CLASSIFICATION OF NATURAL SCENES

Inaugural-Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften
der Fakultät für Naturwissenschaften
der Justus-Liebig-Universität Giessen

vorgelegt von
Jan Drewes
2006

Tag der mündlichen Prüfung: 22. September 2006

Dekan
Prof. Dr. Dr. Jürgen Hennig (Psychologie, Gießen)

1. Berichterstatter
Prof. Karl Gegenfurtner, PhD (Psychologie, Gießen)
Prof. Dr. Frank Bremmer (Neurophysik, Marburg)

2. Berichterstatter
Prof. Dr. Uwe Ilg (Neurobiologie, Tübingen)

ZUSAMMENFASSUNG

Seit einiger Zeit ist bekannt, dass das menschliche visuelle System zu einer erstaunlich schnellen Verarbeitung natürlicher Szenen in der Lage ist. Wenn man einem Beobachter zwei Bilder präsentiert, sei es auch nur für sehr kurze Zeit (z. B. 30ms), so das auf genau einem der Bilder ein Objekt einer bestimmten Objektklasse (z. B. „Tiere") zu sehen ist, so können Menschen dies nicht nur überaus zuverlässig (im Allgemeinen über 90% richtig), sondern auch extrem schnell erkennen – schon ab 150ms entscheiden manche Versuchspersonen über-zufällig richtig. Eine solch schnelle Entscheidung lässt nicht sehr viel Zeit für kognitive Abläufe. Es ist wahrscheinlich, dass diese Fähigkeit zur schnellen Entscheidung nicht etwa auf einem Abgleich mit einem im Gedächtnis gespeicherten Katalog von Tierbildern beruht, sondern aufgrund von sehr grundlegenden Bildeigenschaften geschieht. Eine mögliche Informationsquelle, die zu solch schneller Klassifikation beitragen könnte, ist das globale Amplitudenspektrum. Es ist das Ziel dieser Dissertation, zu untersuchen in wie weit das globale Amplitudenspektrum zur Klassifikation von Bildern beitragen kann, und ob dieses tatsächlich auch im menschlichen visuellen System geschieht. Durchgeführt wird dies am Beispiel der Bildklassen „Tier" und „kein Tier".

Dazu wurde zunächst eine knapp 11000 Bilder umfassende Datenbank geschaffen, die zu jeweils 50% aus „Tier"- und „nicht Tier"-Bildern besteht. Im folgenden wird zunächst ein Computer-Algorithmus ausführlich vorgestellt, der mit einer Trefferquote von ca. 75% in der Lage ist, „Tier"- von „nicht Tier"-Bilder zu unterscheiden, und zwar ausschließlich anhand des globalen Amplitudenspektrums.

Anschließend werden drei Hauptmerkmale des Klassifikationsverhaltens dieses Computer-Algorithmus mit dem Verhalten menschlicher Versuchspersonen verglichen, um Gemeinsamkeiten und Unterschiede herauszuarbeiten.

Im ersten Experiment wird die Anfälligkeit des Computer-Algorithmus auf Rotierten der Bildern mit der Anfälligkeit menschlicher Versuchspersonen verglichen. Aufgrund der Rotationsinvarianz des Computer-Algorithmus wird ein angenommenes Klassifikationsprofil mit den tatsächlich experimentell gemessenen Ergebnissen der Versuchspersonen verglichen. Eingesetzt wurde ein 2AFC-Paradigma, bei dem die Augenbewegungen der Versuchspersonen zur Ermittlung der relevanten Messgrößen werden. Es stellt sich heraus, das Menschen in der Tat eine ähnliches, wenn auch schwächer ausgeprägtes Profil aufzeigen. Bilder in kardinalen Rotationswinkeln (0°, 90°, 180°) werden dabei besser klassifiziert als Bilder in anderen Winkeln (45°, 135°).

Das zweite Experiment behandelt die individuelle „Schwierigkeit" von Bildern. Der Computer-Algorithmus vergibt aufgrund des Abstandes von der Klassifikationsebene eine Wertung der „Tier"-haftigkeit bzw. „nicht Tier"-haftigkeit jedes Bildes. Je höher die „Tier"-haftigkeit eines Bildes ausfällt, desto leichter sollte es einer Versuchsperson fallen, dieses korrekt zu klassifizieren.

Hier wurde ein Go/NoGo-Paradigma eingesetzt, bei dem die Versuchspersonen einen Knopf immer dann schnellstmöglich loslassen sollten, wenn ein „Tier“-Bild gezeigt wurde. Anhand von Reaktionszeit und Trefferquote konnte auch in diesem Experiment ein hohes Maß an Ähnlichkeit zwischen menschlichem Verhalten und Computer-Algorithmus festgestellt werden.

Im dritten Experiment wird die Reaktion auf den Wegfall des globalen Amplitudenspektrums betrachtet. Dazu wurde das individuelle Amplitudenspektrum der Bilder ersetzt durch das gemittelte Amplitudenspektrum ihrer jeweiligen Bildklasse. Eingesetzt wurden sowohl 2AFC- als auch Go/NoGo-Paradigma. Während der Computer-Algorithmus durch das Fehlen des einzigen Klassifikationsmerkmals auf Zufallsniveau abfällt, verringern sich die Leistungen der menschlichen Versuchspersonen nur geringfügig (übereinstimmend in beiden Paradigmen).

Im zweiten Teil dieser Dissertation wird untersucht, welche Datengrundlage zur Klassifikation herangezogen werden könnte, wenn das globale Amplitudenspektrum nicht in Frage kommt. Als Konsequenz wird ein neuer Computer-Algorithmus vorgestellt, der nicht nur Frequenz, Orientierung und Amplitude, sondern auch die Lokalisation der Information berücksichtigt. Als Datengrundlage dient eine Bildpyramide, die mehrere Frequenz- und Orientierungsbänder an jeder Stelle des Bildes beinhaltet. Mit diesem neuen Algorithmus wird eine Klassifikationsleistung von annähernd 78% erreicht. Durch eine genauere Analyse der Verteilung der relevanten Information über die Fläche eines Bildes wird dabei ein zuvor unentdecktes Artefakt aufgezeigt, welches bereits durch den Aufnahmeprozess der Bilder entstanden ist. Dieses Artefakt ist in der Lage, einem Computer-Algorithmus zu einer recht hohen Klassifikationsleistung (ca 74-75%) zu verhelfen, auch wenn das eigentlich relevante Bildzentrum ausgeblendet wird und somit das Objekt der Szenerie nicht mehr vorhanden ist. Dies ist von Bedeutung, da die Bilddatenbank, welche als Quelle sämtlicher Bilder dieser Arbeit dient, im Bereich der Wissenschaft weite Verbreitung genießt.

In einem vierten Experiment wird getestet, ob menschliche Versuchspersonen ebenfalls auf dieses Artefakt zurückgreifen können. Eingesetzt wird das bewährte 2AFC-Paradigma, bei dem selektiv verschiedene Bildausschnitte gezeigt werden. Ein Effekt des Artefaktes auf menschliche Versuchspersonen konnte nicht festgestellt werden.

Abschließend werden die Resultate aller 4 Experimente, sowie der Computer-Algorithmen diskutiert und geschlussfolgert, dass das globale Amplitudenspektrum aller Wahrscheinlichkeit nach keine dominante Rolle für schnelle Bildklassifikation im Menschen dient. Es wird eine Empfehlung ausgesprochen, dies bei zukünftiger Forschung im Bereich menschlicher Klassifikationsleistung zu berücksichtigen.

ABSTRACT

Humans are capable of rapidly classifying scenes by content, even when they are presented only very briefly. Classification accuracy can exceed 90%, while above-chance performance can be achieved in about 150ms. The global amplitude spectrum of an image has repeatedly been suggested to be a possible source of information for such fast classification. The aim of this thesis was to analyze the way in which humans classify images, specifically for the case of scenes which contain an animal or not. Indeed it was found that the information contained in the global amplitude spectrum, even at a rather coarse scale, is quite adequate for successful computer classification. In the first part of this thesis, a computer classifier was developed, capable of correctly classifying 75% of the images in our database. Then, 3 main characteristics of this classifier are identified and then tested against human subjects in 3 experiments:

First, the sensitivity to image rotation is tested. Using a 2AFC paradigm, human subjects were asked to decide which of two displayed images contained an animal. Eye movements were recorded to measure response time and classification accuracy. A high degree of similarity to the behavior of our computer classifier was found, with better performance on cardinal image rotations (0°, 90°, 180°).

Second, the order of the images in terms of classification difficulty is analyzed. We employed both a 2AFC paradigm and a Go/NoGo paradigm. In the latter subjects were asked to release a button as quickly as possible only when an animal image was shown. Here too a high degree of similarity between the results of the human visual system and those of our computer classifier was found.

Third, classification without the amplitude spectrum as a primary clue is tested. We modified our images, replacing the individual amplitude spectrum of each image with the mean amplitude spectrum of its image class. The individual phase spectrum was retained, unaltered. In this case, the computer classifier was “blinded” and would not exceed chance performance, while our human subjects still achieved high classification performance. This clearly contradicts the global amplitude spectrum hypothesis.

In the second part of this thesis, a different approach to computer classification is presented. The images were filtered in a way that allowed to analyze image content for different frequencies and orientations at discrete locations (as opposed to the global amplitude spectrum). The new computer classifier was able to achieve almost 78% correct classification. Also, a previously unreported artifact of the image capturing process was discovered within the image database used. This is remarkable because of the widespread scientific acceptance of the Corel Stock Photo Library used in this thesis.

Finally, the results obtained during all 4 experiments and the computational analysis are integrated and the possible use of the global amplitude spectrum in human visual classification is discussed. The main conclusion of this work is that the global amplitude spectrum is in all likelihood not a dominant factor in human visual classification. This finding should be considered in future research.

ACKNOWLEDGEMENTS

Foremost, I would like to thank my supervisor, Karl. You gave me direction and made sure I was maintaining a steady course, while still allowing me all the freedom a student could wish for. Also, you have been as patient as a single person can possibly be. The experience I gained while I was working for you will have a significant impact on my future life and scientific career. My second supervisor, Felix, you have been a constant source of inspiration and insight, and a shining example of successful cooperation. I wish to express my deepest gratitude to everyone in the “Allpsych” department of the University of Giessen. You have all helped me to make this possible and have encouraged me to hang on and cheer up during times of duress. When the need for conversation arose, there was always more than one of you available. Thorsten Hansen, you have been happily helping me out with any kind of issue, computational or otherwise, and I know it was not only because you “live” next door to mine. A cheerful tanks I would like to send to Nathalie Ziegler, you have taught me much of the art of using an EyeLink. I also owe gratitude to Miriam Spering, Knut Drewing, Volker Franz and Thomas Schmidt, not only because of their statistical aid. Christoph, thank you for your suggestions! And I will not forget Walter Kirchner, you aided me in anything else that a PhD student might need help with.

Cheers everyone!

Table of Contents

1. INTRODUCTION

1.1. Humans Can Do

When we view an image, be it printed, or on TV, or on a computer screen, we can usually understand what is displayed very quickly and easily. This happens seemingly without effort, and we do it countless times every day. Because of this, the human visual system has been the subject of both interest and admiration of many scientists. The ability of humans to grasp the content of an arbitrary image suddenly appearing in the field of view is indeed astounding. When shown, for example, a photography of a polar bear in it's natural surrounding, the question “Is there an animal on this picture?” would be considered trivial by almost any person, even though most of us rarely get to see an actual polar bear in real life. This task, however, is one that is considered highly difficult for todays computer algorithms. Just asking “Is there an animal on this picture?” for an arbitrary image poses a serious challenge to all currently established object recognition approaches. This is mainly due to the fact that there exists a huge number of different animals, and they can be depicted in an infinite manifold of perspectives, orientations, situations and be surrounded by just as many backgrounds. Browsing through a catalog of animal photographs, one might find several images showing the same type of animal, possibly even the very same individual specimen, but still none of the photographs will show the animal in *exactly* the same way (unless the photography was a duplicate). There will always be at least a small difference in posture, perspective, lighting and so on – which does not appear to be a problem for human observers. For classical computer algorithms, however, those slight changes are extremely difficult to deal with. As there is virtually an infinite number of individual views of animals, it is not possible to compile a catalog of animal images and compare them one-by-one. Aside from the near-infinite space required for this hypothetical catalog, the amount of time required to search through it would also be almost infinite.

Illustration 1: Polar Bear

Still, humans perform admirably well in this task. It does not really matter how an animal is shown to us – be it from its front, side, rear or even from the bottom, we will usually be able to find the animal in the scene quickly and easily. The same would hold true for other object detection / localization tasks, e. g. cars or vehicles in a city scene, or furniture in an indoor photograph. Object detection in humans is largely invariant to general scene properties such as perspective, lighting, object posture, object rotation and so on. This is not true for most of todays algorithms in computer vision.

1.2. Exactly What Can Humans Do?

The exact way in which humans detect and identify objects in scenes is still unknown, though interesting advances have been made recently. It is currently believed that, when presented with a novel photograph as mentioned above, a human observer will assemble the scene's abstract contents (objects, their positions and postures etc.) in several stages, starting from low-level features such as contrast, edges, orientation to higher level features such as contours and segmented shapes with textures to a final representation of abstract objects, their positions and meanings in the scene. This theory is supported by the pattern of eye-movements that humans use to attend different positions in the presented photograph. When observing a scene, eye movements occur because of the distribution of receptors on the human retina. While there is a very high density of color receptors in the center of the retina, the so-called "fovea", the density of receptors and especially that of color receptors drops dramatically with increasing eccentricity. The number of photo receptors is in the order of 200000-250000 receptors within the central 1° of the visual field, but there are far less in the outer periphery (see Illustration 2). This is reflected also in the number of ganglion cells at the fovea, and extends to the general representation of the fovea in the primary visual cortex (V1), about 50% of which is dedicated to the fovea. In order to perceive an entire scene, which typically spans at least several degrees in both width and height, the eye will need to scan it in order to sample the most important locations of the presented photograph with the higher foveal resolution. Those perceived patches which will then be stitched together when our brain processes the information delivered through the optical nerve. These point-to-point eye-movements are called "saccades" and happen very rapidly, up to 3-4 times per second. These very first eye-movements on a novel scene are believed to be mostly directed by a data-driven bottom-up strategy, designed to efficiently encode the images' most important locations to facilitate further processing. Since they happen very quickly, they are also assumed to be controlled mainly by early stages of the information processing, and therefore are attracted mainly by targets with a high low-level saliency.

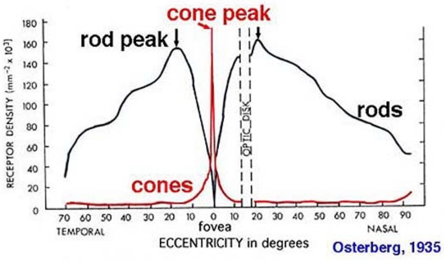

Illustration 2: Distribution of receptor density on the retina (www.webvision.org)

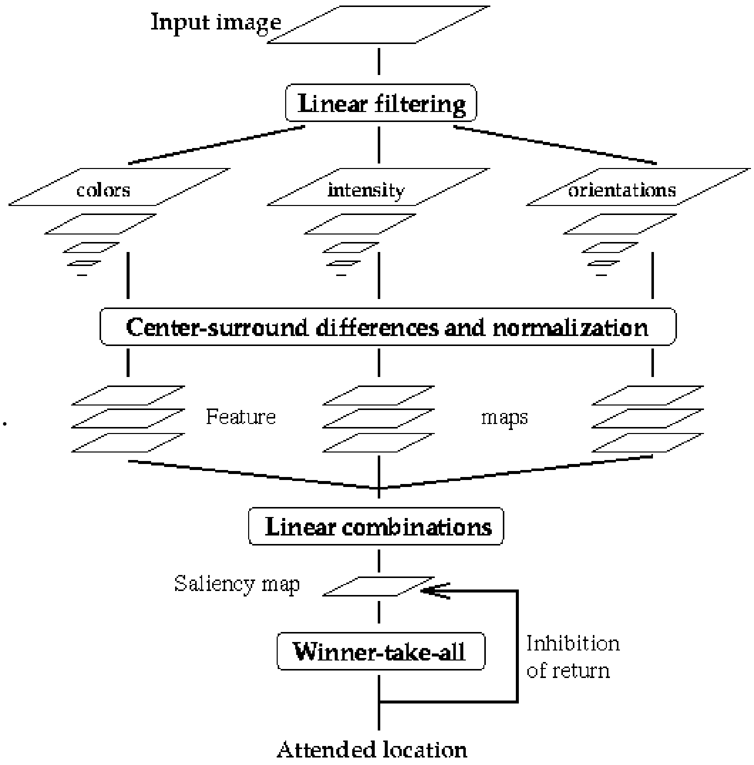

Illustration 3: The model of saliency as proposed by Itti, Koch and Niebur [Itti Koch Niebur 1998]

Conservatively, one would assume that successful detection of complex objects such as animals will only be possible after a rather high level of processing has been completed, e. g. after an abstract object representation has been formed.

In contrast to this assumption, Thorpe and colleagues [Thorpe Fize Marlot 1996] have discovered that humans are capable of detecting classes of objects in image presentations very quickly when experimental conditions are designed accordingly. The experimental setup was a so called “Go/No-Go” paradigm (see Illustration 4) in which the subjects were given the task to produce a “Go” response by releasing a pushbutton as quickly as possible (with a maximum allowed reaction time of 1000ms) whenever the presented picture contained any kind of animal, and to produce a “No-Go” response by not releasing the button for 1000ms whenever there was no animal on the presented picture.

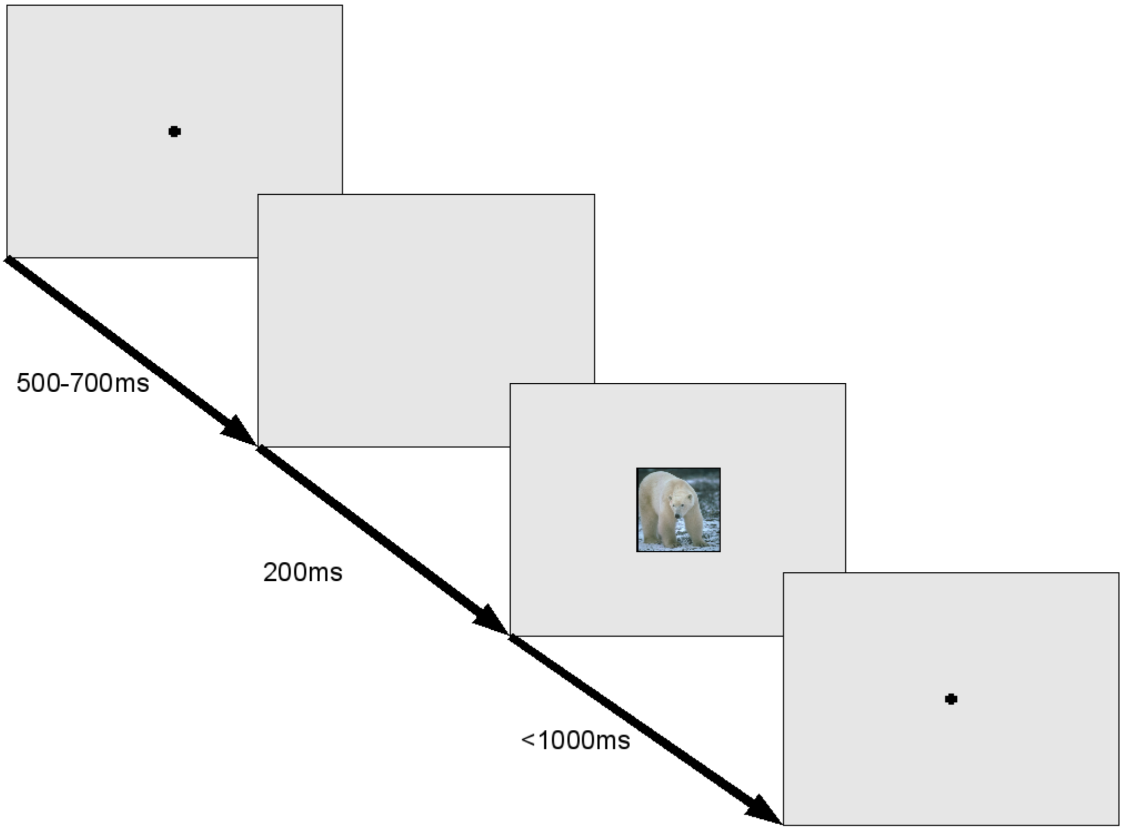

Illustration 4: The course of events during the Go / No-Go paradigm

Even though the images were presented only for a very brief moment (20ms), subjects were able to respond on average in 445ms, with the fastest subjects at 382ms median latency and the slowest at 567ms. Strikingly, the average percentage of correct responses was at 94% despite of the very fast responses. Also, the very short presentation time of the images effectively ruled out the factor of eye movements, meaning that the presence of an animal must have been detected literally at the first glance – even though they had no a priori knowledge about the type, size, position or even the number of animals shown. All images were novel to the subjects. These findings posed a serious challenge to many established models of human vision and raised a number of questions.

1.3. How Long Do We Actually Need to “See” an Image?

A detailed analysis of a scene usually involves several fixations and therefore takes much longer than the results in [Thorpe Fize Marlot 1996] would suggest. However, the task the observers had to perform was actually a simpler one. Just to tell *if there is any animal* present in the scene with a binary choice of “yes” or “no” might require less processing than the precise localization of the animal within the presented scene and its complete identification within arbitrary choices – the conclusion “Yes, there is some animal somewhere in the picture” might be easier to find than “There is a Polar Bear at the center of the image”. Still, the very short time needed to respond correctly to a presented image can not be explained well with the assumed stages of processing commonly associated with the detection of complex objects in human vision, especially since these <500ms even include the time necessary to perform the motor output response. In an attempt to better tell apart motor preparation and execution from the actual decision process (which necessarily has to precede the motor output), Thorpe and colleagues measured their subject's event related potentials (ERPs) [Thorpe Fize Marlot 1996] [Bacon-Mace et al 2005]. ERP recordings showed a statistically significant difference between target- and distractor-trials in form of a frontal negativity specific to No-Go animal detection trials, which started as early as 150ms after stimulus onset (the beginning of the display of the image).

