patcog.2016.06.005

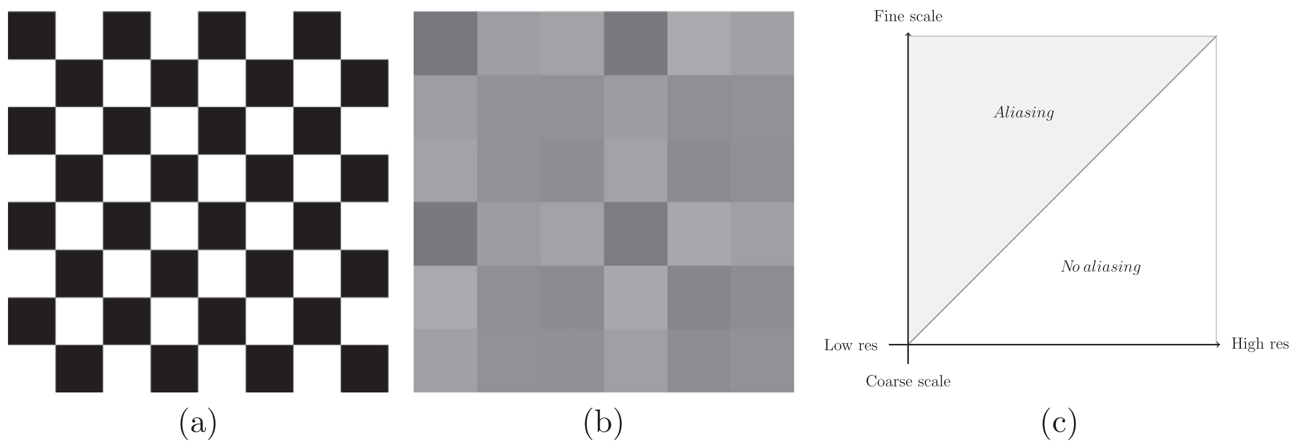

Fig. 1. Illustration of aliasing. (a) Image of a chessboard. (b) Reproductions of the chessboard with an image of insufficient resolution (6×6 pixels). The reproduction is obtained by applying bicubic interpolation. (c) The space spanned by image resolution and image scale. Images defined by resolution-scale combinations in the shaded area suffer from aliasing. See text for details.

removed and low spatial frequencies are retained. So, spatial smoothing a fine scale image yields a coarser scale image.

The relation between the resolution and the scale of an image follows from the observation that in order to represent visual details, an image should have a resolution that is sufficiently high to accommodate the representation of the details. For instance, we consider the chessboard pattern shown in Fig. 1a. Fig. 1b shows a 6×6 pixel reproduction of the chessboard pattern. The resolution of the reproduction is insufficient to represent the fine structure of the chessboard pattern. The distortion of an original image due to insufficient resolution (or sampling) is called *aliasing* [7].

As this example illustrates, image resolution imposes a limit to the scale at which visual structure can be represented. Fig. 1c displays the space spanned by resolution (horizontal axis) and scale (vertical axis). The limit is represented by separation of the shaded and unshaded regions. Any image combining a scale and resolution in the shaded area suffers from aliasing. The sharpest images are located at the shaded-unshaded boundary. Blurring an image corresponds to a vertical downward movement into the unshaded region

Having discussed the relation between resolution and scale, we now turn to the discussion of the relation of object size to resolution and scale. Real-world images with a given scale and resolution contain objects and structures at a range of sizes [8], for example, the image of the artwork shown in Fig. 2, depicts large-sized objects (people and animals) and small-sized objects (hairs and branches). In addition, it may contain visual texture associated with the paper it was printed on and with the tools that were used to create the artwork. Importantly, the same object may appear at different sizes. For instance, in the artwork shown there persons depicted at different sizes. The three persons in the middle are much larger in size than the one at the lower right corner. The relation between image resolution and object size is that the resolution puts a lower bound on the size of objects that can be represented in the image. If the resolution is too low, the smaller objects cannot be distinguished anymore. Similarly, the relation between image scale and object size is that if the scale becomes too coarse, the smaller objects cannot be distinguished anymore. Image smoothing removes the high-spatial frequencies associated with the visual characteristics of small objects.

Fig. 2. Artwork '*Hoefsmid bij een ezel*' by Jan de Visscher.

Please cite this article as: N. van Noord, E. Postma, Learning scale-variant and scale-invariant features for deep image classification, Pattern Recognition (2016), http://dx.doi.org/10.1016/j.patcog.2016.06.005