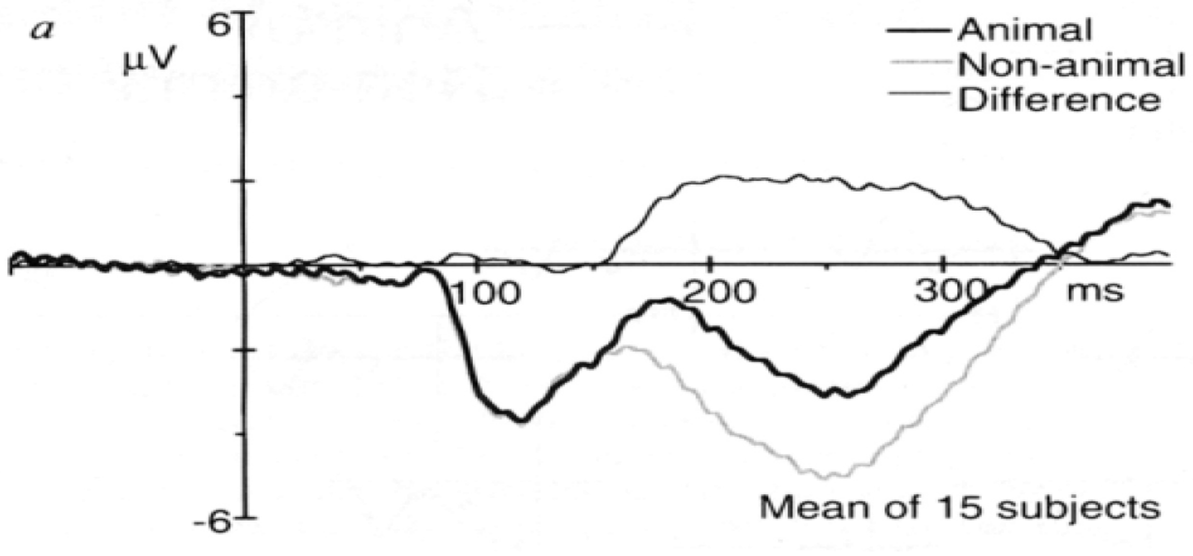

Illustration 5: Average brainwave patterns (ERPs) [Thorpe Fize Marlot 1996]

This enormously short response time is apparently related to the decision process only, since the onset latency of the frontal negativity and that of the following motor reaction showed no correlation. One can therefore conclude that the necessary visual processing for the decision process must already have been sufficiently completed at this very early time. This is supported by the finding of Kirchner, Thorpe and colleagues that Human reaction times can be further enhanced when using a 2 alternatives forced choice paradigm (2AFC) instead of the “Go/No-Go” one [Kirchner Thorpe 2006]. In this, the subject is presented with 2 alternative images, one of which being a target image, the other one being a distractor. They are usually shown side-by-side, modifying the subjects critical decision in the task into not *if* there is an animal, but *where* is the animal (binary choice: left or right). In this setup there were no “No-Go” responses, hence the tag “forced choice”, requiring the subjects to “prime” themselves for a quick reaction, which would always be triggered in one of two possible ways. This caused a dramatic drop in median reaction times of about 100ms, lowering the median reaction time of all subjects to about 350ms, without degrading classification performance (on average, 94% of the images were correctly classified). Even at only 250ms, some subjects were still able to perform statistically significantly better than chance performance, which complies with the reported minimum time required to generate a “reaching” command being around 80-100ms [Kalaska Crammond 1992].

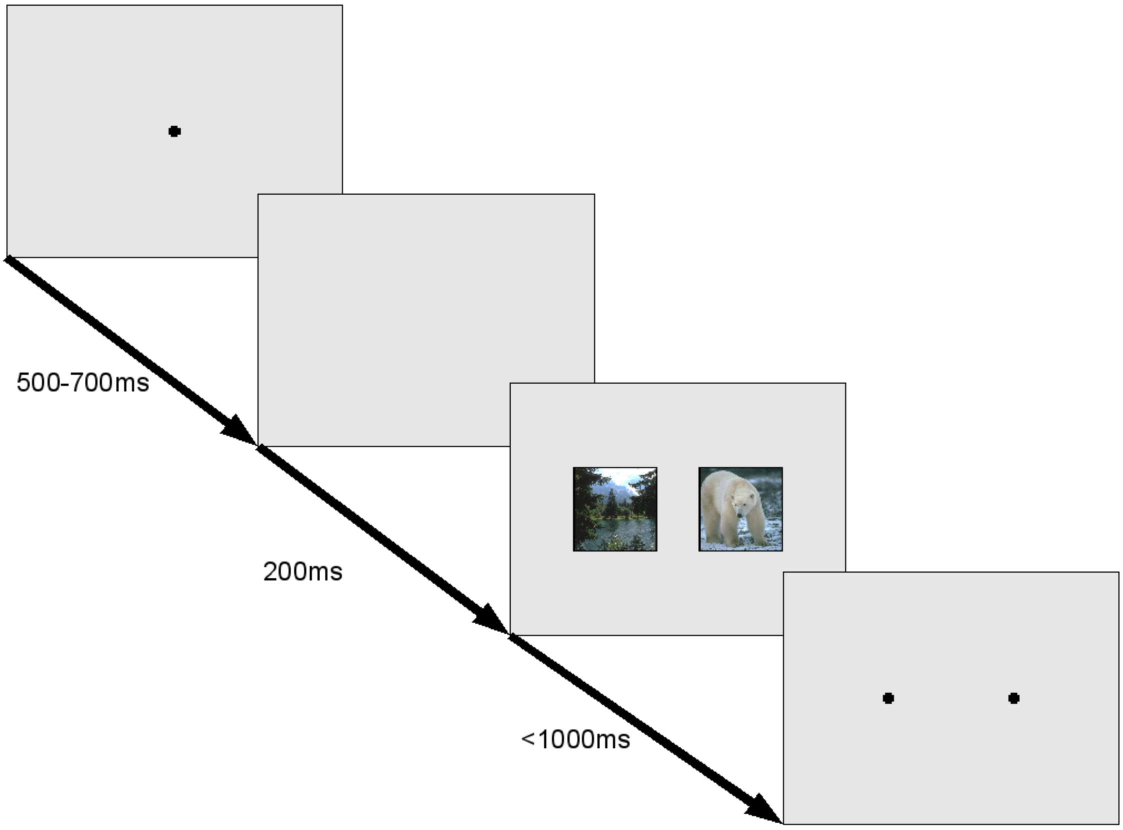

Illustration 6: The course of events during a 2AFC gap paradigm similar to the one in [Kirchner Thorpe 2006]

Another set of ERP recordings performed during this new experiment showed the same frontal negativity, emerging about 150ms after stimulus onset. This unchanged negativity onset time in combination with the dramatically shortened median reaction time supports the assumption that the measured ERP difference is not related to motor planning/preparation, but to decision making, which is also supported by the location of the most intense negativity, over the Pre-Frontal Cortex (PFC). This area of the brain is commonly assumed to play a major role in categorical judgments and decision making (see Illustration 7).

The ability for this quick detection process, called “ultra rapid visual classification”, was also shown to exist in rhesus monkeys [Fabre-Thorpe et al 1998]. In fact, the monkeys were able to perform a similar task with different categories (food vs. non-food) at a very high level of correctness (about 90,5%), while in the animal vs. non-animal task, the monkeys performed at about 84% correct. Their reaction times in this Go/No-Go paradigm were somewhere between 100 and 180ms faster than that of the human response times mentioned above, a speed advantage which may be based on the smaller size of their brains: the transduction velocity of spikes in the cranial nervous system is commonly put at roughly 2m/s, which might account for the faster responses of the monkeys. Aside from the similarities between humans and monkeys both in processing speed and accuracy, it is notable that about 90% of the errors made by humans and monkeys are the same, giving reason to assume that the mechanisms involved in object detection both in monkeys and in humans are essentially the same as well.

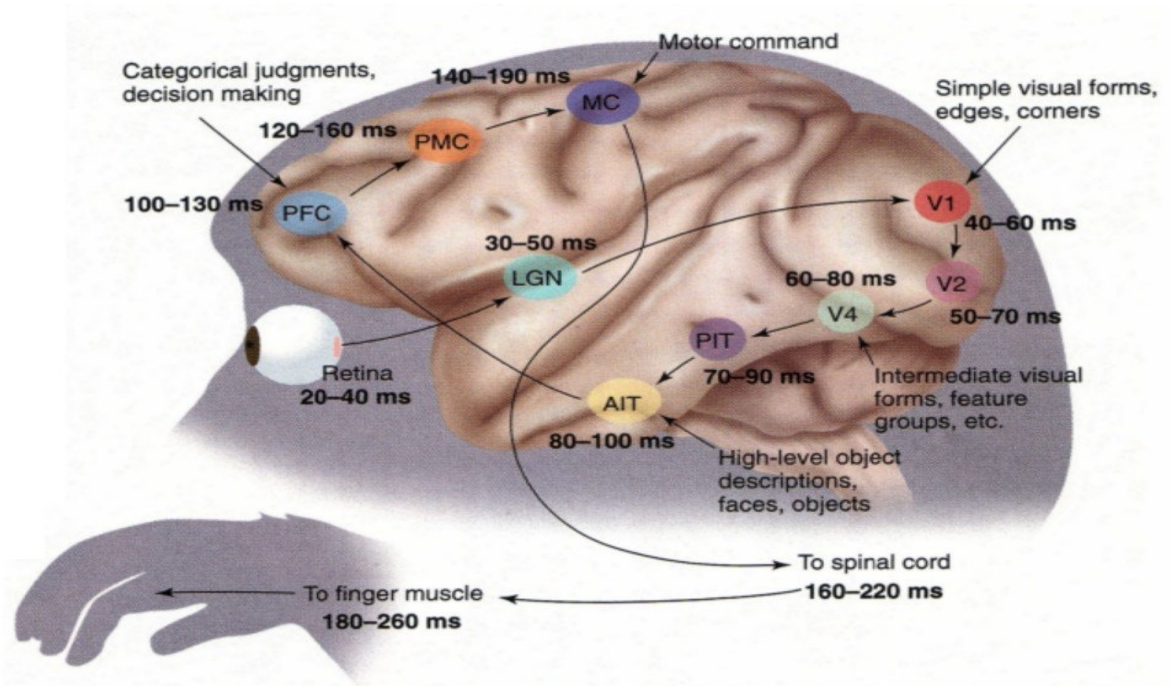

Illustration 7: Timecourse of visual processing in monkeys (Thorpe et al.)

Now since all these experiments have dealt with animal vs. non-animal (or food-nonfood) detection tasks, one might argue that the quickness of the responses is due to some hard-wired mechanism that evolved over the time course of several milleniae. The biological survival relevance of the quick detection of animals, be it from the view of a predator as possible food or from the view of the prey as a possible threat, is undisputed. This theory of a hardwired mechanism has been swiftly disproved by another experiment of Van Rullen and Thorpe [VanRullen Thorpe 2001]. In this experiment, applied the Go/No-Go paradigm to a man-made category: “means of transport”. The target images contained cars, airplanes, boats and many more vehicles, all of which are certain to not possess a significant biological relevance. As a cross-check, they also tested their new subjects on the known animal/non-animal task, and interchanged some of the images to serve as distractors in their respective opposite tasks: in the animal task, ½ of the distractors where randomly selected from the “means of transport” category, and in the “means of transport” task, ½ of the distractors came from the set of animal images.

Illustration 8: Images similar to the categories used in [VanRullen Thorpe 2001]

The percentage of correct responses was similar to the former experiments, as were the reaction times. This effectively rules out the possibility of a general, unintentional difference between the sample images to be responsible for the results; it truly had to be a detection of the desired object class that was responsible for the subjects performance. No significant differences were found between the two categories; this suggests that there exists no category-specific hardwired mechanism in our visual system. This would also explain the findings that their subjects overall performance was slightly reduced when they switched between object categories from block to block as one would not expect such a reduction in performance if subjects were merely switching between different hardwired decision mechanisms. However, there might still exist a partially fixed mechanism, which can be tuned to different image categories, probably including any acquired category. This is, however, still subject to current research.

1.4. A Limit to the Speed of Processing

Another experimental setup by Kirchner et al. [Kirchner Thorpe 2006] modified the 2 alternatives forced choice paradigm to use eye movements instead of button presses as motor responses. In this setup, the fastest response times were measured to be around 150ms and were still far above chance performance. As even eye movements need a certain time to be prepared after the decision has been made, this implies that the actual detection and decision have been computed before the 150ms, thus even before the onset of the frontal negativity that Thorpe and colleagues reported [Thorpe Fize Marlot 1996].

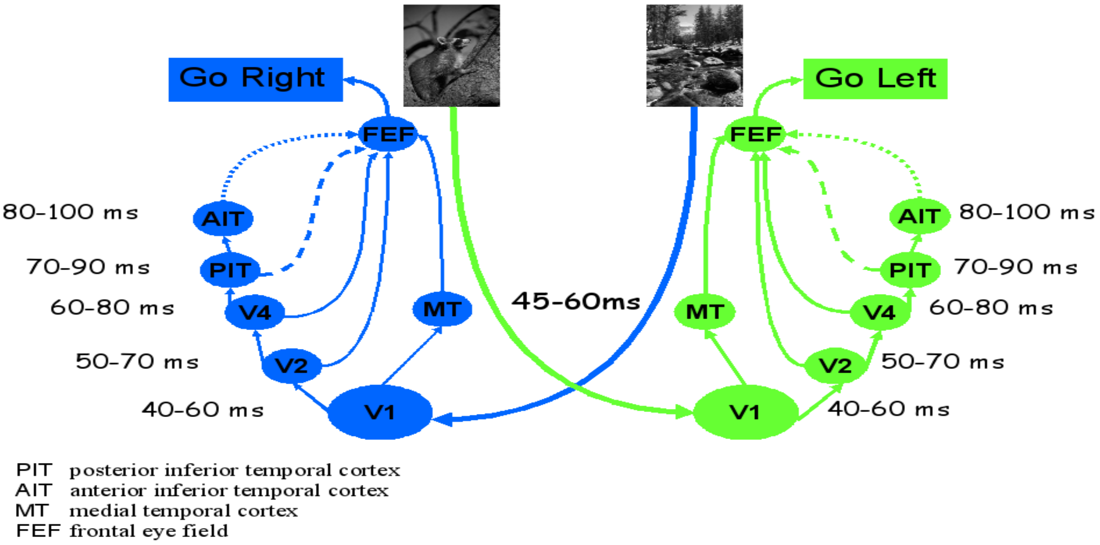

Illustration 9: Taken from [Kirchner Thorpe 2006]: timecourse of visual processing

When combined with the approximated speed of the brain (see Illustrations 7 and 9), this leaves only about 45-60ms for the actual information processing, which strongly supports the theory that ultra-rapid object detection is based mainly on feed-forward networks, not involving extensive feedback or top-down interaction. However, this does of course not exclude the possibility that the class of objects to be detected can in some way be selected or tuned via means of a top-down mechanism. More evidence for the feed-forward nature of visual object categorization comes from another experiment [Fabre-Thorpe et al. 2001]: the subjects were not able to further shorten their fastest responses on familiar images, even after 3 weeks of training – afterwards, novel scenes were still classified equally as good as the well-known ones that had been used in training over and over again. This, however, holds only for the images that were classified within the time interval typical for ultra-rapid image categorization. The more difficult images, which had longer latencies (sometimes > 500ms) could be classified significantly faster after the 3 weeks of training, suggesting that there in fact is a top-down or memory-feedback involved at a later processing stage that is not applied to the easier, ultra-rapidly classifiable images, or is at least not necessary with these to find a reliable decision.

During the ultra-rapid categorization experiments performed by Thorpe, Gegenfurtner and colleagues, more interesting discoveries were made: Subjects were able to respond to images shown for the usual 20ms, but in unpredictable locations across the entire extent of the horizontal visual field [Thorpe Gegenfurtner et al 2001]. Even at 70° eccentricity, subjects were still able to categorize the shown images at over 60% correct, being significantly above chance level – they knew they had seen an animal even though they were sometimes unable to identify what kind of animal it had been. The fact that the images appeared at unpredictable locations and disappeared again too quickly to allow for any sort of eye movement showed that locally focused visual attention is apparently not essential, contrasting once more the conventional view of the human visual system. This again was further supported by Rousselet and colleagues, who showed that the time required to process several images shown simultaneously does not increase linearly with the number of images shown [Rousselet et al 2004]. There exists a high level of parallelism during the earliest stages of vision, with competition arising only later on at frontal sites, an idea that is partially carried by the known fact that in V1, several low-level features like edges, corners and orientations are all computed in parallel over the entire visual field. Also, they found possible evidence for a intra-hemispherical parallelism (each hemisphere might be analyzing a different scene in parallel), as also suggested in [Kirchner Thorpe 2006] (see Illustration 9).

Another facet that might explain some of the very short latency of the visual system is pipelining. While later stages of the visual system are still processing previous visual input, earlier stages might already begin to work on newer, more recent impressions. To think that the highly efficient and optimized human brain would actually allow a large portion of its subsystems to be inactive most of the time, as the visual system as we know it would be without pipelining, is probably a far-fetched thought all by itself. Still, in a recent rapid serial visual presentation (RSVP) study Keysers and Perrett showed that human subjects can detect images of certain categories in sequences presented at rates of up to 75 images per second [Keysers Perrett 2002]. As this would only leave about 13ms for each image to process, it is highly unlikely that any mechanism in the human visual system could perform at such level without extensive use of pipelining. It is therefore an interesting question just how long a picture actually needs to be displayed for the visual system to gather sufficient information to allow reliable processing.

The retina and the very early stages of the visual system can perform like a "sample-and-hold" circuitry. This means that the perceived input as it is delivered to the higher stages of the visual system can persist even after the original stimulus, for example an image, has already vanished. Because of this, a simple variation in stimulus presentation time will not be able to produce significant performance differences, as the visual system would reach a saturation point too early (20ms, as used in the above studies, already appears to be sufficient). The sample-and-hold function can be disrupted by the appearance of a sufficiently dominant new stimulus that differs from the original one very much. Such a stimulus would then be called a mask, and the time between the onset of the stimulus and the onset of the mask, replacing the stimulus, is the stimulus onset asynchrony (SOA), posing as a more suitable measure of the time available for a stimulus to be processed within the earliest stages of the visual system.

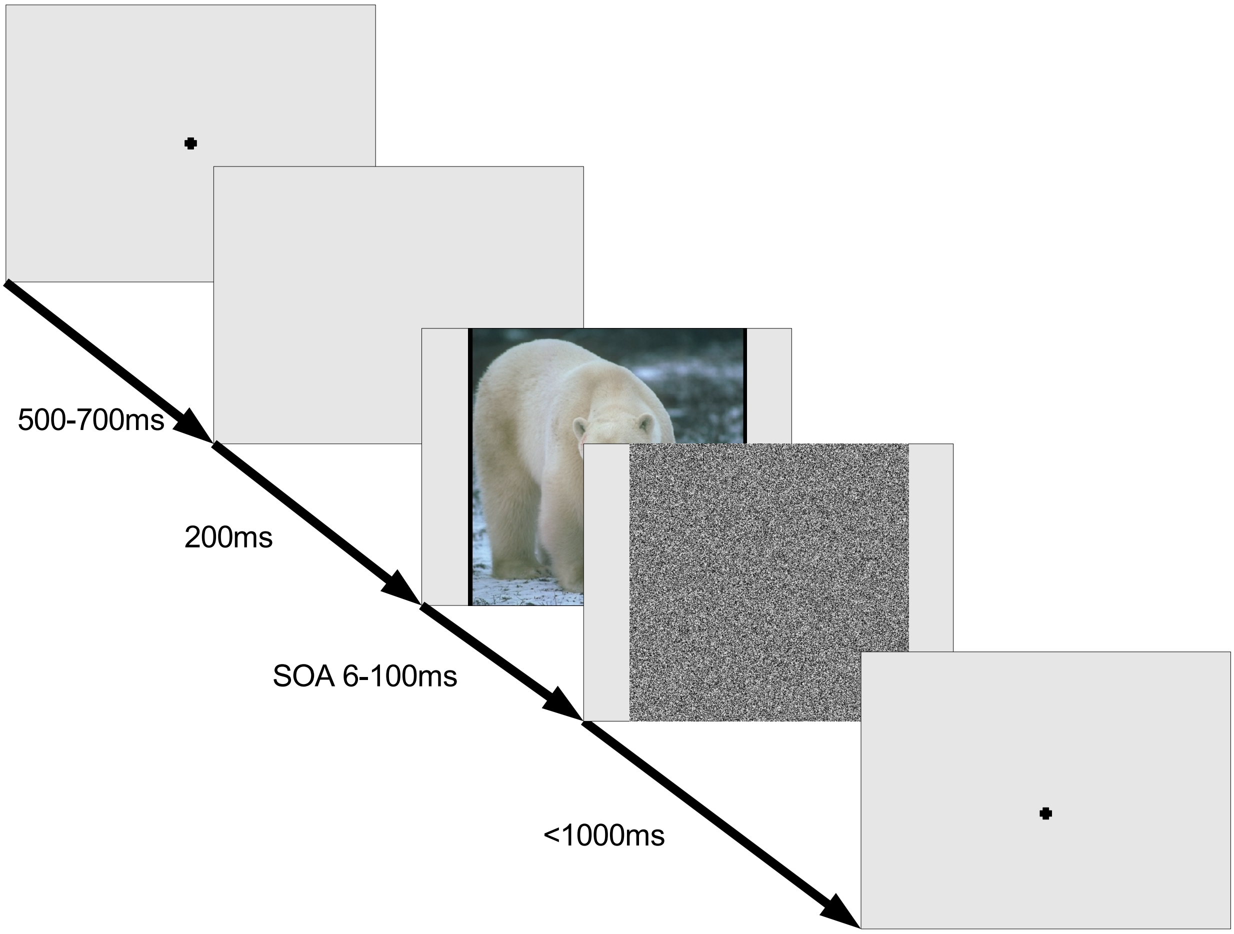

Illustration 10: Mask paradigm similar to the one used by Bacon-Macé et al.

Bacon-Mace and colleagues showed that without masking, the stimulus presentation time can be as short as 6,25ms without a decline in categorization performance [Bacon-Mace et al 2005]. When masked, however, an SOA of also 6,25ms resulted in near chance performance: the subjects behaved as if no image (at least none that contained anything recognizable) had been presented at all, which means that at this short SOA the mask effectively overrides the perception of the preceding image. Categorization performance steadily increased with longer SOAs, along with a decrease in median reaction times, until at about 80ms SOA, the mask had no major effect on the subjects performance anymore.

1.5. Simpler Than Expected?

This ultra-rapid categorization mechanism has given reason to several examinations as to what kind of information could be used to reach a decision after such limited processing time. Vogels reported that in a “tree” vs. “non-tree” categorization task, simple low-level features such as texture, size and color could not account for the subjects categorization performance [Vogels 1999a][Vogels 1999b]. The use of color specifically in the “animal” vs. “non-animal” task can be ruled out since the removal of chromatic information from the images shown does not essentially affect the performance either in humans or monkeys. This is commonly accredited to the fact that the chromatic input from the parvocellular pathways reaches the visual cortex slightly (~20ms) after the luminance based information of the magnocellular pathway, so the earliest responses to visual input would not be influenced by chromatic information [Fabre-Thorpe et al. 2001]. It has further been found that reaction times to simple stimuli such as geometric shapes can be up to 50ms faster (for mean reaction times) than the responses to target images mentioned here [Aubertin et al 1999]. If animal- or object-detection in general would be based on such simple stimuli alone, one would not expect such a difference in response times.

The stimulus properties used for categorization must therefore be more complex than a single low-level feature, yet must still be processable in a single feed-forward sweep through the visual system. The unique ultra-rapid categorization ability works best with complex, real-life scenes, and has been tested most thoroughly with natural scenes. Images of abstract, artificial objects such as letters or geometric shapes usually need active attention to be recognized, unless the presentation has been simplified a lot (e. g., single letters or words instead of entire paragraphs, or unique sine grids instead of complex structures). Evolution has obviously tuned the abilities of our visual systems to the needs of everyday vision.

1.6. The Properties of Natural Scenes

A term that will frequently be referred to in the following pages is “natural scene”. A scene in general is an arrangement of objects and non-objects in such way that they are contained in a single view, which is then usually represented by a photography. A “natural” scene in this context is a photography of some real-world view, be it a mountain, desert or a forest, any kind of natural environment, sometimes with certain animate or inanimate objects such as rocks, animals, or plants, in the studies referred to mostly at a central, focused position. When artificial (meaning “man-made”) scenes are examined, they frequently contain long, straight edges and sharp angles (e. g. buildings with windows, balconies etc.). While the true meaning “natural” would not allow for any artificial and thus not natural objects to be depicted, we will expand our use of the term to all kinds of real-world photographs, showing man-made objects or even entire surroundings as well as truly natural elements.

One thing often mentioned in literature is that natural scenes share a common spectral property: their power spectrum roughly follows a 1/f curve [Field 1987], and generally, horizontal and vertical components are more frequent and thus more prominent in the power spectrum than oblique angles. This could be due to the force of gravity: If something wishes to rise towards the sun, for example a tree, then its main body will have to sustain the least static stress when growing straight up, balancing out the gravitational pull, therefore resembling a vertical structure, which results in a horizontal frequency. On the other hand, when something is laying down, it will usually follow the shape of the ground, which in most cases will be flat, parallel to the horizon, resembling a vertical frequency. The effect is emphasized even more in images showing man-made objects, as these tend to include sharp and rectangular edges (e. g., windows and balconies of buildings etc.). It has been proposed that this prominence of vertical and horizontal structures may have allowed for a specialization of the visual system enabling us to efficiently process scenes that resemble natural images, as these are the ones we need to process most frequently in everyday life. There are, for example, more neurons in the visual cortex that respond to horizontal and vertical orientations, than oblique ones [DeValois & DeValois 1988].

Torralba and Oliva have also found manifestations of several other scene aspects in the amplitude (or power) spectrum, like openness, average depth and so on [Torralba, Oliva 2003]. The common spectral properties of natural vs. man-made scenes have been thoroughly described in [Torralba, Oliva 2003], as shown in the figure below.

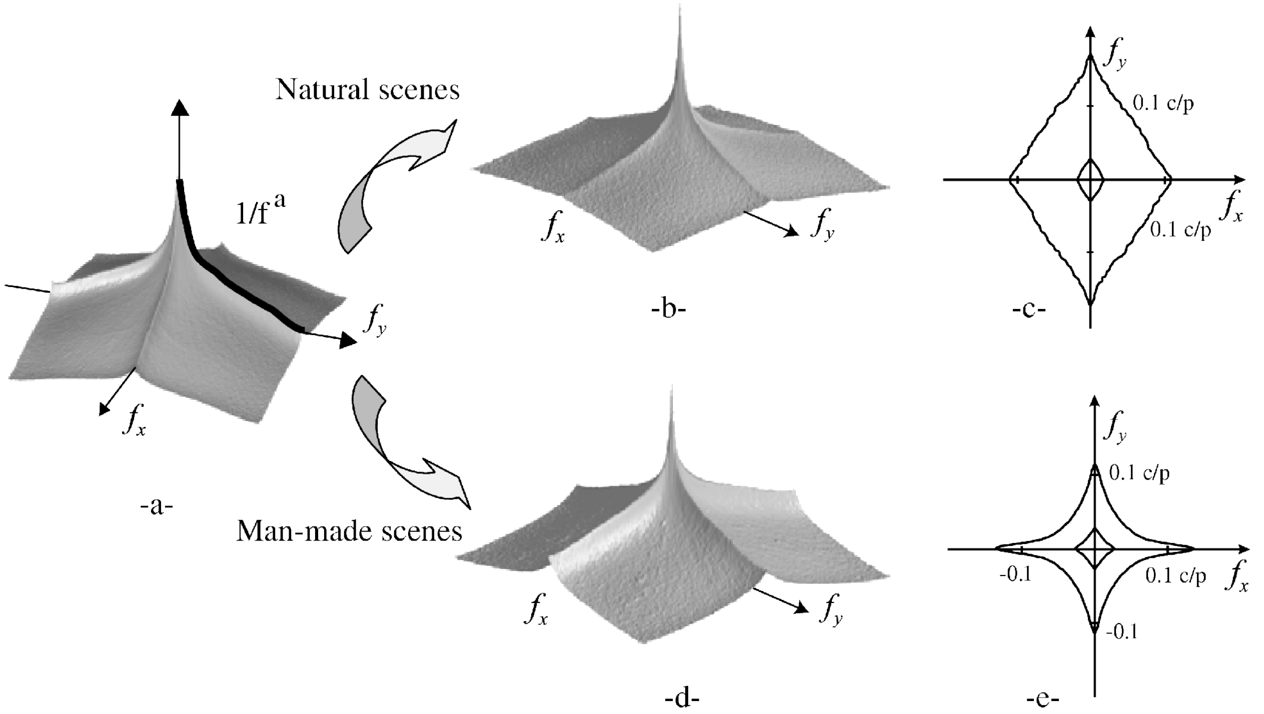

Figure 2. (a) Mean power spectrum averaged from 12 000 images (vertical axis is in logarithmic units). Mean power spectra computed with 6000 pictures of man-made scenes (b) and 6000 pictures of natural scenes (d); (c) and (e) are their respective spectral signatures. The contour plots represent 50 and 80% of the energy of the spectral signature. The contour is selected so that the sum of the components inside the section represents 50% (and 80%) of the total. Units are in cycles per pixel (cf also Baddeley 1996).

Illustration 11: Mean power spectra of natural scenes (taken from [Torralba, Oliva 2003])

The amount of general similarity between images of the same class can be illustrated by averaging the spectrum of many images of one class into spectral signature-like contour plots. The differentiation among various classes of man-made scenes shows a strong bias towards horizontal and vertical frequencies, while natural (here: meaning not man-made) scenes show a broad variation in spectral shapes.

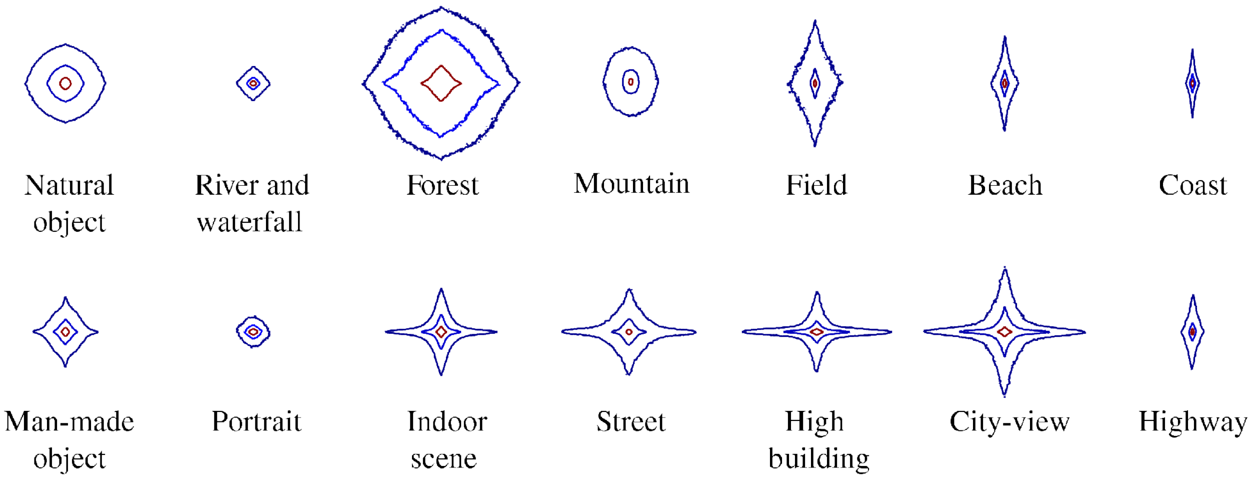

Figure 3. Spectral signatures of 14 different image categories. Each spectral signature is obtained by averaging the power spectra of a few hundred images per category. The contour plots represent 60, 80 and 90% of the energy of the spectral signatures (energy is obtained by adding the square of the Fourier components). The size of the spectral signature is correlated with the slope (α). A large value of α produces a fast decay of the energy at high spatial frequencies, which produces a smaller contour. The overall shape is a function of both $\alpha(\theta)$ and $A(\theta)$.

Illustration 12: Spectral signatures of different image categories (taken from [Torralba, Oliva 2003])

In a recent article, Johnson and Ohlshausen [Johnson, Ohlshausen 2003] report another finding about a difference in the average spectral signature of two classes of images: images of animals and images of nature without animals (see Illustration 13). Images containing animals appear to have a more even distribution of power among different orientations, while the strong peak in vertical frequencies (or horizontal edges) in both animal and non-animal images is claimed to be related to the appearance of the horizon in many natural images. These findings support those in [Torralba, Oliva 2003] and suggest that the spectral differences between the two classes of images might be used to determine the category of an otherwise unknown scene through an evaluation of its spectral shape.

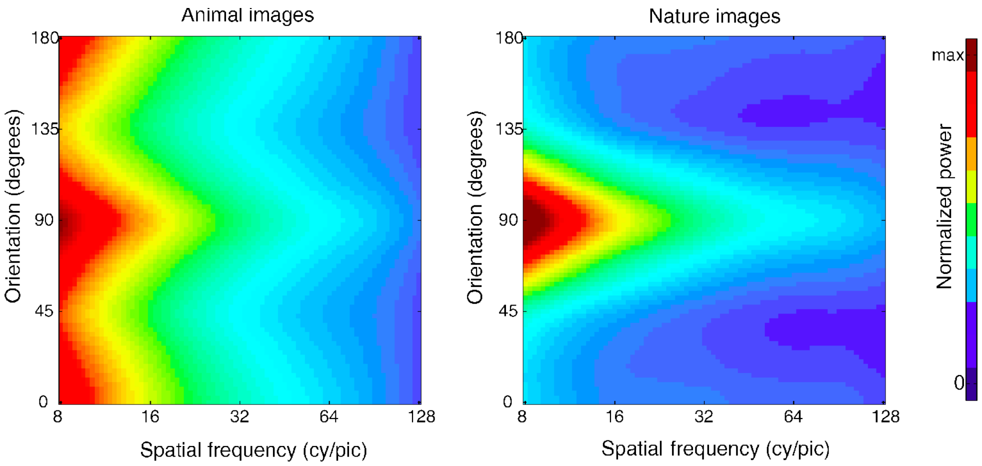

Illustration 13: Mean amplitude spectra of "Animal" and "Nature" image (Taken from: [Johnson, Ohlshausen 2003])

1.7. The Black Box Approach

The challenge of understanding aspects of the human visual system resembles a black box problem: there exists a system of unknown inner functionality $o=f_b(i), i\in I, o\in O$, which is capable of transforming a set of input data I into a set of associated output values O .

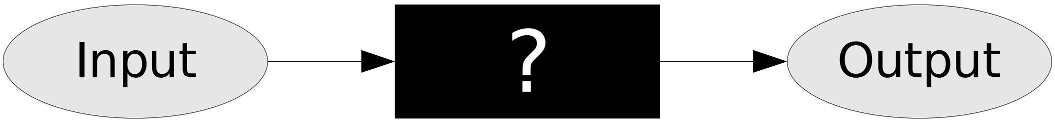

Illustration 14: Symbolic blackbox system

In our particular case, I is comprised of all possible natural scenes, whereas O is comprised of only the two possible labels “animal” and “non-animal”. The black box itself, the human visual system, can not be disassembled to discover its working mechanism; therefore, the only way to derive an equivalent to its inner functionality f_b is to find an emulation function $o=f_e(i), i\in I, o\in O$ such that $\forall i\in I_b : f_e(i)=f_b(i)$ (any given input value must always be translated into the same output value by both f_b and f_e). The emulation function is not necessarily identical to the black box function; however, it is mathematically equivalent. In many complex problems, like image classification, there is one major problem that prevents us from comparing input and output in their entirety: for all practical purposes, the input set is of infinite magnitude. Also, the intricate mechanisms of the human visual system are extremely sophisticated and show a high degree of inter-individual variance; discovering a perfect emulation function seems very unlikely. We will have to content ourselves with an emulation function that exhibits at least a reasonably high degree of similarity with the human visual system.

1.8. The Task Ahead

To proceed, we will first need to assemble a suitable subset of the input set $I_s \subset I$; this will be done in chapter 2. Then, we need to develop an emulation function and evaluate it on the input subset; this will be done in chapter 3. Last, we need to compare the output of our emulation function with the output of the original black box to determine whether our emulation function behaves similar to the original one; this will be done in chapter 4.

2. THE IMAGE DATABASES

As a first step in our attempt to emulate the black box functionality, we need to assemble a suitable set of test images that can be used in experiments involving visual presentation with human subjects as well as computer / algorithmic classification trials. The following chapters introduce the two sets of images used in this work.

2.1. The APG Database

The images used in the classification experiments were taken from the formerly commercially available Corel Stock Photo Library, which contains about 60000 images of both natural and man-made objects and scenes. For the APG animal / non-animal database, roughly 11000 images were selected, one half of which contained animals (and possibly humans, but never humans alone) and served as targets, the other half of which contained no animals and no humans and served as distractors. Out of the Corel Stock Photo Library, the original images measure 768x512 pixels, in either portrait or landscape orientation, and are stored as individual JPEG image files (see Illustration 15, step I). To facilitate the application of the digital Fourier transform that most of the following sections make use of, a square-shaped section of the rectangular original image was cropped in such way that the scene's main object, if there was one, was centered (step II). The resulting square images were of size 512x512 pixels and were cut by 16 pixels on either of the four borders to remove boundary artifacts (black bars that are not part of the actual scene), which were apparently introduced during the original scanning process (step III). The remainders of the images were then shrunk down to 256x256 pixels to limit processing time (step IV). This also helps to minimize JPEG artifacts. In some of the psychophysics experiments, the images were shown in color. In all of the computational experiments, grayscale versions of the images were used (step V, see Illustrations 15, 16 and 17).

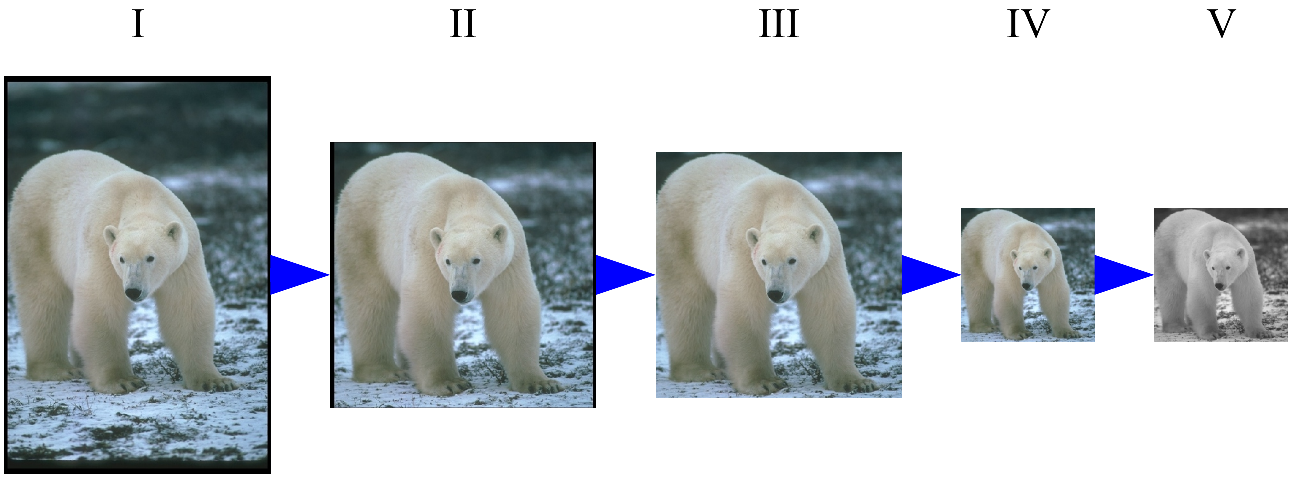

Illustration 15: Basic image preprocessing

2.2. Samples from the APG Database

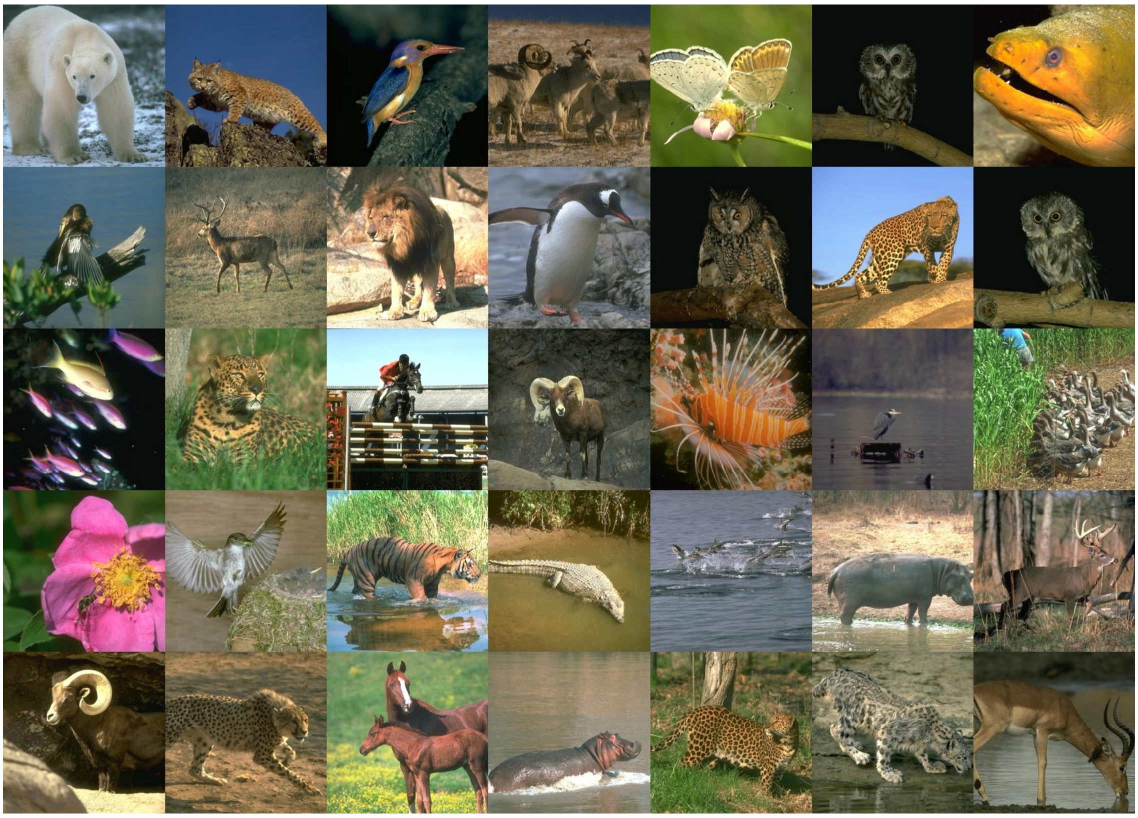

Illustration 16: APG database, samples of animal images (as in step IV of Illustration 15

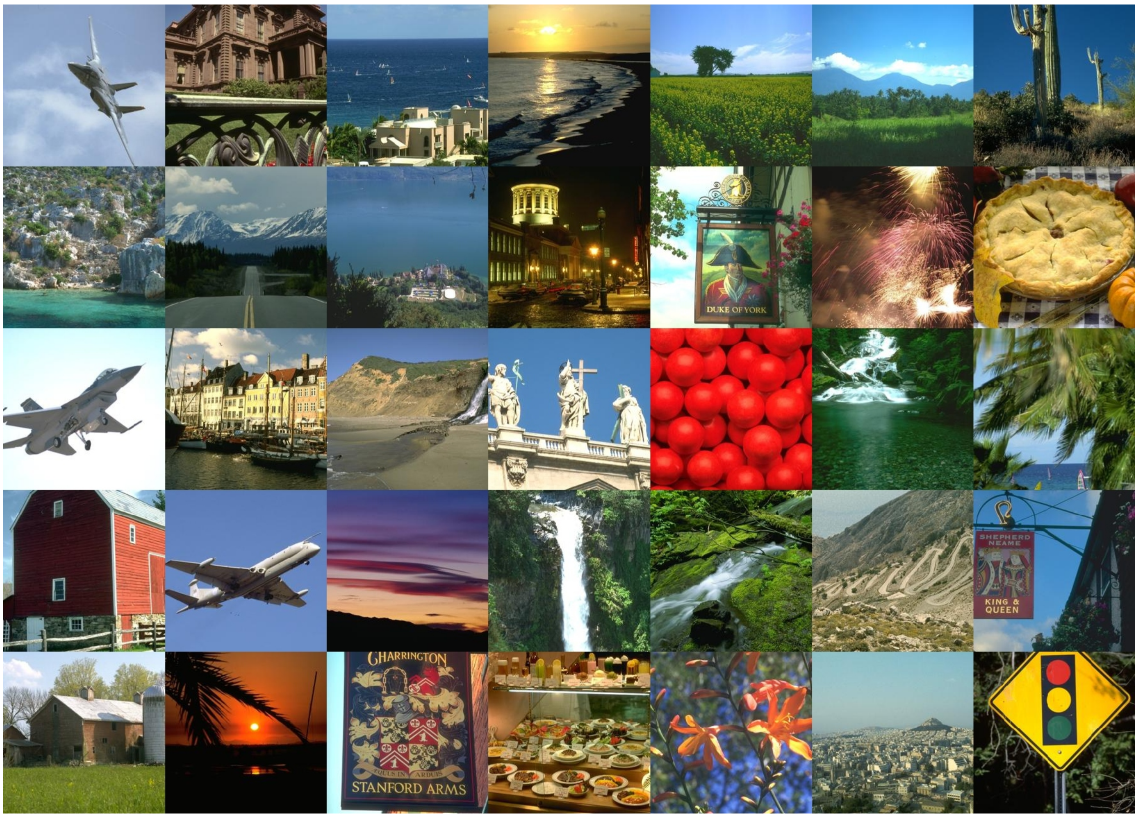

Illustration 17: APG database, samples of non-animal images (as in step IV of Illustration 15)

2.3. General Statistics of the APG Image Database

To obtain a general idea of the differences between the two image classes, some aspects of general image statistics were analyzed. We started with averaging the actual pixels of our images, separately for the “animal” and the “non-animal” class. This was performed by adding up the intensity values independently for all 3 color channels (Red, Green, Blue) of every pixel of all images, then dividing the results by the number of images. As the resulting image would be a monotonous medium gray due to the averaging effects, all 3 color planes of the resulting mean images were scaled with a common scaling factor to span the full range of possible RGB values (so-called “24 bit” RGB images allow for 256 shades per color channel, spanning the value range of an unsigned 8-bit integer $[0\ \ 255]$).

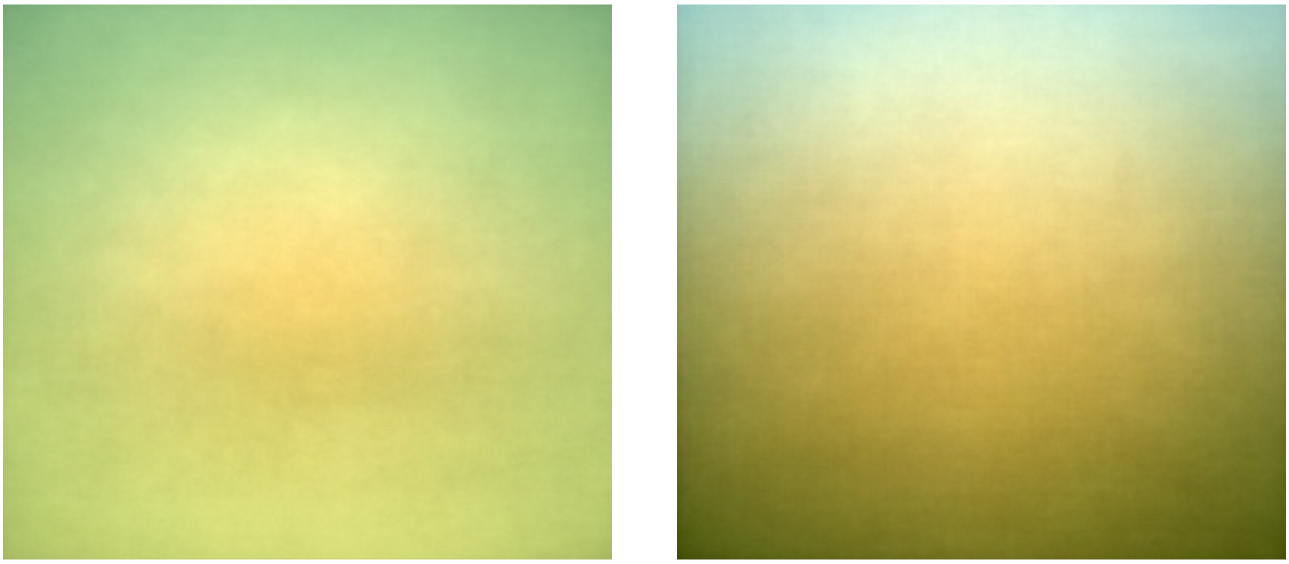

Illustration 18: Per-Pixel averages of the image database. To the left: animal images, to the right: non-animal images

Both mean images show an orange-brownish blob at the center of the image, surrounded by a mostly brownish green with a slight shade of blue at the top, which appears to be somewhat stronger on the “non-animal” category. The blue tones in the top corners can be assumed to be generated by the sky, depicted in a (dominantly) large number of images in both classes. The greenish-brown at the bottom of the images would be the ground, and the orange-brownish blob in the center of the image represents the main object of the scene, which was arranged to be mostly at the center of each image when the images were cut to squares (see chapter 2.1). The mean “non-animal” image appears to be a bit darker towards the bottom and at the corners, which might be an effect of the photographer's camera settings.

As a second aspect, the amplitude spectra of the images were analyzed. In their network paper, Torralba and Oliva reported a particular difference in the slope of the amplitude spectrum of “natural” vs. “man-made” scenes [Torralba, Oliva 2003]. The definition given in their work declares a scene “natural” if its contents depict merely naturally occurring objects and areas, whereas “man-made” scenes contain any kind of human-built structures or otherwise artificially created environments. We analyzed our image categories “animal” and “non-animal” in a similar manner, finding the same difference in spectral slope that Torralba and Oliva did, even though our image categories differ in their definition. On average, both image classes showed more energy on the horizontal and vertical axis than on the oblique angles. This difference in energy distribution is significantly stronger in the “non-animal” image class (see Illustrations 19 and 20). For both image classes, the mean slope of the amplitude spectrum approximately follows the well-known 1/f distribution commonly found in natural images.

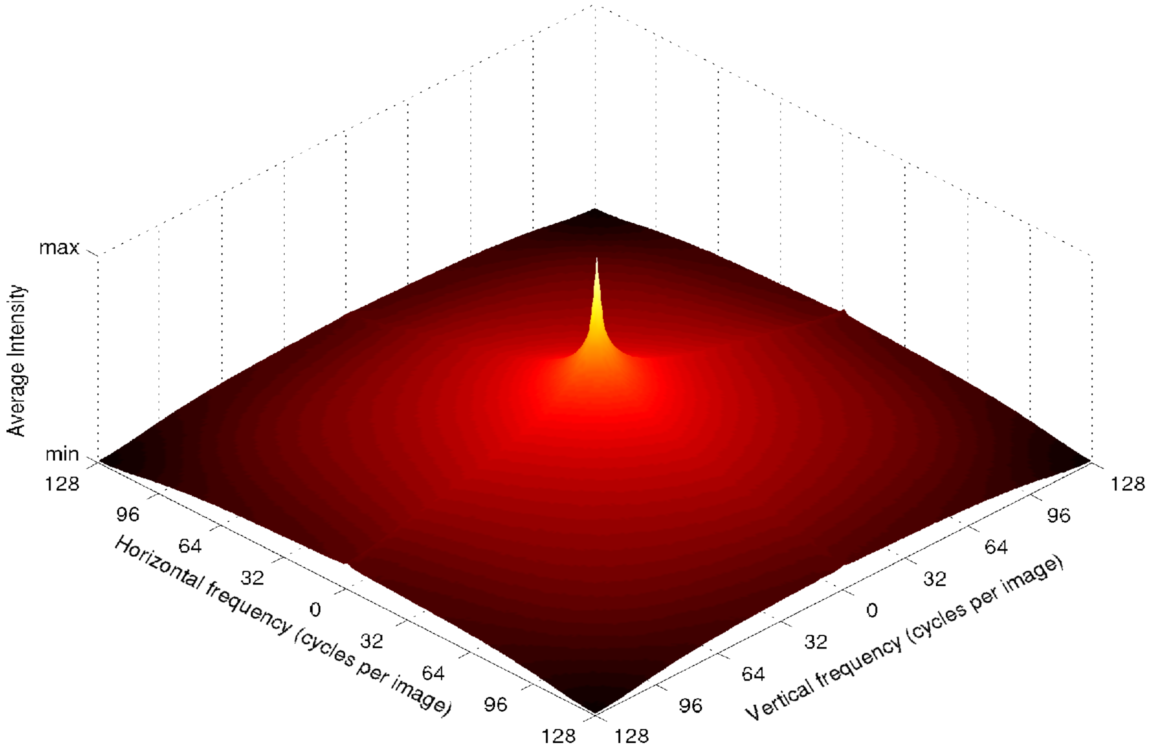

Illustration 19: Mean amplitude spectrum of animal images

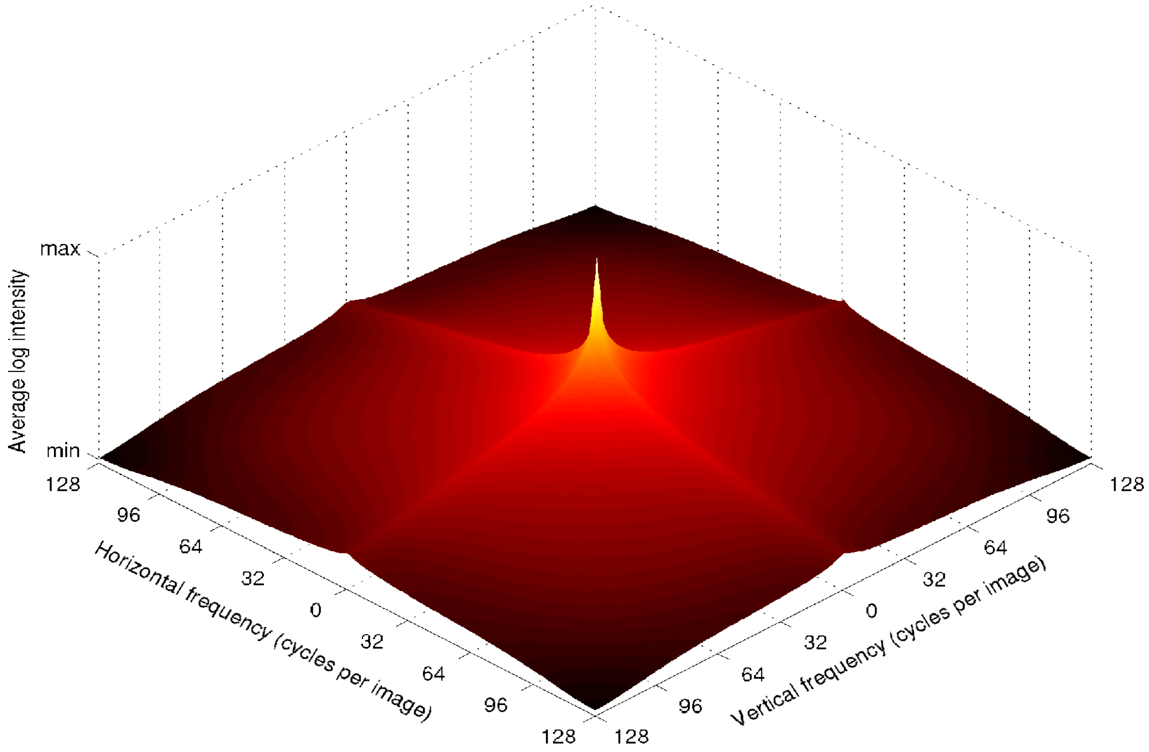

Illustration 20: Mean amplitude spectrum of non-animal images

2.4. The Torralba Database

The image database kindly provided by A. Torralba[1] contains 1200 images, one half of which show various animals, the other half showing no animals (see Illustrations 21 and 22). The images have been selected and cropped from the Corel Stock Photo Library, the same library our own "APG" image database was selected from. Though this means that most original images that have been used in the image collection by Torralba have also been used in our own image database, none of the images are truly identical due to different ways of cropping and resizing (see chapter 2.1). The selection of images has been balanced to include an equal number of images from four distances (called "head", "body", "medium" and "far") and two types of environments ("natural" and "man-made"). Animals, when present, appear to be centered most of the time.

[1] Antonio Torralba, PhD; MIT CSAIL, MIT 32-D462, 32 Vassar Street,Cambridge, MA 02139, torralba@csail.mit.edu

2.5. Samples from the Torralba Database

Illustration 21: Torralba database, samples of animal images

Illustration 22: Torralba database, samples of non-animal images

2.6. General Statistics of the Torralba Database

Not surprisingly, the per-pixel averages of the Torralba image database look very similar to our own – though more coarse, due to the smaller number of images used. The difference in luminance and overall contrast compared to the images in our own database is a result of the normalization process applied to coarser images. We find the mean amplitude spectra of the images to be very similar to those of our own APG database as well; they follow the approximate 1/f curve, as has already been shown extensively in [Torralba, Oliva 2003].

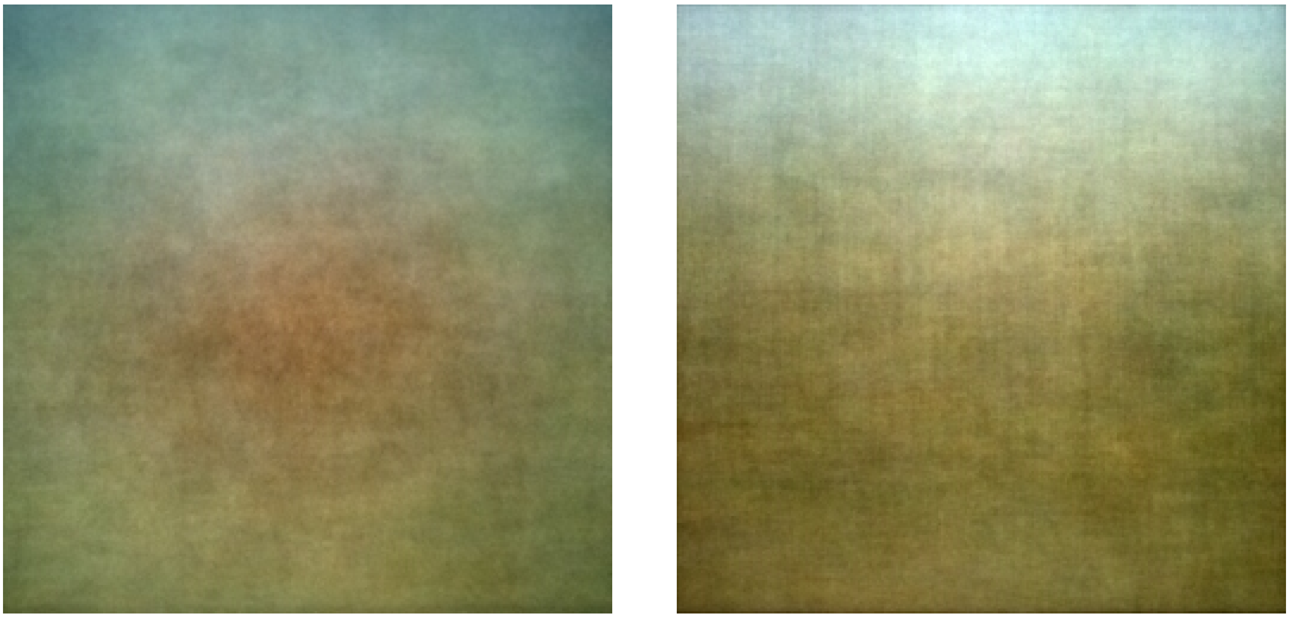

Illustration 23: Per-Pixel averages of the image database. To the left: animal images, to the right: non-animal images

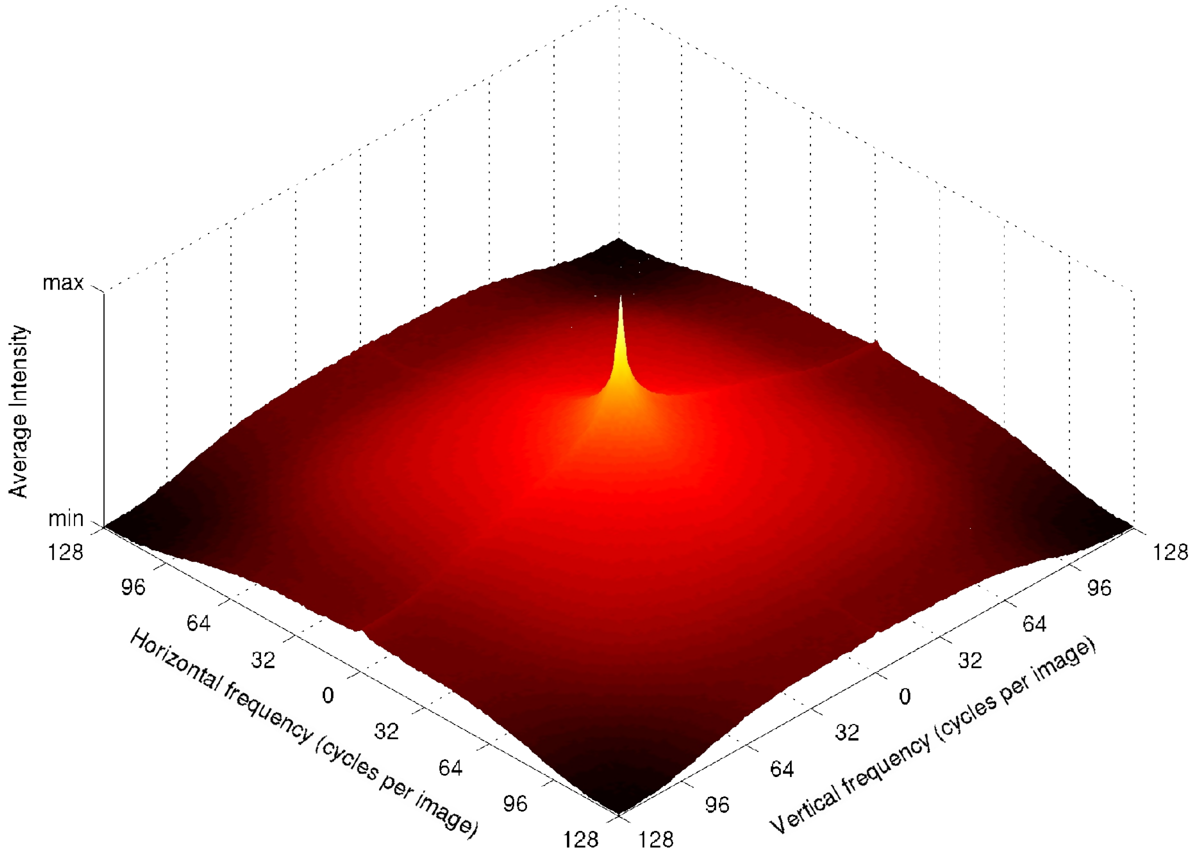

Illustration 24: Mean amplitude spectrum of animal images

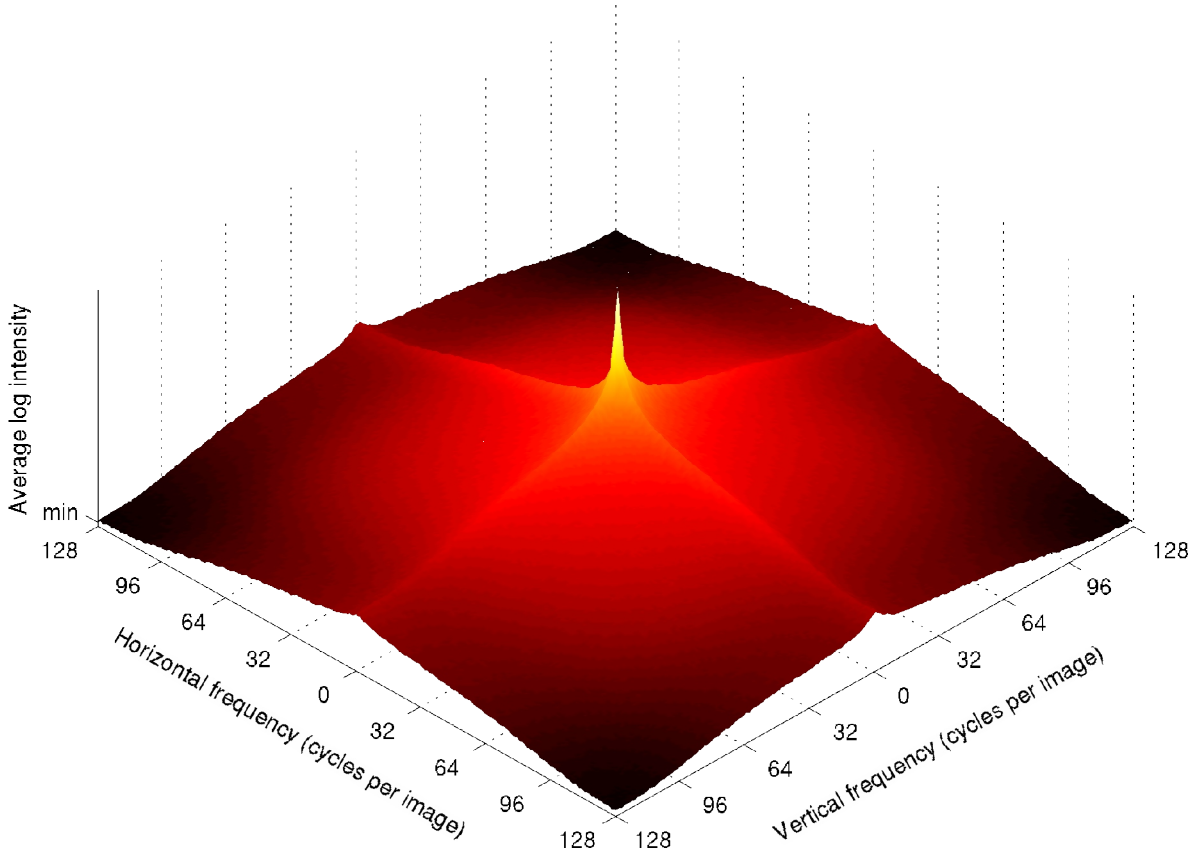

Illustration 25: Mean amplitude spectrum of non-animal images

3. COMPUTATIONAL CLASSIFICATION

As the second step in our attempt to understand the black box functionality, we need to develop an appropriate emulation function and test it on our image database. Attending this need, we now present results from our computer-based classifiers performing the before mentioned task of animal vs. non-animal discrimination. We start with the simplest form of classifier using only raw pixel information as input. Subsequently, more complex data reduction tools are tested and even preprocessing steps are applied; also, a more sophisticated classification mechanism is tested. We also compare our own image database against the one provided by A. Torralba, and we evaluate the classification procedure proposed in [Torralba, Oliva 2003] on our APG database.

Whenever classification performance is to be measured, one usually expects results to be somewhat less than perfect. The interesting part is to see how close to perfection the results really are, and, in particular, whether one approach can yield significantly better results than another.

In computer vision, the key to successful classification is usually to find a suitable way to reduce the dimensionality of the original image data to an amount one can handle with the available resources, while retaining as much as possible of the information required for the classification. This is even more difficult to realize as the “information required for classification” is not clearly known in the type of data used in this work. An important factor to be considered for the adjustment of the dimensional reduction step is the number of training samples available. The higher the number of training samples, the higher the number of dimensions that can be learned successfully. In the “learnable” (or trainable) classification mechanisms applied in this work (linear discrimination analysis and support vector classification, see chapter 7), the number of training samples required for a given number of data space dimensions can not easily be predicted ahead of time, though one will usually need at least several times as many samples as there are dimensions. As a baseline, we used the pixels of the images directly for classification, with the only data reduction applied being a reduction of image size.

All evaluation of data and all computations were performed in Linux (mostly Debian), using Matlab 7 Release 14, with varying service packs (No. 1-3) on a variety of x86-32 and x86-64 systems. Some computations were performed using a number of lab PCs as a computation farm, some were performed using the “Two Towers” computation cluster at the Max Planck Institute for Biological Cybernetics, Tübingen.

3.1. Classification and Supervised Learning

The fact that there exist differences in the spectral properties of different image categories suggests that these differences could be used to determine the category of an otherwise unknown scene through an evaluation of its spectral profile. This leads us towards the field of computer classification. With modern machine learning techniques, the general approach is relatively straight forward: First, the selected classification algorithm is trained on a known database. “Known” in this case means that all the samples (images) in this database have been correctly labeled, maybe by the operator or, with artificial data, during their original generation. The classification algorithm is supposed to derive from the training examples a compilation of the necessary information that can be used to tell the different classes of samples apart, so that future, unknown samples can be related to the compiled knowledge and therefore be (hopefully) correctly classified, even without carrying a preassigned label. This general procedure is called “supervised learning”. One of the advantages of this procedure is that the classifier is assumed to automatically recognize the relevant information, thereby sorting out irrelevant or misleading portions of the training dataset. The main problem with complex types of data, such as images, however is that the amount of data necessary to perform a certain classification task can be anything between minimal (e. g. actually requiring just one or very few pixels of the image, $O(1)$) or huge (e. g. requiring every single pixel of the image, $O(N)$, N being the number of pixels in the image). The minimal amount of training data required to successfully learn a classifier is directly dependent on the amount of data necessary to actually describe the difference between the classes. At least two samples are required to perform a linear discrimination (or to learn a linear classifier on a very low level); however, this would be the ideal case were the two sample images would perfectly mark the border between the two subsets, a case which is unlikely to happen with non-synthetic data – and if it would happen, the “learning” of the classifier would be unnecessary because the plane of discrimination would have to be perfectly clear already in order to pick the right pair of samples. In a realistic scenario, it is much more likely that the samples will be distributed in a pseudo-random fashion, with no or very few “obvious” differences between the classes. In fact, many of the more complex classification algorithms assume a Gaussian distribution of the samples. Due to the random-like distribution, one cannot say for sure whether a certain number of samples in the training set will be sufficient to successfully learn an arbitrary classifier. It is therefore always desirable to have as many samples as possible for training; usually, in complex problems at least several times as many samples as there are dimensions in the feature space are considered to be appropriate. When working on high-dimensional items such as images, this can easily exceed the volume of even the vastest of training data collections. It is therefore imperative to reduce the size of the data that is actually used for classification. This is commonly referred to as “dimensionality reduction”.

3.2. Reduction of Dimensionality

The key issue in dimensional reduction is to leave away only such information that does not concern the problem at hand. Simply cutting away an arbitrary number of the pixels of an image is generally not a good idea. The main effort in the development of a well-performing classifier is therefor to identify the properties of the data samples, called “features”, that are relevant for classification and to compile a compact set of these features.

In the kind of classification problems that is to be analyzed in this context, one usually does not know the nature of the features relevant to the problem, at least not exactly. As it is therefore impossible to simply isolate and extract the desired features in some straightforward, deterministic way, one needs to turn to more general, holistic approaches in dimensional reduction. One particularly useful and well-known technique is called “Principal Component Analysis” (PCA). PCA is used to recompute the axis of the coordinate system of the data space in such way that the first axis will be along the direction of the greatest variance of the data, the second axis will be along the direction of the greatest variance after the first axis has been eliminated, and so forth. The result of a PCA will typically be a set of orthogonal (and thus linearly independent) vectors describing the new axis system, and a set of weights defining the original data points transformed into the new coordinate system. The consideration now is that the higher the number of an axis, the lower the amount of variance covered by it, and so the loss of information will also be lower should this axis be ignored in the computations to follow. Using the first n principal components is therefore a way to reduce the dimensionality of the data to n while maintaining a data representation that still covers most of the original information. PCA has been extensively and successfully used in image compression and many other fields of digital signal processing.

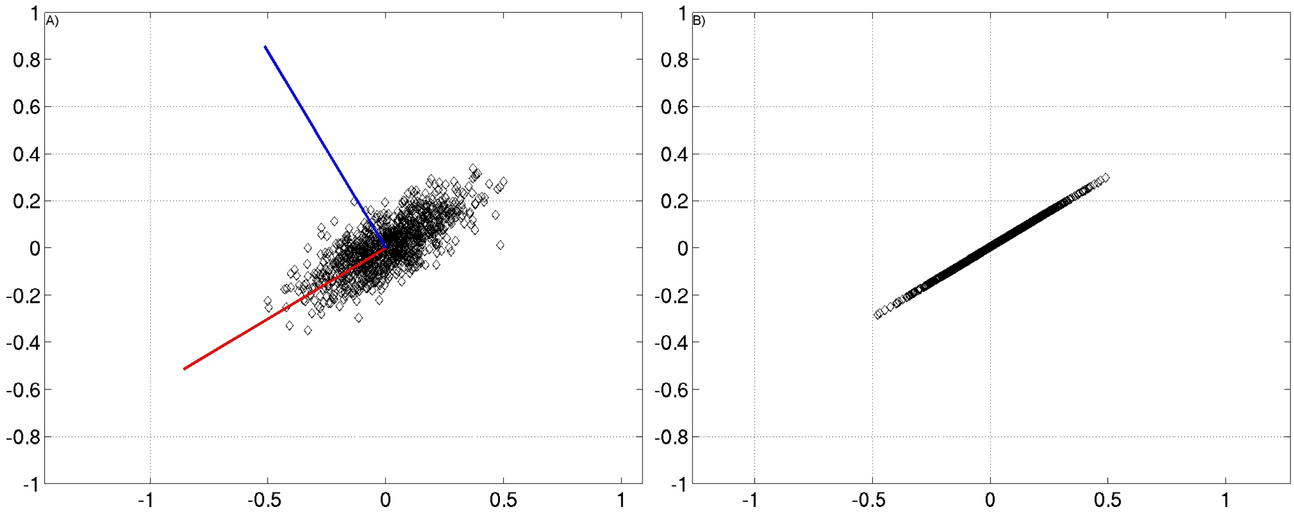

Illustration 26: PCA on a random 2D sample dataset.

Figure A) illustrates a pseudo-random dataset which is not aligned with the axis of the coordinate system. The colored lines are the new primary (red) and secondary (blue) axis (the principal components). The primary one clearly covers the larger part of the variance within the dataset. Figure B) shows the appearance of the dataset after elimination of the secondary axis – the dimensionality has been reduced from 2 to 1, cutting the amount of data to process in half, while still retaining the larger portion of the variance within the dataset.

In datasets of higher dimensionality, a printable representation of the data space can be very confusing. Likewise, the representation of the principal components is not always possible in an intuitive way. In the processing of 2-dimensional images, however, the principal components can be displayed as images themselves.

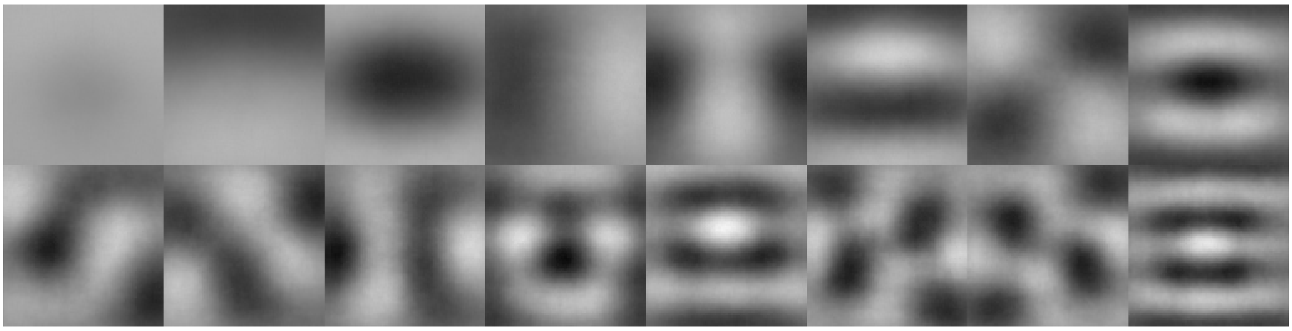

Illustration 27: The First 16 principal components of natural scenes

Illustration 27 shows the first 16 Principal components of natural images (see chapter 3.3.2). Some of these principal components of natural scenes exhibit a striking similarity to some of the receptive fields discovered in the visual system of primates and humans. The fact that these apparently regular structures actually have some kind of meaning can be shown when comparing them to the principal components of purely random data, which by themselves appear totally random.

All or an arbitrary subset of these components may be shown inverted as they represent the axis of a coordinate system; inversion of an axis can easily be compensated by flipping the sign of the corresponding weight for each data sample. It can be argued what the individual components actually represent – or if they even represent one abstract scene aspect at all. For example, it would appear plausible that the very first component is a DC component, while the second one could be related to the horizon; the third one might actually be a vignette-effect introduced by the lenses of the cameras used to capture the images, or it might simply be related to the fact that the core element of the depicted scene was centered in most images.

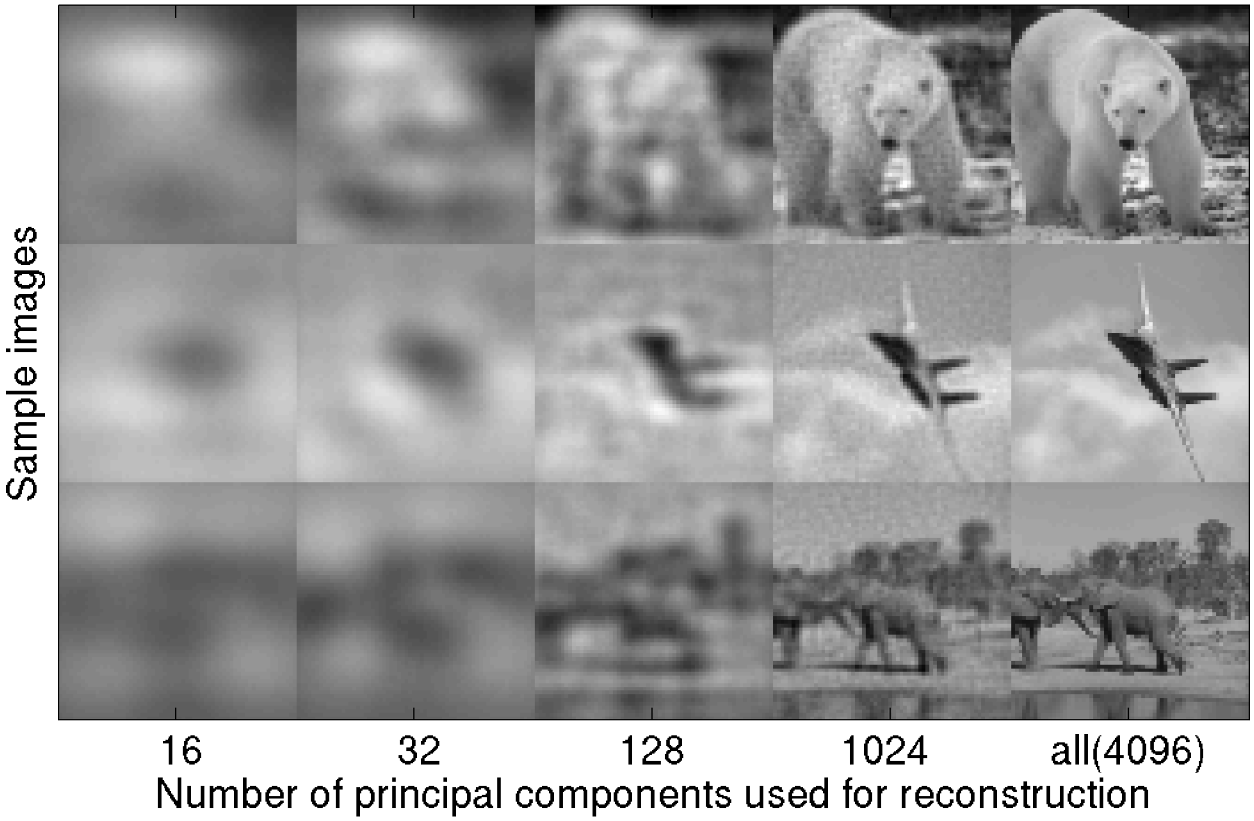

Illustration 28: Images reconstructed from various numbers of principal components

Just like images, the power spectra of natural scenes can be decomposed into their principal components. These represent elemental groupings of frequencies and orientations, and unlike the image principal components they do not contain spatial information and thus are invariant to the position of objects within a scene; they therefore do not resemble anything like the horizon or a centered vignette like the image principal components do, but instead they represent general aspects of the entire scene. In the example from [Torralba, Oliva 2003] shown below the second and third spectral principal components are used to successfully organize images along their "openness" and “naturalness” axes.

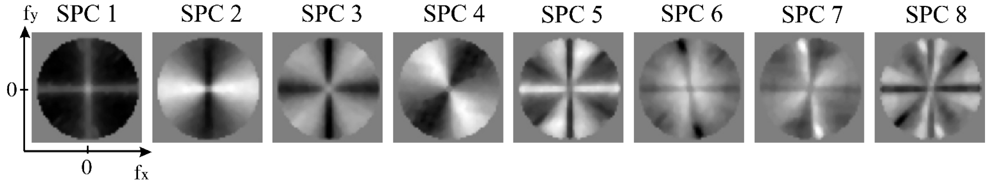

Illustration 29: Spectral Principal Components, from [Torralba, Oliva 2003]

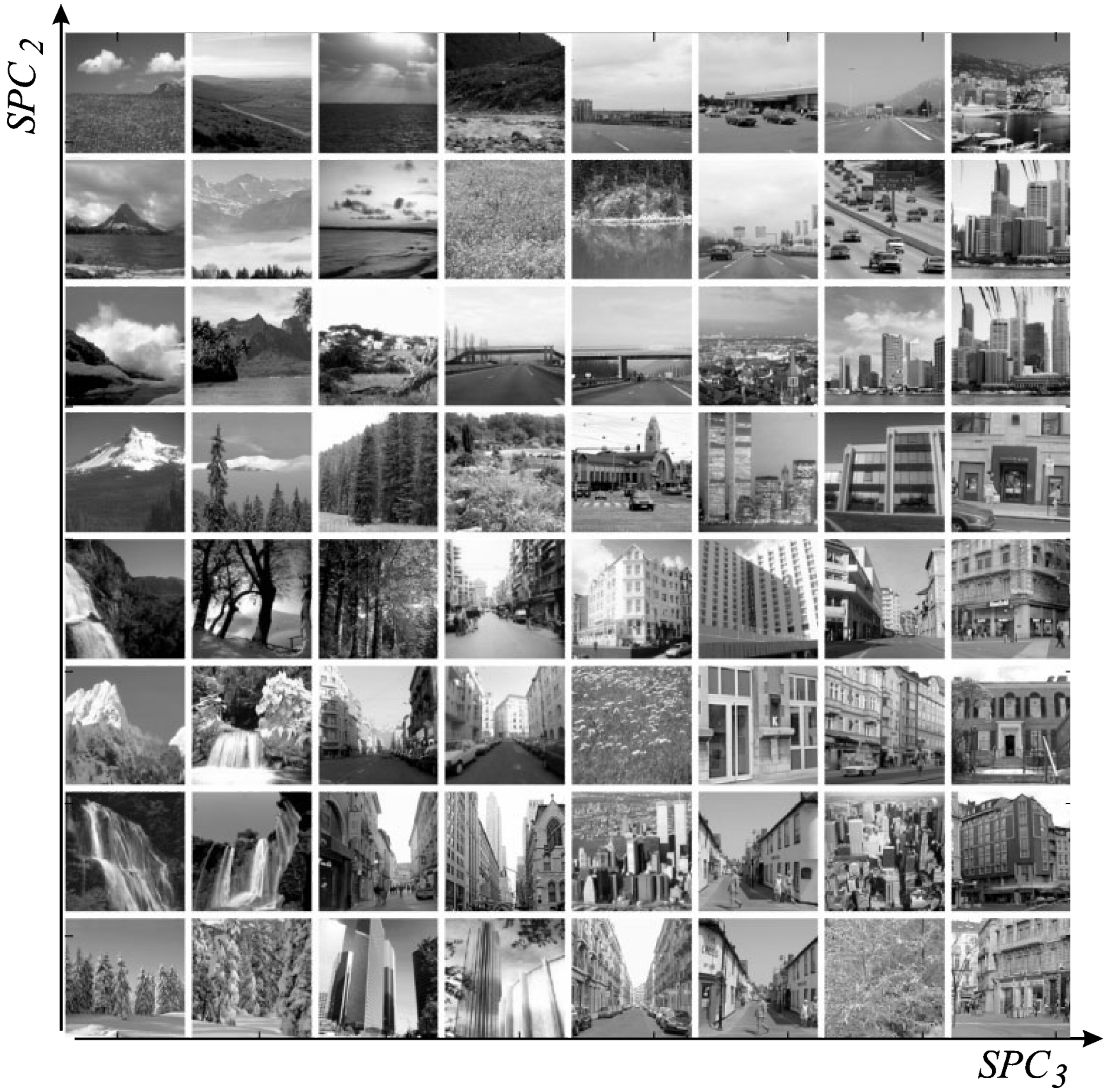

Figure 9. Projection of images into the space represented by the second and the third principal components of the power spectra. Images are organized according to spectral properties: SPC$_2$ puts images with dominant energy in the f_y axis on top of the figure opposed to images with dominant energy in the f_x axis which are at the bottom. SPC$_3$ opposes images with energy in the f_x and f_y axis (cross shape) with respect to images with energy at oblique orientations. A coarse organization of scenes emerges: man-made versus natural scenes and open versus closed environments.

Illustration 30: Organization of scenes with spectral principal components, [Torralba, Oliva 2003]

Employing these techniques, [Torralba, Oliva 2003] have been able to tell images containing animals apart from images containing no animals with an average accuracy better than 80% using only the first 16 spectral principal components.

3.3. Classification in the Spatial Domain

When humans classify images, they look at the plain pixels of the image. As humans are very successful at classifying images, the information necessary for classification has to be encoded somehow into the intensity values of the pixels of our images. As a baseline for naivety, one may try to use these intensity values directly for an attempt at classification. Naturally, the dimensionality of the data space will be extremely high, with the largest images (256x256) even exceeding the number of sample images available to us by a factor of 6 (10864 images with 65536 pixels each). Considering the number of samples usually required to successfully learn a classifier (see chapter 7), it will not be surprising to find classification performance to be just above chance level.

3.3.1. Classification on Plain Images (Raw Pixels)

The only preprocessing applied to our images was a per-pixel scaling of the intensity values to the interval [0 1]. Thereafter, the intensity value of each pixel gave one of the dimensions of the input space. The dimensionality of the input space was therefore directly dependent on image size. Images of different sizes were tested as a (naive) means of dimensional reduction to at least partially compensate for the huge number of dimensions involved (see Illustration 31). We used a linear classifier with a 20-fold cross-validation.

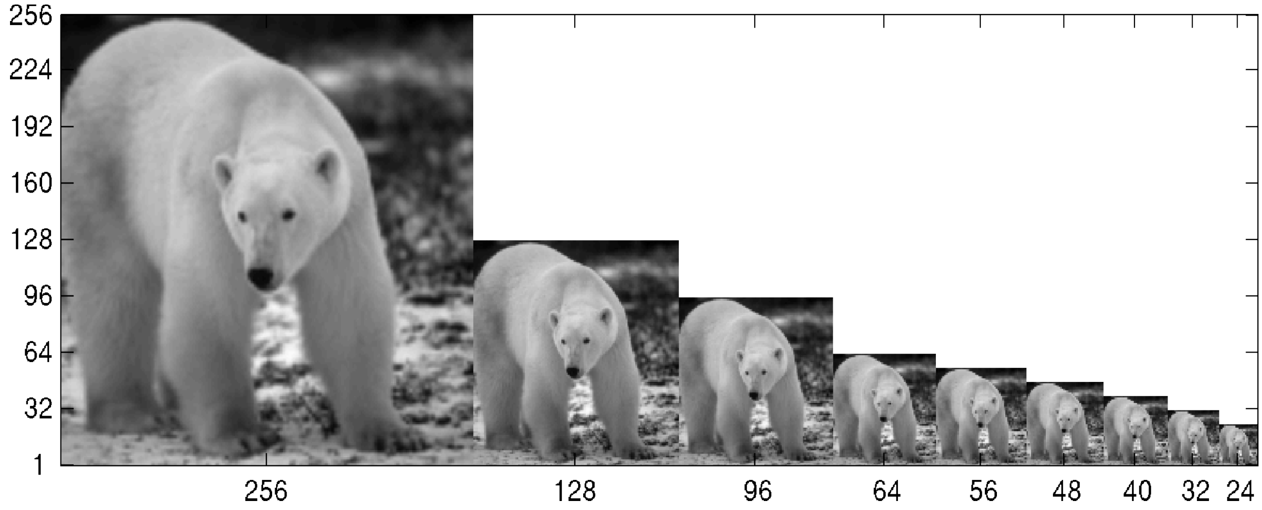

Illustration 31: Sample image in different sizes, as used for classification

Image size in pixels	*Resulting dimensionality*	*Image size in pixels*	*Resulting dimensionality*
256x256	65536	48x48	2304
128x128	16384	40x40	1600
96x96	9216	32x32	1024
64x64	4096	24x24	576
56x56	3136		

Table 1: Image sizes and resulting dimensionality of data space

3.3.1.1. Results

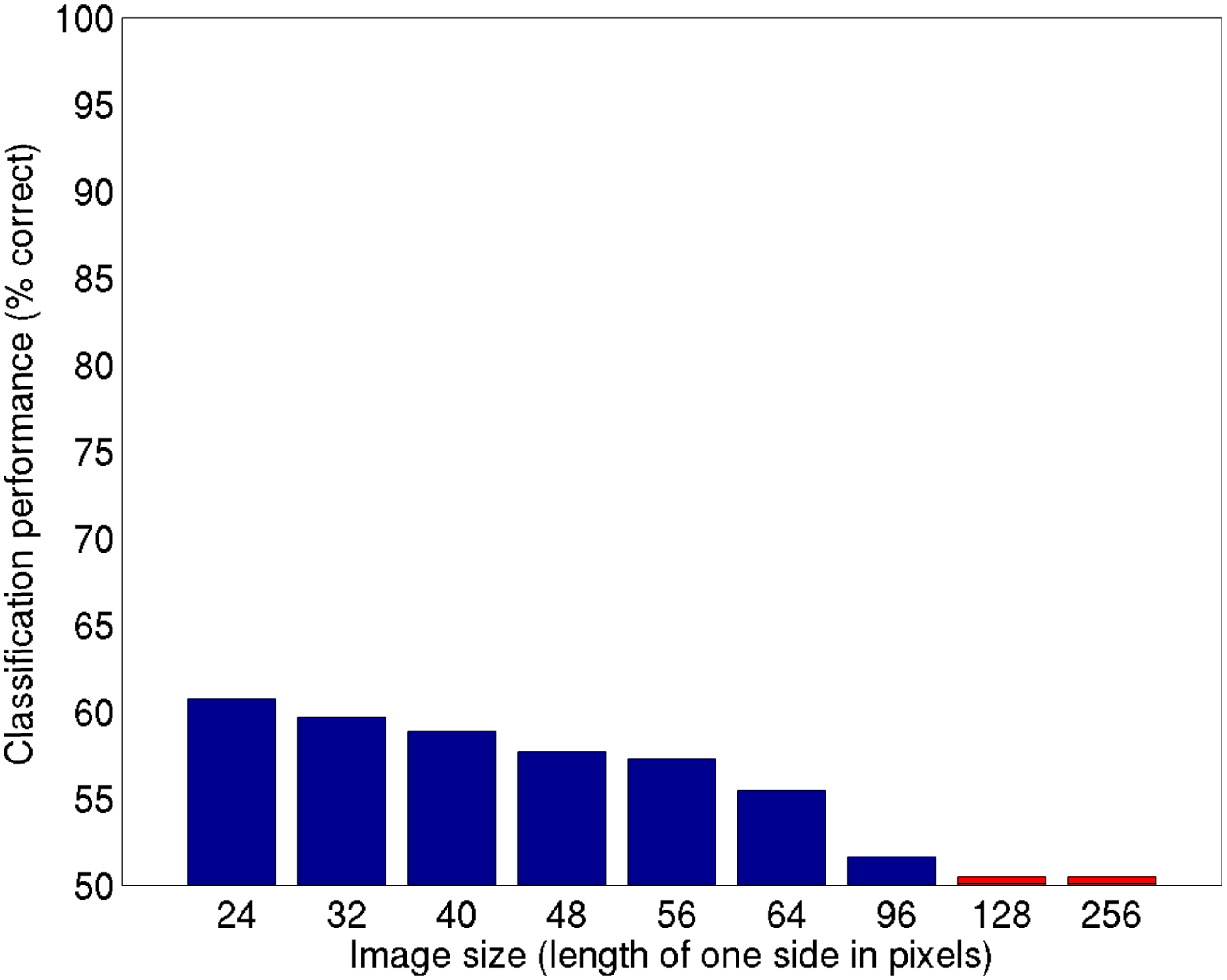

Illustration 32: Classification performance on pixel intensity values of various image sizes

Our results show that classification performance reaches just over 61% with the smallest images (24x24 pixels), then continuously declines with increasing image size, finally reaching 52.6%, which effectively is chance performance, at 96x96 pixels image size.

With image sizes 128 and 256 a computational classification was not possible, as the covariation matrix that the linear classifier uses, differed from zero only to the extent of machine precision, prohibiting the classifier from computing any results – which would have resulted in chance performance, for just the same reason, anyway. We therefore added the hypothetical results to the above plot, shown in red, for the reader's convenience.

Despite of the poor results, this does not actually prove that the pixel intensity values can not at all be used for successful computer classification; it does not even prove that it can not be carried out with our linear classifier. The reason for this is the following:

When assuming that the information necessary for classification is distributed evenly across all dimensions of the data space, as we must since we have no reason to assume otherwise at this point, the number of samples needed for successfully learning a classifier is directly proportional to the number of dimensions of the data space. Generally one assumes his data to be approximately of normal distribution. The number of data points necessary to reliably estimate the true parameters of the distribution depends on the mostly unknown noise level and can therefore not be safely determined beforehand. At least several samples are required, sometimes several hundred or even thousands. Taking that aspect into account, at least several times as many samples as there are data space dimensions are required to successfully train a classifier – the more samples, the more reliable the results of the training will be. When considering the large variance of our dataset (see chapter), one can not expect successful classification with only a few samples per dimension. It is impossible to predict precisely how many dimensions of data space are to be considered feasible, but it seems highly unlikely that any number above 1000 dimensions (resulting in roughly 10 samples per dimension for training) can be successfully used to classify our dataset; a dimensionality below 100, however, is to be preferred. We conclude that the raw pixel intensity values will not allow us a successful computational classification with the available sample data.

3.3.2. Classification on Principal Components of Plain Images

The attempt to classify on the untreated pixels of our natural scenes has not lead us to a successful classification, the most obvious reason for this being the huge number of dimensions to deal with.

Whenever a dataset like our natural scenes, which could be seen as being of “unexplainable” structure, is to be reduced in its dimensionality, one can not simply eliminate individual dimensions (meaning pixels in our case) at random: the individual importance of every one of these dimensions is unknown.

A commonplace approach to overcome this challenge is to re-code the dataset into a different coordinate system, hoping that the new coding will help to understand the nature of the data. One such recoding is the well-known Principal Component Analysis, which recodes a dataset into an orthogonal coordinate system with axes chosen by order of occurring variance (see PCA chapter). Since principal components are ordered by the amount of variance covered by each component, it is possible to cover a large part of the variance in the dataset with only the first few components, omitting the remaining dimensions and thus reducing the effective dimensionality of the data space significantly while still retaining a description for most of the differences between our images. In other words, we can greatly reduce dimensionality while only slightly reducing informational content.

When looking at the principal components of natural scenes (illustrations 33 and 34), we notice that they very much resemble a Fourier base of the image space. Also, the principal components computed from (the same) images of different sizes are of virtually identical content (except, of course, for their sizes), and should therefore also have an almost identical meaning – which in our case should manifest through almost identical classification results over the various image sizes (see below).

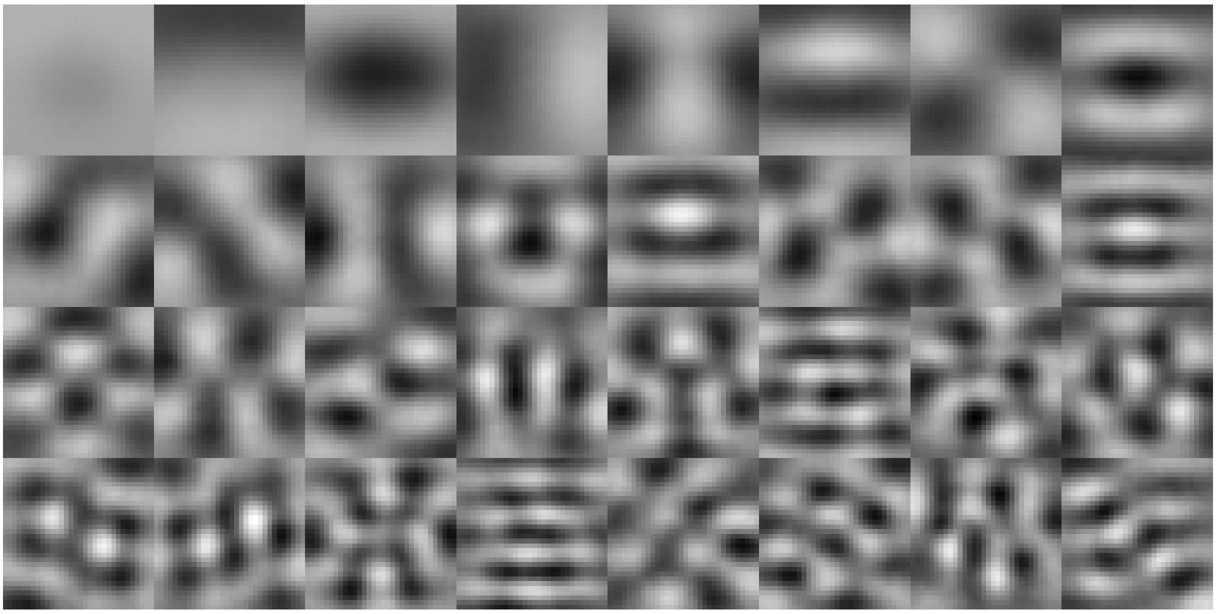

Illustration 33: The first 32 principal components of all images, size 32x32 pixels

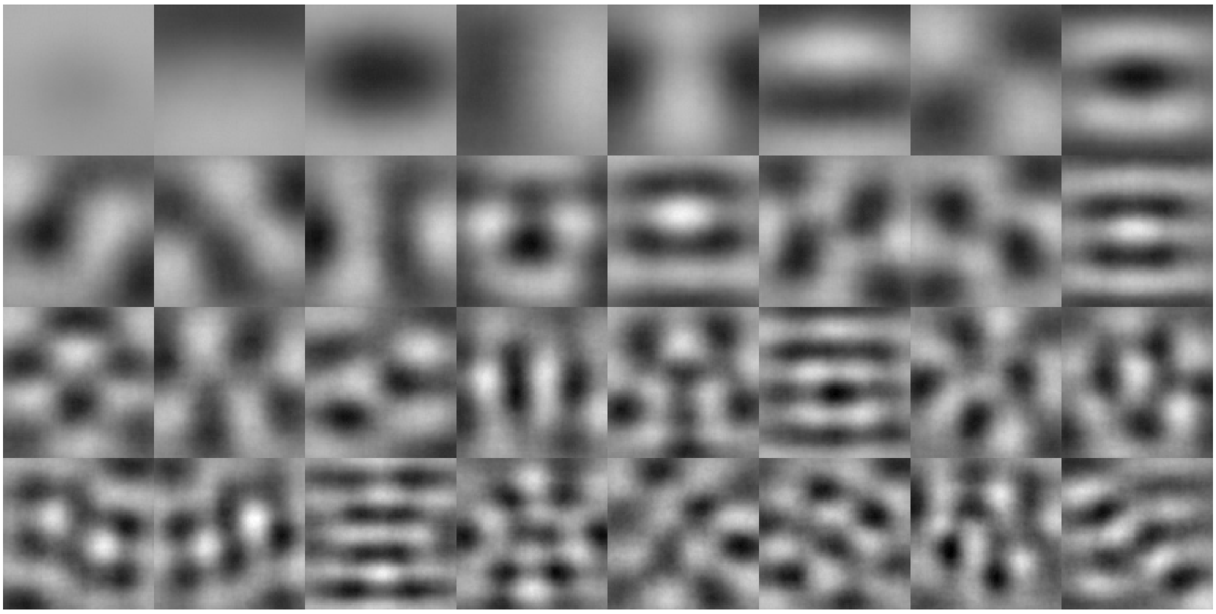

Illustration 34: The first 32 principal components of all images, size 128x128 pixels

3.3.2.1. Results

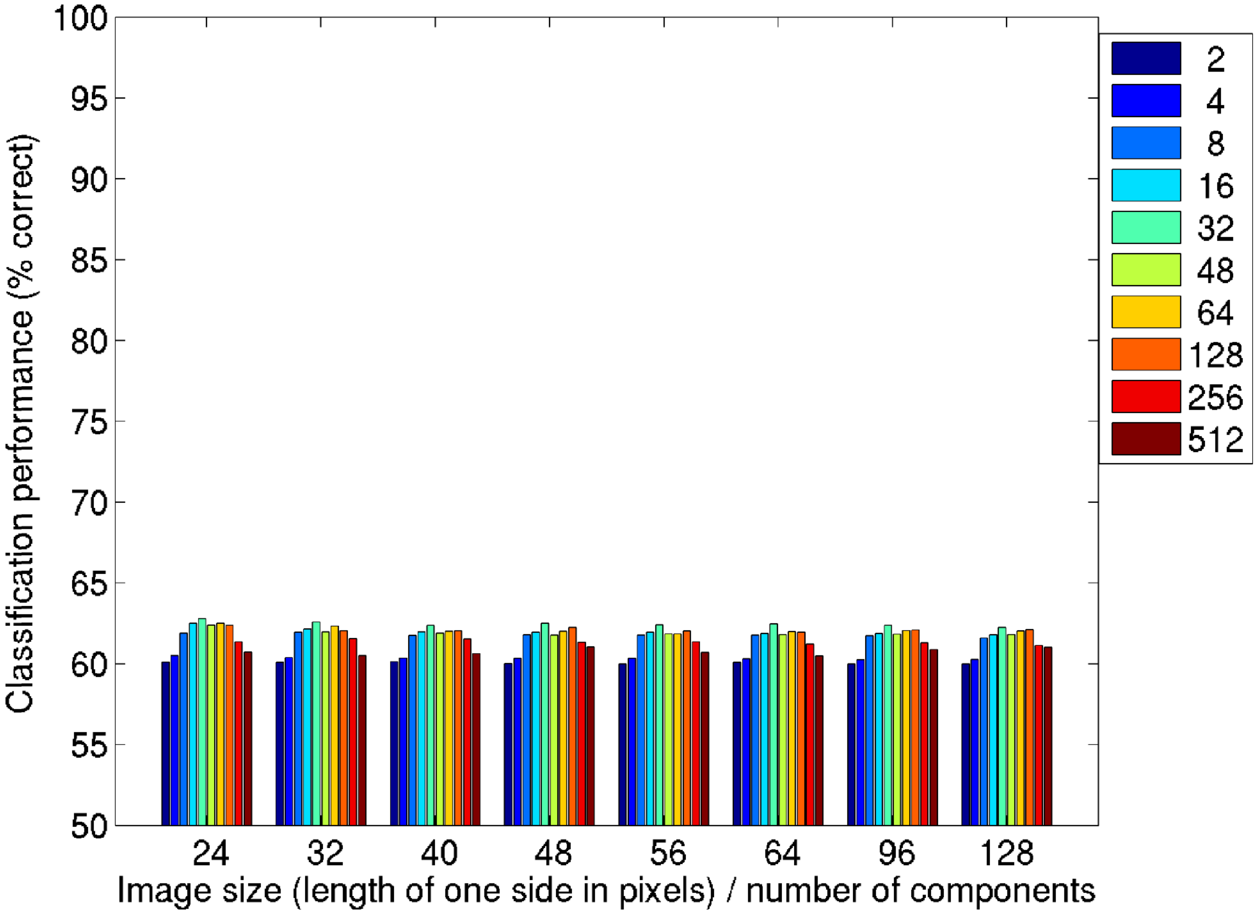

Illustration 35: Classification performance on various numbers of principal components of natural images

As predicted from the nature of the principal components, the results for different image sizes are virtually identical; the classification results are slightly better than those achieved on direct pixels (see previous chapter). We abstained from computing the results on images of size 256x256 both because of the prohibitive computational cost and the absence of an expected difference to the other image sizes. The slight improvement over the results of the attempt on plain image pixels is assumed to be a result of the "concentration" of variance on the first PCs. Also, the number of dimensions classified on was never higher than 512, with classification performance peaking when the first 32 components were used. The negative effect of an insufficient number of samples was therefore significantly reduced relative to the previous chapter. However, the maximum classification performance stayed below 64%, still being largely unsatisfying. Most of the variance covered by the first principal components therefore seems to be only loosely correlated to the presence or absence of an animal within the individual scenes. The use of principal components per se thus can not significantly improve the results of our classifier.

3.4. Classification with Global Image Statistics

Since we were unable to achieve a satisfying classification performance using the raw intensity values of the pixels of our images, we need to find another way to extract the relevant information from our data. As mentioned in the introduction, it has been suggested that the global amplitude spectrum can be used to categorize the content of scenes. In the following chapters we will try to exploit this approach in our quest for efficient classification of animal / non-animal images.

3.4.1. Classification using Spectral Principal Components

It has already been reported that the mean amplitude spectra of “animal” and “non-animal” images differ in shape ([Torralba, Oliva 2003], [Johnson, Ohlshausen 2003], see also chapter 2.3). Of course, the mean of the pixel intensity values of the images from our database also differ – still we have been unable to use this to our advantage in our classification attempts. Between the two mean amplitude spectra, however, the difference is much more localized than between the two mean intensity images. One more advantage of using the amplitude spectrum is the fact that it is symmetric to the origin, requiring only one half of each individual amplitude spectrum to be used for classification. While reducing the number of dimensions involved by a factor of two, this will still not allow us to use the raw values of the amplitude spectrum for classification due to the number of dimensions in the remaining half. When working with our images at their full size of 256x256 pixels, the number of dimensions in one half of the amplitude spectrum is still 32768, far too many for a successful classification with our 10864 examples. We will therefore reduce the dimensionality of our data space by computing the principal components of the amplitude spectrum (called “spectral principal components” or “SPCs” by Torralba et al.), and then use only the first few for classification. The structure of the differences in the amplitude spectra should make it possible to find a pattern within the loadings of the spectral principal components of the images that will enable us to achieve a better classification performance on our dataset than with the raw intensity values. This was previously demonstrated by Torralba and Oliva, who used the spectral principal components of their images to achieve a classification performance of up to 85% in the “animal” vs. “non-animal” task. On the following pages, a reproduction of the results of Torralba and Oliva shall be attempted, both on the original image database that Torralba and Oliva used (the “Torralba dataset”) and our own database.

3.4.1.1. Method

To be able to reproduce the results of Torralba et al. as precisely as possible, it is important to follow the same preprocessing steps as the original authors. The steps as reconstructed from [Oliva et al. 1999] and [Torralba, Oliva 2003] are the following:

The original images as used in their work were kindly provided by A. Torralba, a detailed description of their general statistics can be found in [Torralba, Oliva 2003]. They are JPEG files, with a size of 256x256 pixels in RGB color and were transformed to grayscale, thus being represented by their individual intensity functions $I(x,y)$.
After this, the logarithm was computed on the intensity distribution of the images: $I_l(x,y,k)=\log(I(x,y,k))$, with $I(x,y,k)$ being the intensity value of the image no. k at position (x,y) .
The next step was the application of a high-pass filter in order to “attenuate the very low spatial frequencies” [Oliva et al. 1999]. This was approximated by multiplying the amplitude spectrum of each image with a narrow, inverted Gaussian mask (d=256 pixels, σ =1.5 pixels), with the negative maximum centered on the zero-frequency component of the amplitude spectrum. This seemed to produce similar results to the otherwise unreported filter that Torralba et al. used.
Finally, Torralba et al. reported that they applied “an adjustment of the local standard deviation at each pixel of the image” [Oliva et al. 1999], yet they did not precisely report in which way. We therefore assumed that the standard deviation was adjusted to 1, which is a common practice in image preprocessing: $I_{ls}(x,y,k)=\frac{I_l(x,y,k)}{std_{n=1}^{N} I_l(x,y,n)}$, with N being the number of images in our database (10864).
To minimize aliasing in the following Fourier transform, we multiplied the images with a Gaussian hamming-window of the same size as the image (d=256 pixels, σ =1/4d). Subsequently, we computed the DFT of every image, retaining the amplitude spectrum and discarding the phase spectrum. As we received only the 600 animal- and 600 non-animal images of the original paper, we added these images to our own database prior to the computation of the principal components. This achieved two things: on the one hand, the principal component base used to recode the amplitude spectra is much cleaner since there are more samples available, and second, both the Torralba images and our own database are encoded with the very same spectral principal components (“SPCs”), allowing for a direct comparison of the classification performance on the two sets of images. We also

performed these steps on the smaller versions of the images in our database to be able to examine the effect of image size on classification performance. We evaluated the suitability of the SPCs for classification both with our already familiar linear classifier and with a support vector – based classifier (see chapter 7), using a RBF kernel.

3.4.1.2. Principal Components of the Amplitude Spectrum

The Spectral principal components of different sizes are shown below. While it is obvious that the higher resolution version of the SPCs show cleaner, more distinct features, it is also obvious that the general shape of the SPCs of various sizes is the same. The similarity is not as unique as with the principal components of the plain images, however. Also, the order and the orientation (the sign) of the principal components is not always the same. Hence, one may expect the classification performance to differ somewhat over the various image sizes.

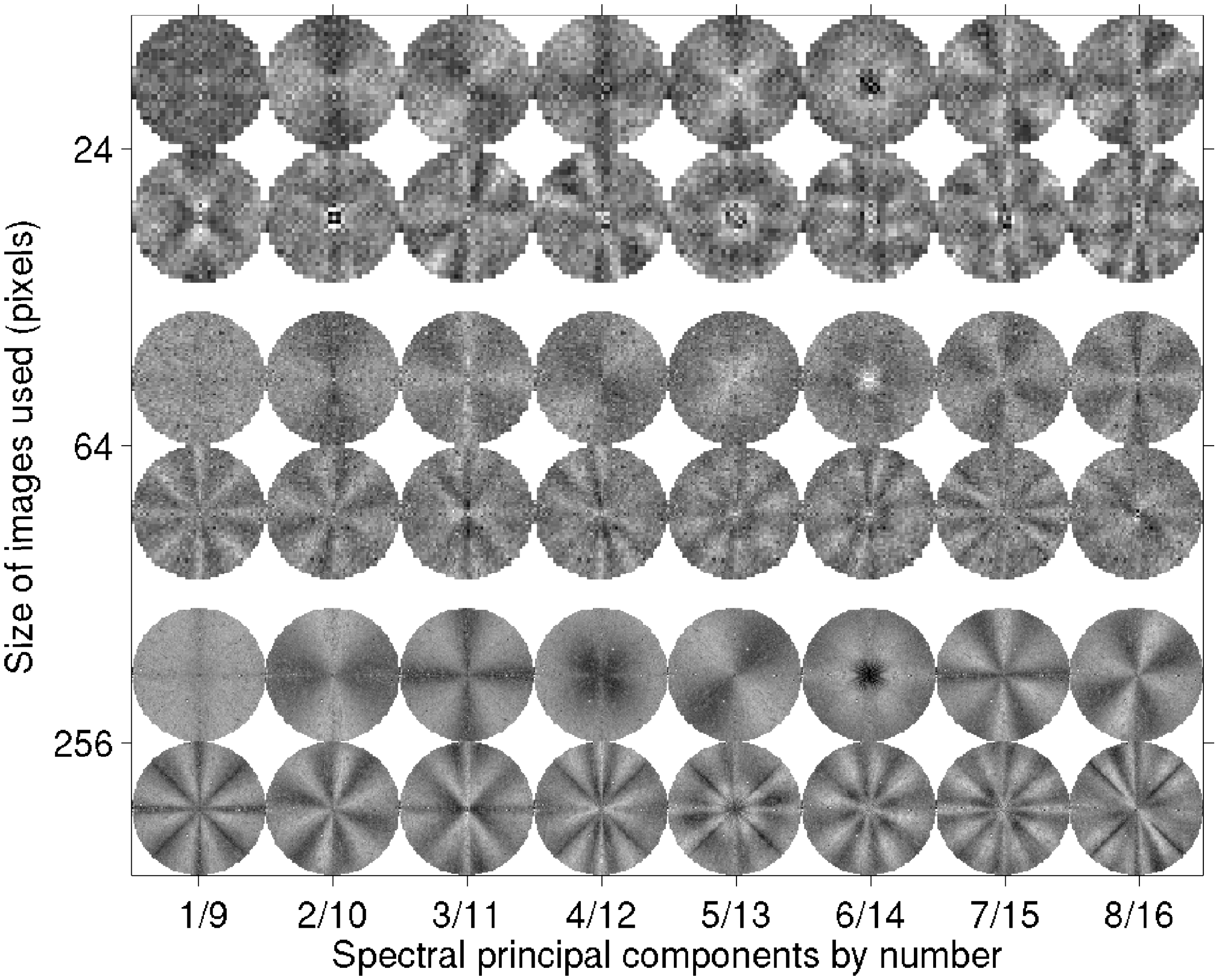

Illustration 36: The first 16 SPCs of different image sizes

3.4.1.3. Results Using the Linear Classifier

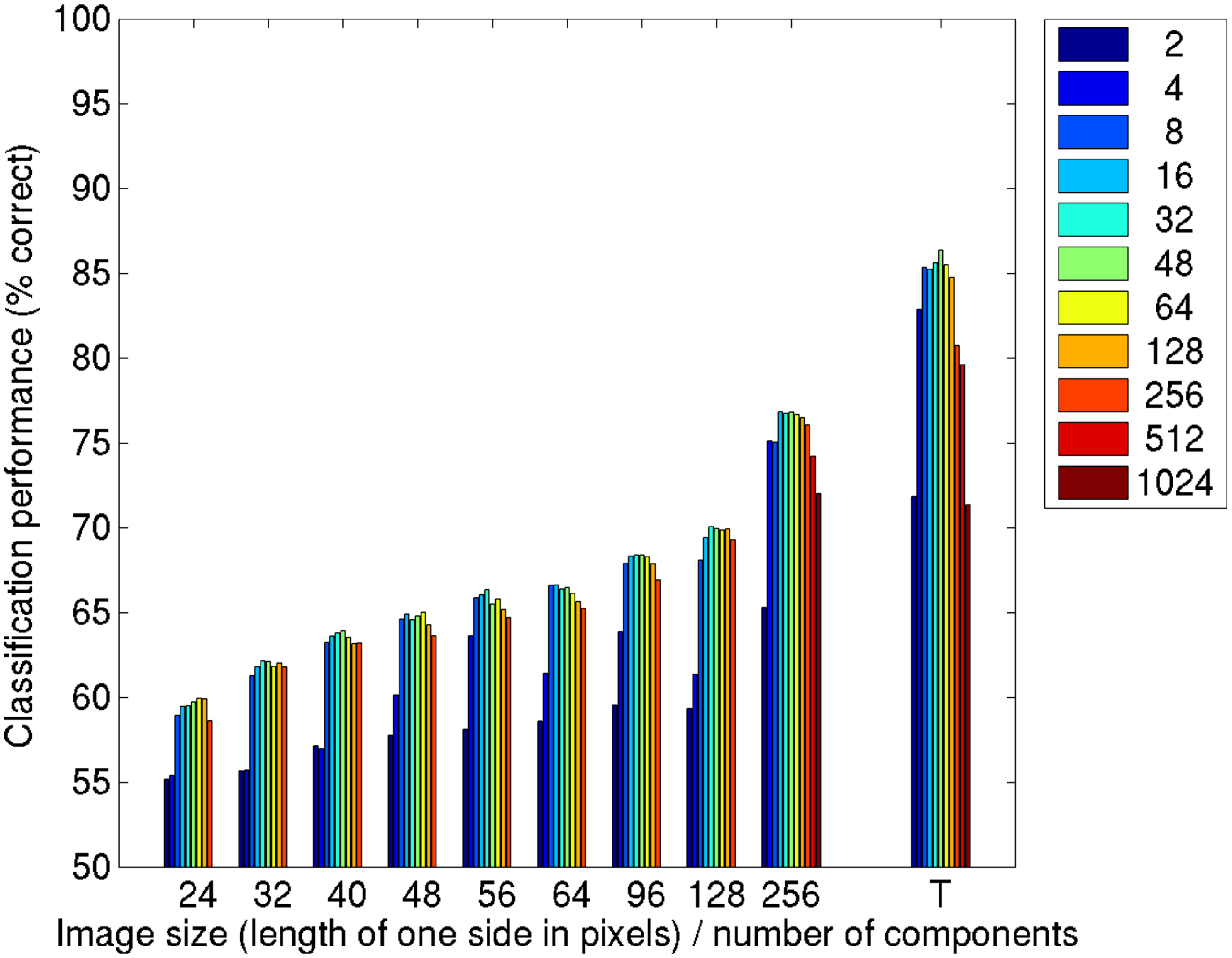

Illustration 37: Classification results on various image sizes and dimensionalities, using the linear classifier

Shown above are the results of the linear classifier, using various numbers of principal components on various image sizes, with a 200-fold cross-validation. Overall, classification performance increases with image size. As the actual dimensionality of the classified dataset is no longer directly dependent on the amount of pixels that comprise an image, this means that the usefulness of the information in the first n SPCs actually increases along with the size of the image. This might lead to the assumption that there is more useful information in the high spatial frequencies, as larger versions (really meaning “higher resolution” in this case) of the same image can contain higher spatial frequencies than smaller versions (Nyquist's Theorem). Apart from the classification gain correlated to image size, the number of SPCs used also clearly has an effect. Classification performance is best with relatively few SPCs, the maximum being between 16 and 64. The exact number of maximum performance differs between image sizes, but the difference between 16, 32, 48 and 64 is generally not large enough to determine a clear “winner”. Starting with about 128 components, the classification performance starts to decline, indicating that the number of samples available is not sufficient to successfully train the linear classifier on datasets of higher dimensionality.

3.4.1.4. Comparison Between Datasets

From the results shown in Illustration 37, it becomes obvious that the linear classifier performs significantly better on the dataset provided by A. Torralba (labeled “T”). The data itself is based on the same image sizes (256 pixels), the same preprocessing, even the very same principal components as our own dataset. The images are from the same original image database as our own; most of them also appear in our own dataset, though most of the time, different areas have been cropped. If we had expected our classifier to perform differently on the Torralba dataset at all, we would have expected it to perform worse, as there are only 1200 images in the Torralba dataset as opposed to our 10864 images. This does not allow us to train the classifier as extensively on the Torralba dataset. The increase in classification performance can therefore only result from image content, which again can only differ significantly through a “lucky” choice of images. We chose our own images to be as diverse as possible in order to be a reliable ground truth; when a classifier performs well on a large and diverse dataset, this suggests good generalization of the classifier. The Torralba dataset apparently is a less challenging selection of images to classify upon. For the remainder of this work, we will therefore continue to use our own image database exclusively.

3.4.1.5. Results Using the SVM/RBF Classifier

The fact that classification performance increases with image size leads us to the conclusion that it makes sense to continue classification experiments with the largest images (256x256 pixels) only. This causes the highest computational cost during the principal component analysis, but once this step has been performed, computational complexity is no longer dependent on image size. Since we have already computed the SPCs of all available image sizes, it would not make sense to continue using any but those that promise the best results in our proceedings. We apply this conclusion to the evaluation of the SVM classifier, testing it only on the SPCs of the largest images (256 pixels). With the RBF kernel used for our classification experiments, we generally find the same results as with the linear classifier: performance first increases with increasing number of SPCs, then a decrease can be seen with SPC numbers over 128. We also notice that despite the much higher computational cost of the SVM/RBF classifier (see chapter 7), the classification performance achieved is just slightly higher than that of the computationally cheap linear classifier (linear: 76.9%, 200-fold cross-validation vs. SVM: 77.5%, 200+200 fold C-search, 50-fold crossvalidation, see Illustration 38). For the purposes of the following pages and chapters, the use of the linear classifier shall therefore suffice.

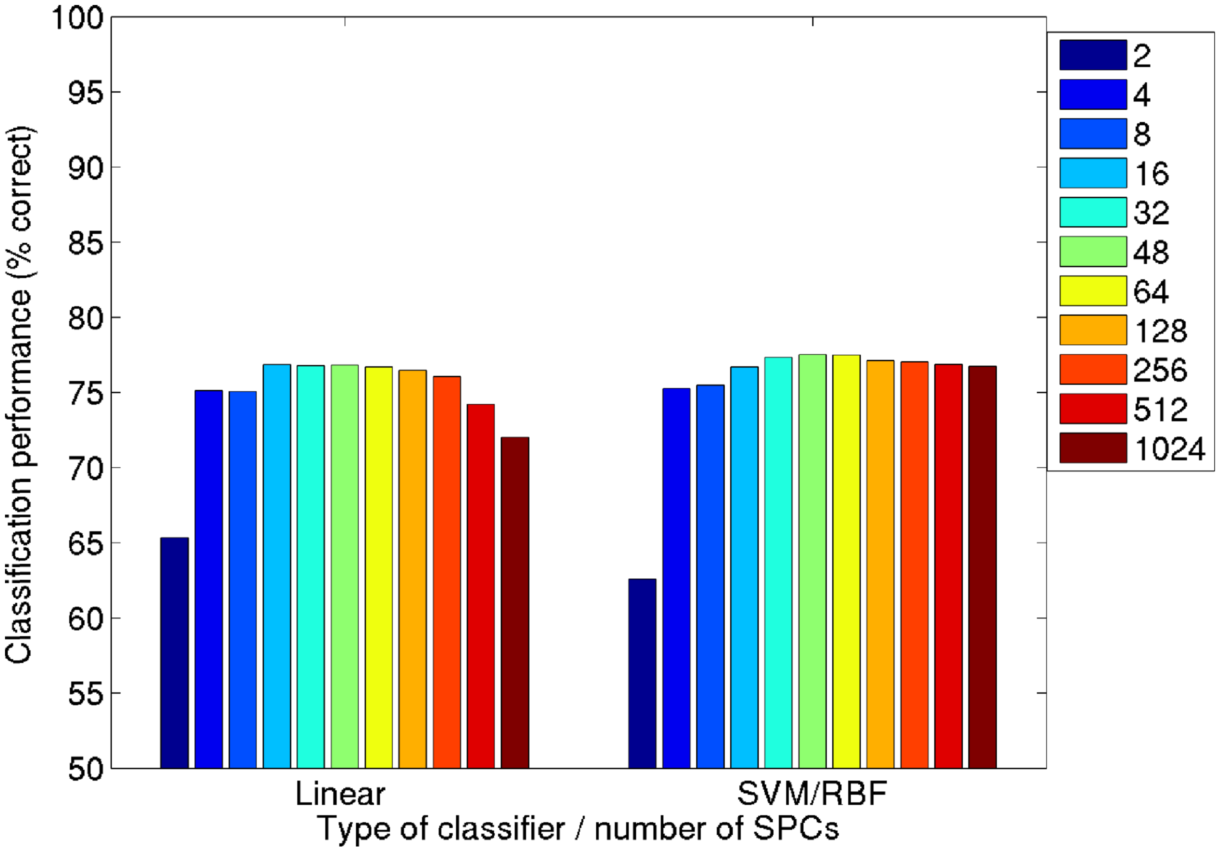

Illustration 38: Classification performance on SPCs of size 256 pixels, linear vs. support vector classifier

3.4.2. Classification Based on the “Fourier Fingerprint”

3.4.2.1. Motivation of the “Fourier Fingerprint” Approach

In the previous chapters, we showed that classification performance just above 77% can be reached when using the first n principal components of the amplitude spectrum. The first few of the principal components are those that cover the most variance over the entire set of images, both “animal” and “non-animal”. It is usually assumed that the large variance along the direction of the first principal components also corresponds to high classification performance – due to the higher amount of variance covered, more difference between individual images can be accounted for. There is, however, no computationally feasible way of proving that the first n components are also the best n components for classification; in fact, the order of the principal components in terms of variance might in theory be quite different from their order in terms of classification relevance. The relatively costly step of computing the principal components might therefore be replaced with something computationally simpler, without necessarily compromising classification performance.

As found by Torralba et al., and supported also by our own analysis of the general statistics of our image database (see chapter 2.3), the mean difference between the “animal” and “non-animal” images are more apparent in the horizontal and vertical orientations than in the oblique ones, and while there is significantly more overall energy in the low frequencies, the relative difference between oblique and non-oblique orientations appears to be strongest in the higher frequencies. It is therefore acceptable to assume that some frequencies and orientations will be more useful to our classification task than others. To illustrate this point, we will try to employ a more direct approach than before, subdividing the amplitude spectrum into a number of areas by means of frequency and orientation, without computing principal components first (or thereafter). If this approach performs well enough, the relative importance of each of these areas for the classification might then be used to infer the location of the classification-relevant information within the amplitude spectrum.

3.4.2.2. Method, Data Preparation / Preprocessing

We again accessed our image database (256x256 in 8bit grayscale) and superimposed our images with a Gaussian window of the same size (sigma $64\text{pixels} \simeq 1/4d$) to minimize boundary artifacts in the following Fourier transform (Illustration 39). As our images consist exclusively of real values, it is safe to ignore one half of the resulting amplitude spectrum due to the symmetry inherent to it (positive and negative frequencies are always center-symmetric for real-valued data, see [Butz 1998]). This helps a lot to reduce the amount of data that needs to be processed. In analogy to the previous approach by Torralba et al., we discarded the phase spectrum.

To decompose the spectrum into bands of orientation and frequency, the contents of the remaining half of the amplitude spectrum of each individual image were then collected in a grid mask, arranged as 8 orientations with 6 frequency bands each (see Illustration 40), together forming 48 bins, which might be regarded as the “Fourier fingerprint” of the original image. All of the values of the amplitude spectrum that fell into the same bin were then summed up to form exactly one value per bin, effectively reducing each image into a matrix of 48 numbers. The size of the bins was chosen in octaves to compensate for the average 1/f energy distribution found in natural scenes (see also chapter), so that on average the same overall amount of spectral energy resides in every bin (the area covered by any given bin and the 1/f spectral slope compensate each other). The actual sum of values in every bin may, however, still be somewhat different than the 1/f slope one might expect. The zero-frequency-component, equivalent to the DC offset of the image, was ignored, as the mean intensity of an image does not really contribute to the actual content of the image. The above preprocessing provides us with 10864 images (5432 “animal” images and an equal number of “non-animal” images) represented by 48 values (bins) each. Prior to classification, each of the bins was normalized to $[0\ 1]$ over all images.

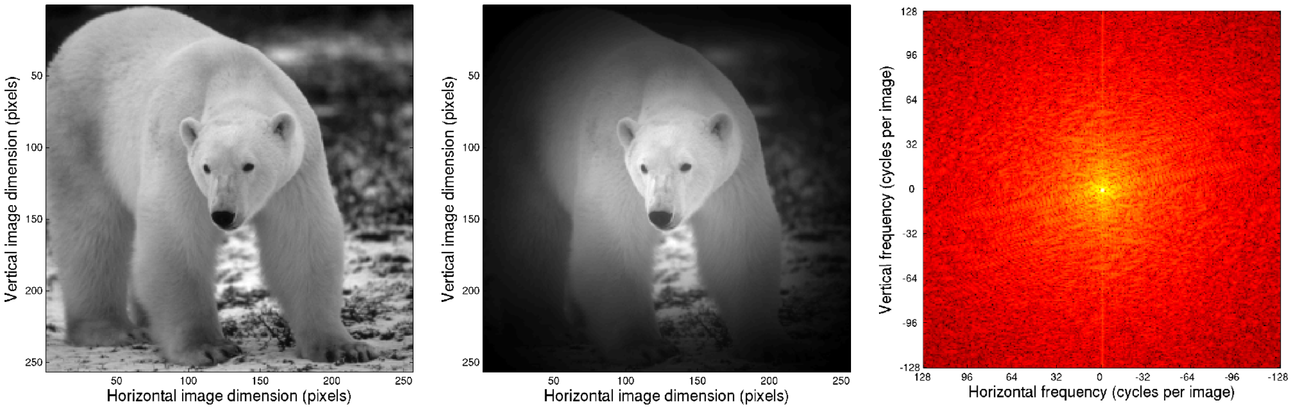

Illustration 39: Image preparation and fourier transform

left: *The original Polar Bear, size 256x256, 8bit grayscale*

center: *The same Polar Bear, superimposed with the Gaussian window*

right *The amplitude spectrum of the Polar Bear*

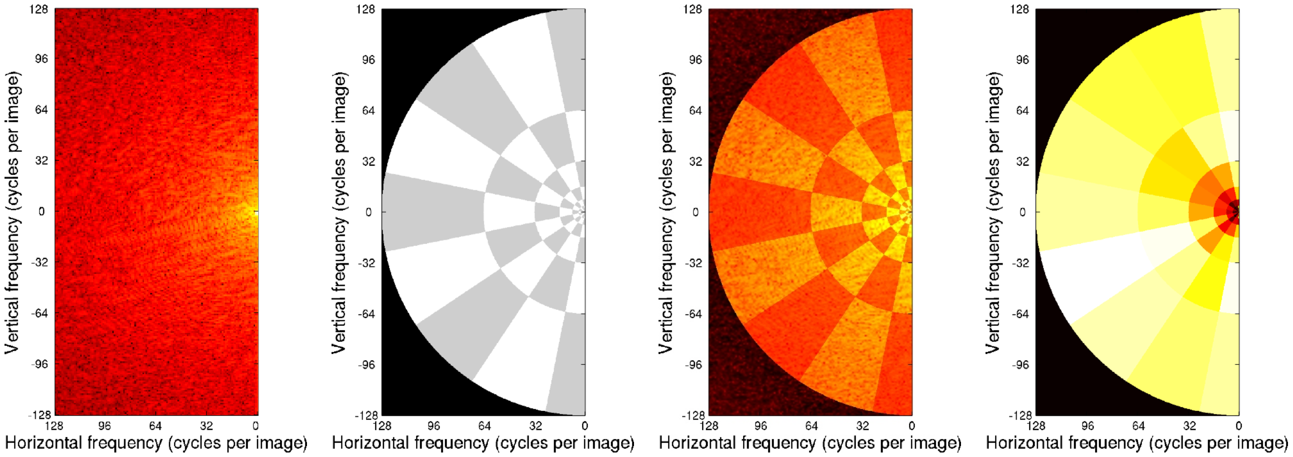

Illustration 40: Summing the amplitude spectrum into 48 bins

left: *One half of the Polar Bear's amplitude spectrum*

left center: *The grid mask, shown in alternating gray and white for illustration purpose*

right center: *The grid mask being applied to the Polar Bear's amplitude spectrum*

right: *The resulting 48 bin (orientation / frequency bands) "Fourier fingerprint"*

3.4.2.3. Classification

Initially, we classified on the entire Fourier fingerprints (all 48 bins) using the linear classifier (see chapter 7) with a 200-fold cross-validation. Classification performance in this basic setup exceeded 74% (74.29% equaling 8071 of 10864 images correctly classified), showing that the replacing of the Principal Component Analysis with our simple bin structure did only slightly lower the efficiency of our classifier.

The general idea, however, was to discover which of the bins and, thus, which of the frequency / orientation bands are more or less important for the separation between animal and non-animal images. We therefore need to find a way to put the bins into some kind of ranking. Simply classifying on every single lonely bin and later sorting the bins by order of classification performance may not lead to success because of interaction effects – sometimes, the classification performance using several (e. g. 2) bins simultaneously can be different from the sum of the individual classification performances of those bins. The only theoretically correct approach would therefore be to try out all possible combinations of any subset of our 48 bins, an undertaking that, while methodologically sound, is not feasible due to the enormous amount of computations necessary to perform this task: we would have to perform $48!=1{,}51\cdot10^{51}$ classification runs, with 200 cross-validations each if we want to stay recent with our single classification run on all 48 bins. This can not be done within an acceptable time frame even with the most advanced of todays computers - we therefore need an approach that, while still producing a plausible result, will not require quite as many computations.

3.4.2.4. Iterative Bin Elimination Procedure

Our Approach is as follows:

In an iterative procedure, we will remove exactly one bin per iteration cycle.
We first classify on all our dimensions (meaning all our bins). Then, we take all 48 possible combinations of 47 bins, and select that combination that shows the best classification performance of all of these 48 combinations. There now is exactly one bin that was not included in this combination, which is considered the one that can be removed with the minimal loss of classification accuracy. This bin is then considered "permanently eliminated" - it will never be used again for the remainder of this procedure.
We continue in our next iteration, working on all 47 possible combinations of 46 bins, and so on, until there is only one bin left.

The description of the algorithm in pseudo-code is as follows:

N	dimensionality of the data (in current iteration)
i	number of iteration (the "i-th" iteration is always the current one)
P(x)	classification performance of the dataset "x"

```
for i=1 to 48 do
    1. select all $N$ possible combinations (" $C_n$ ") of $N-1$ bins from Dataset $D$
    2. classify on all of these $C_n$
    3. select the combination $k$ where $P(C_k)=\max_{n=1}^{N}(P(C_n))$ , with $C_k$ being the
       Combination where the $k$ -th element has been removed
    4. store the eliminated element in a storage vector
    5. let $C_k$ be the new $D$
continue
```

When sorting all bins by the number of the iteration they have been eliminated in, we get the desired ranking of the bins. The last n surviving bins are considered to be the most important ones for the classification task. Computing this minimum-loss ranking for our 48 bins requires only

$\sum_{i=1}^{48} i = 1176$ classification runs with varying dimensionality (from 48 down to 1), which could have been computed in about 1-2 weeks on a single machine. The process was accelerated by using a computation grid. In every iteration, all of the possible combinations can theoretically be computed in parallel, as they are completely independent of each other. It has therefore been efficient to distribute the individual combinations as one computation job each, accelerating the overall process almost linearly with the number of computation nodes available, which varied between 3 an 12 depending on the availability of the grid nodes. It becomes apparent from the results of this algorithm that during the first few iterations (the ones with still over 40 bins left), the maximum classification performance achieved by the classifier is indeed increasing with the number of eliminated bins, even exceeding the performance achieved with all 48 bins. This is caused by the harmful effect that a "useless" bin can have on the performance of the linear classifier: If one or several dimensions of the dataset can by themselves not be classified with a performance significantly higher than chance, then it might be optimal for the classifier to actually ignore this particular dimension altogether, as there is no useful information coming from it. The linear classifier, however, does not posses the ability of ignoring one or several data dimensions, and therefore needs to include this information in its decision despite of the possible negative implications: as such a "useless" dimension per se only provides chance performance, it will ill affect the overall performance of the classifier, as if dragging it down.. In other words, since the linear classifier can not ignore data dimensions by itself, eliminating one or several data dimensions "manually" (through the elimination algorithm used) can actually improve classification performance.

The above algorithm can also be performed in an inverse fashion: We proceed just like before, but we select those bins for elimination that will cause a *maximum* loss of classification accuracy, instead of a minimum loss. In a perfect world, this would produce exactly the same ranking as would the minimum-loss approach, just in reverse order. Due to the interaction effects between bins, however, the order may be slightly different. If we look at the classification performance graph in Illustration 41, we see that after the small gain during the first iterations, the classification performance of the minimum-loss approach stays very much the same for a long period, forming a plateau. As there are only slight differences between individual combinations of bins within this plateau, the true "distance" between bins in the resulting ranking will be just as minor. In other words, even very small effects may trigger a slightly different ranking, and as we remove the most useful bins at the very beginning of the maximum-loss approach, the resulting differences between the "weaker", less useful bins are emphasized. The general trend, however, will be the same on both approaches: the "strongest", most useful bins stand out against the majority of "weaker", less useful bins. If one wants to see only which bins are actually the most important ones, the mean of the two rankings is going to be the most reliable one, representing those bins that are on average the "strongest" of both approaches.

3.4.2.5. Results of the Iterative Bin Elimination Procedure

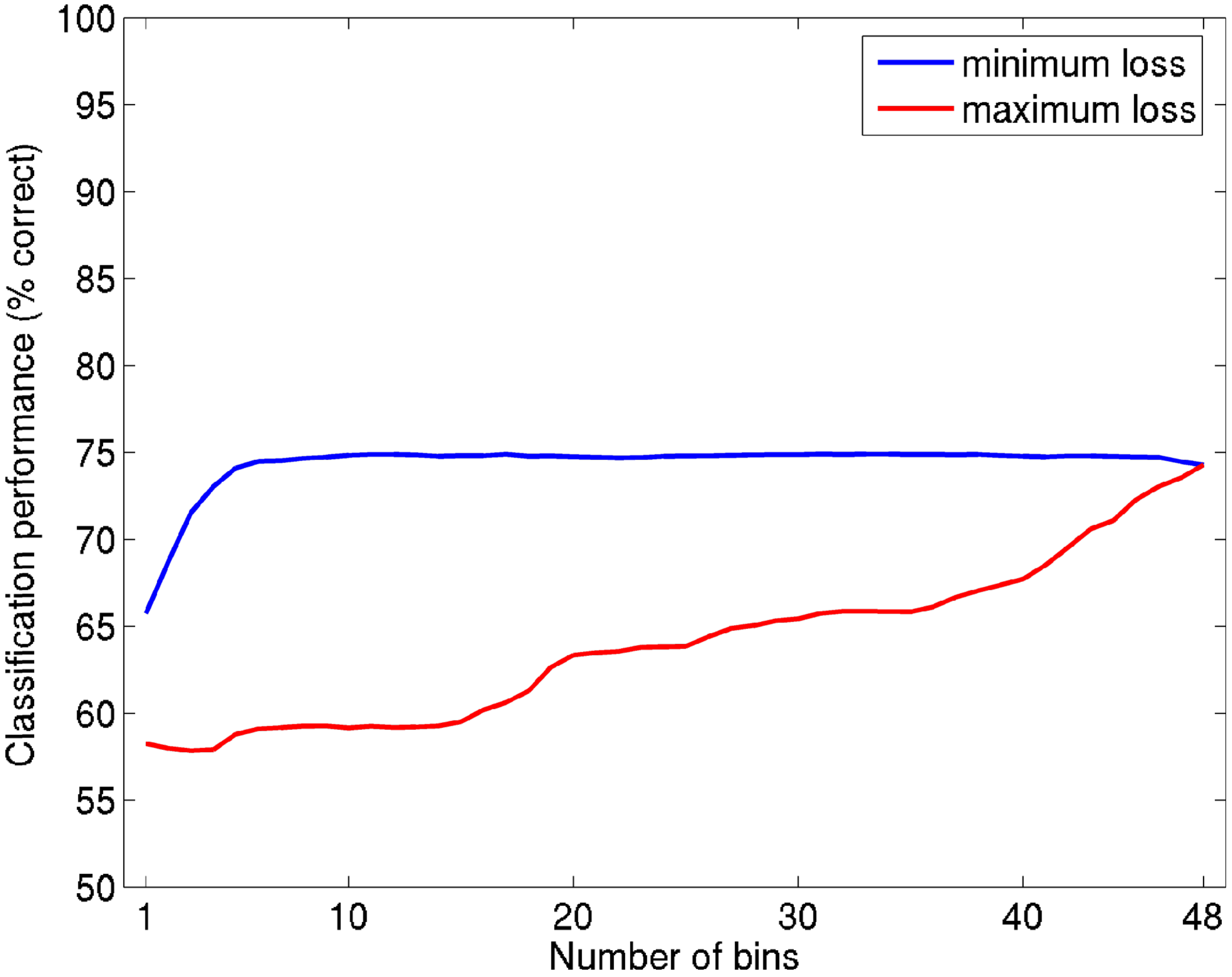

Illustration 41: Classification performance during incremental bin elimination

In the minimum-loss approach we find that at first, with most bins still present, there is a small increase in classification performance (see above). Then, there is a very long plateau, going from about 45 bins down to as few as 6 or 7 bins. During the entire plateau, the performance stays just below 75%, only falling below the 74% marker when less than 5 bins are left. This means that when using only the last ("strongest") 5 bins, a little more than 74% classification performance can be achieved. The decline in classification performance during the maximum-loss approach varies more through the duration of the procedure than the during the minimum-loss approach, with performance decreasing monotonically until the last 3-4 bins are left, after which the performance actually increases just slightly. This final behavior is similar to the minimum-loss-approach, where the elimination of the first few bins led to a slight increase in classification performance as well.

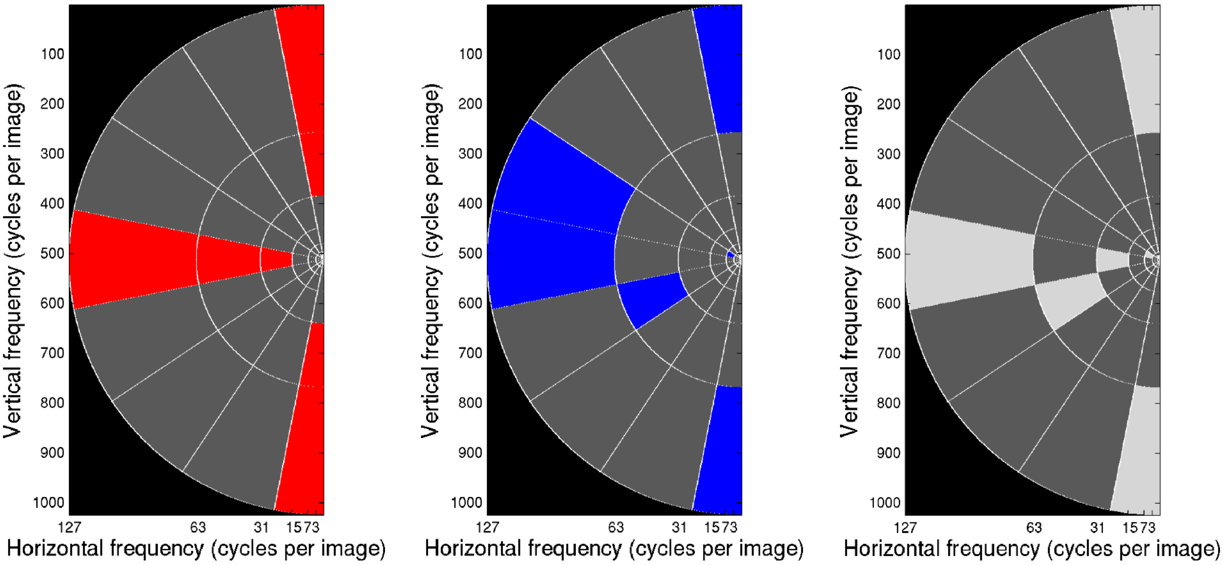

Illustration 42: Iterative bin elimination: the 5 "strongest" bins

left: maximum-loss approach, center: minimum-loss approach, right: mean

Both on average and in the individual approaches, the bins that represent the highest frequencies in the horizontal and vertical orientations are among the very top of the ranking, while the bins on the oblique angles play no major role in any of them. This is no surprise, as the high frequencies in the vertical and horizontal orientations are the ones that differ the most in the general average of the image categories (see Chapter 2.3). The only bin that is located in the lower frequencies is also the one closest in terms of ranking to the bins of the plateau and therefore could be exchanged with any of the plateau bins with only very minor (though, of course, not minimal) loss in classification accuracy.

3.4.2.6. Ranking vs. Rating

While the order of the most important bins in this ranking is quite reliable, there is no real metric to determine the actual distance between two bins. This is a drawback, since it will not allow us to tell how much more or less important one bin really is compared with another. To accurately measure this, the complete evaluation of all possible bin combinations would be necessary, but is not computationally feasible, as mentioned before. One might attempt, however, to compute something that approximates the complete evaluation on a coarse scale, resulting in a quasi-continuous metric of bin importance that includes at least some of the interaction effects between bins. To achieve this, a subset of all the possible bin combinations needs to be computed, for example the subset with all combinations of exactly k bins. The resulting classification performance achieved on each individual combination gets accredited to all of the include bins. This results in a matrix M_k of size N^k , each entry representing the classification performance achieved when using the k -tuplet specified by the indices of the k axis of the matrix. Populating the matrix will require $\frac{(N^k)}{k!}$ classifications – the division by $k!$ is due to the fact that the permutations of each individual bin combination do not differ in their classification behavior and therefore need to be computed only once. After computing the classification performance on all of the possible subsets, the average of all accredited classification performances is computed separately for each bin. The resulting mean value is a semi-continuous rating of the bin's general importance for the classification. To keep the number of computations on a feasible level, this was done for $k=2$, thus requiring $\frac{48*47}{2}=1128$ classifications with 2 dimensions each, plus an additional 48 classifications with 1 dimension (every bin by themselves, located on the diagonal of the resulting matrix), totaling 1176 classifications. By its general computational structure, this approach would scale very well on a computation cluster or grid, as all the individual k -tuplets can be classified upon independently; however, the individual computations are only 2 dimensions wide and therefore complete very rapidly. The computations have therefore been done on a single machine, completing in mere minutes.

3.4.2.7. Results of the Pairwise Ranking Computation

The pairwise classification results show, very much alike the iterative elimination approach, that the horizontal and vertical orientations, and within these the higher frequencies are most important for classification. When taking the per-column-mean of the above matrix, we get a vector of size 48, which we can reshape into the already familiar bin structure. From the rating of the bins we can see that the relative distance between the 4-5 most important bins and the rest of the field is actually quite large (see Illustration 43), while the distance between the first and the second most important bin is rather small (the mean classification relevance of the top 5 bins is 72,2%, whereas the mean of all remaining bins is merely 15,3%).

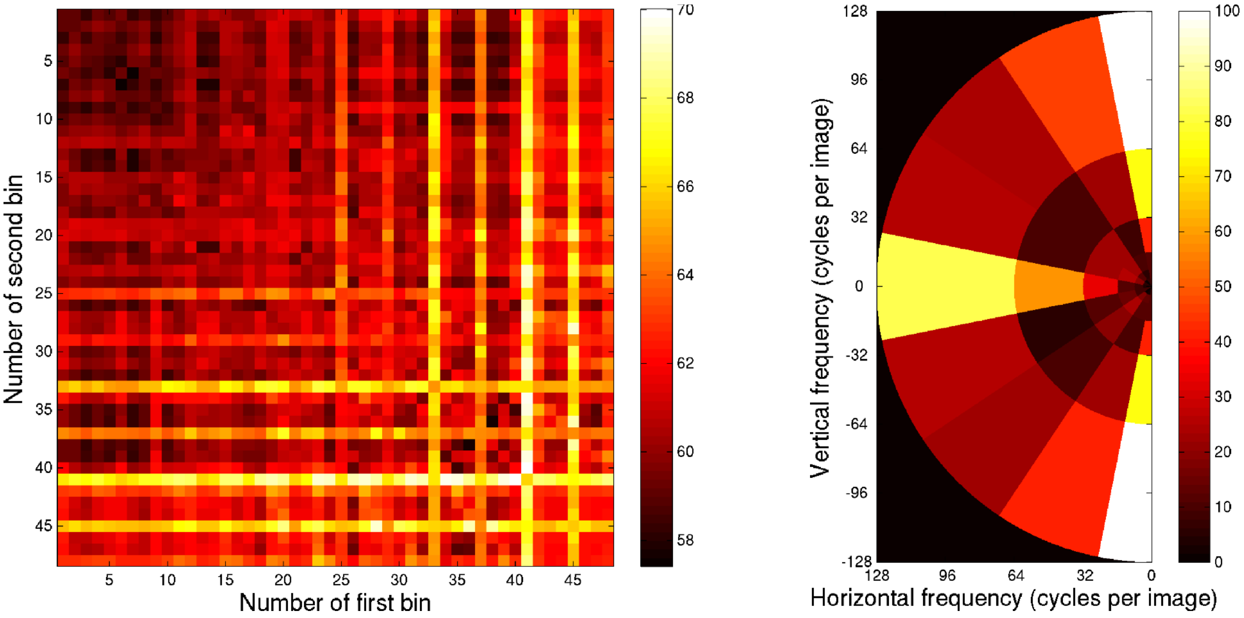

Illustration 43: Results of rating with k=2

To the left, the resulting matrix M_2 *of size 48x48, unit is classification performance in % correct.*

To the right, reprojection of the mean of the matrix into the bin shape, rating in % importance relative to the single most important bin.

3.4.2.8. Classification on Frequency Bands

The result of the evaluation in the previous chapter showed the importance of the vertical and horizontal orientations of the highest frequency band for classification. This, of course, also suggests that the highest frequency band generally has more meaningful content in terms of classification relevance. To examine how the information relevant for classification is distributed amongst the frequency bands, we split up our Fourier fingerprint data into the 6 included frequency bands, resulting in 6 datasets with 8 bins each (one for each orientation band). We then classified again using the linear classifier, with 200 cross-validations on every classification run.

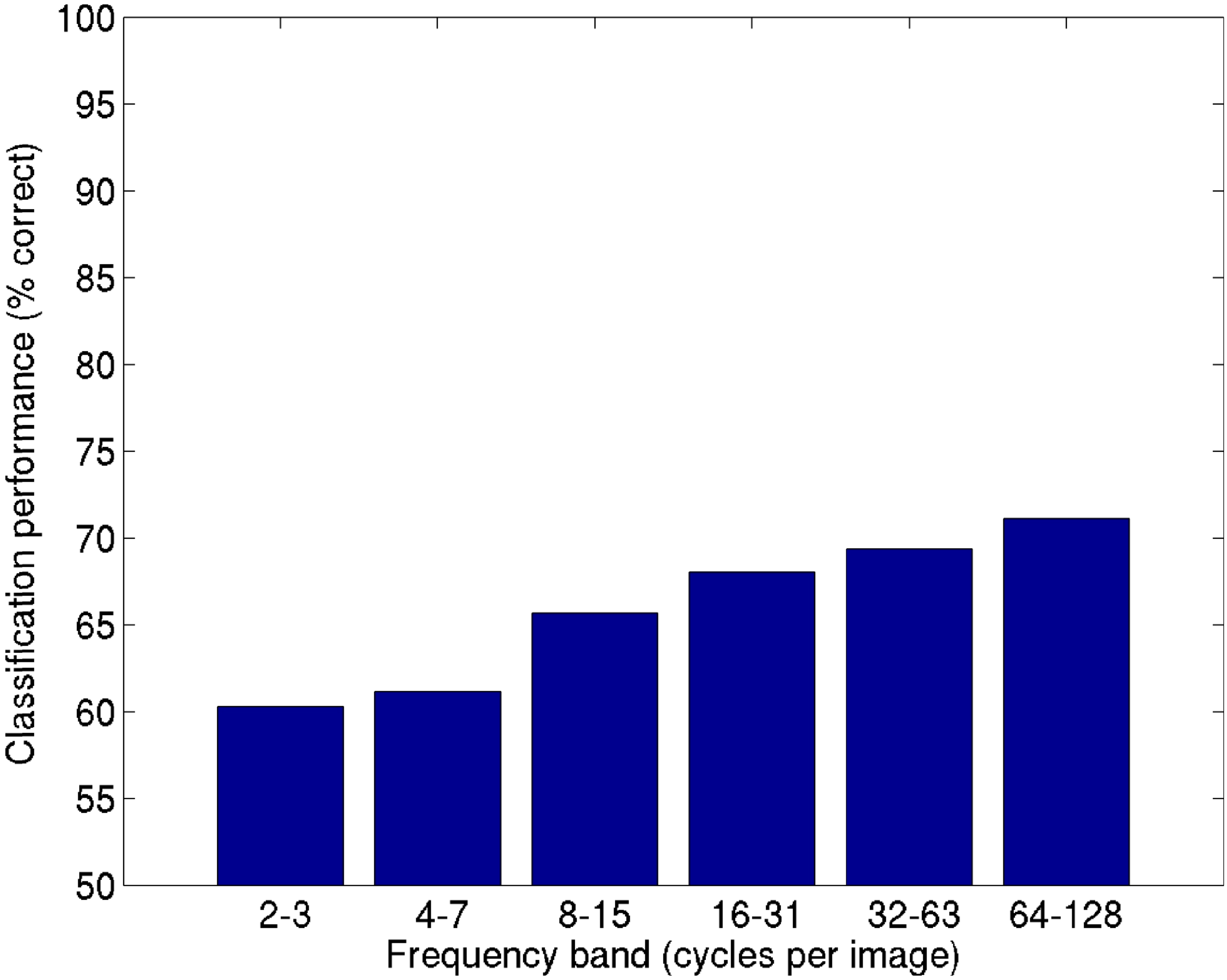

Illustration 44: Classification performance on frequency bands

The results show that indeed, the highest frequency band appears to contain the most relevant information for the classification process., exceeding 70% accuracy, while the lowest frequencies barely pass the 60% line. The ratio of performance increase over the frequency bands is quite even except for a stronger than average increase between the 4-7cpi and the 8-15 cpi frequency bands. The image structures corresponding to the frequencies below 8cpi are therefore the least useful ones for classification, while the frequencies above that apparently contain more and more information the higher the frequency.

3.4.2.9. Selecting Images by Classification Difficulty

The output of the linear classifier is a list of assigned labels and associated probabilities, one for being “animal” and one for being “non-animal”. Browsing through this list allows one to select a number of images based on their associated probabilities (as a measure of distance from the plane of discrimination). The images within the “animal” class that score the highest “animal”-probability and the “non-animal” images with the highest “non-animal” probability could be seen as the ones most easily classified by the algorithm, the “animal” images with the lowest “animal” probability and the “non-animal” images with the lowest “non-animal” probability could be seen as the ones most difficult for our classifier. Whenever an image scores less than 50% in its class (automatically scoring higher than 50% in the other class), this means that is has been misclassified.

	animal probability	***distractor probability***
“easy” animals	>>50%	<<50%
“difficult” animals	<50%	>50%
“easy” distractors	<<50%	>>50%
“difficult” distractors	>50%	<50%

Table 2: Image selection by classification difficulty

On the following pages, those images are being displayed that the linear classifier, based on all of the 48 bins, chose to be the 200 most class-typical (“easy”) or the 200 most class-atypical (“difficult”), for both classes. All of the atypical images have class probabilities below 50%, while the typical ones have probabilities far above 50%, usually above 80%.

As mentioned before, the discrimination plain is located at 50%, so all images labeled “easy” were classified correctly, while all images labeled “difficult” were classified incorrectly.

3.4.2.10. The 200 "Easy" Animal Images

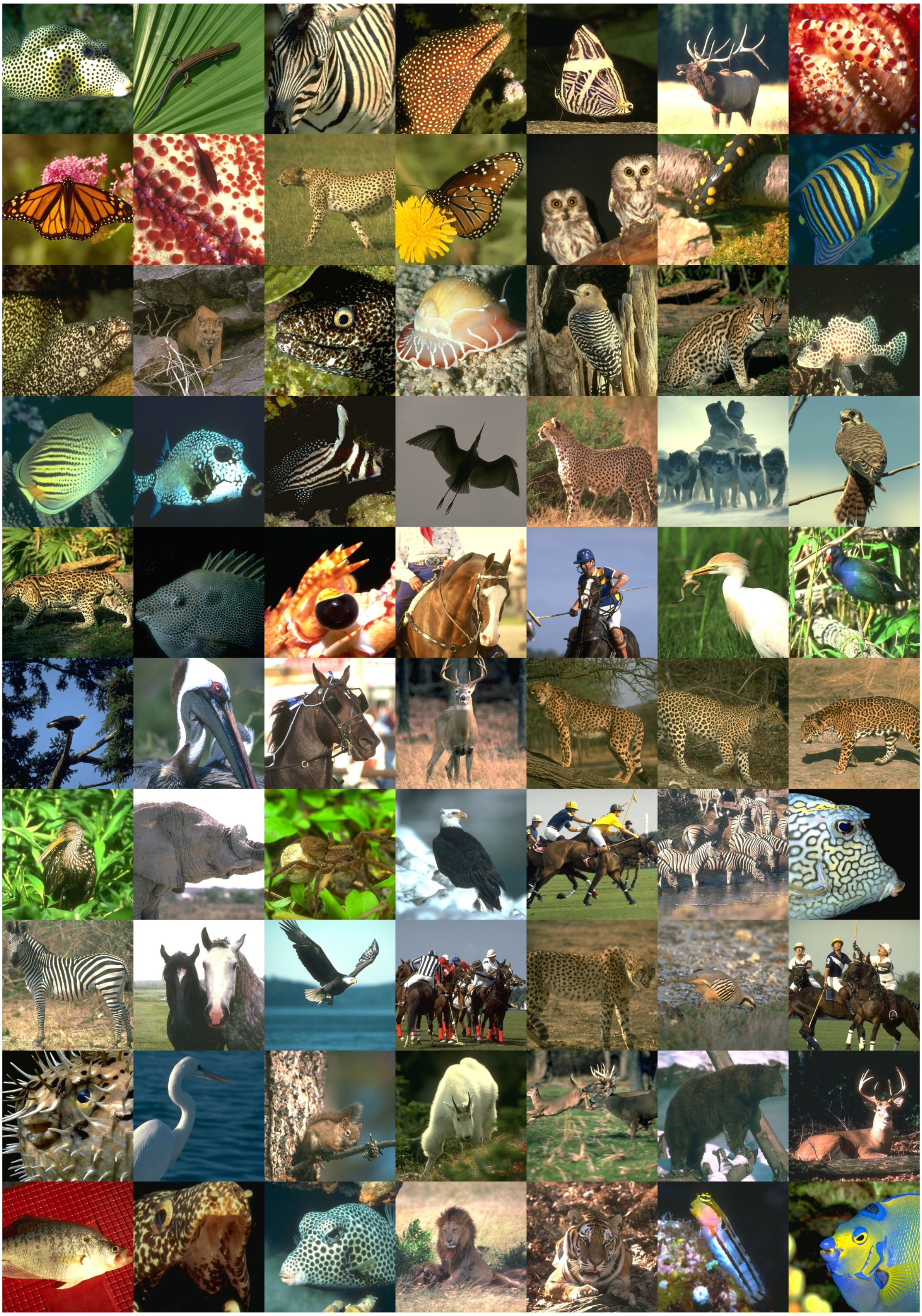

Illustration 45: The 200 "easy" animal images, no. 1-70

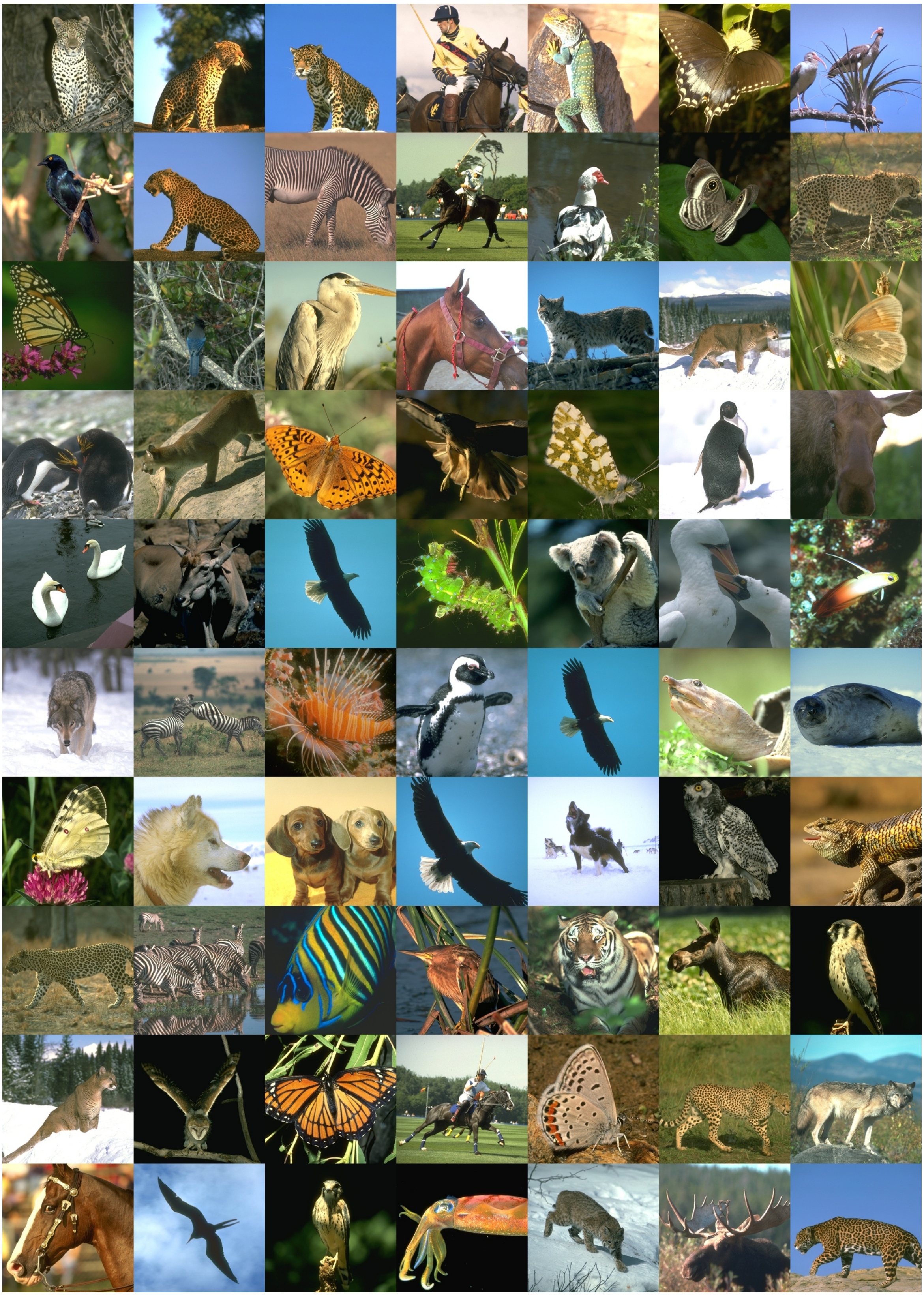

Illustration 46: The 200 "easy" animal images, no. 71-140

Illustration 47: The 200 "easy" animal images, no. 141-200

3.4.2.11. The 200 “Difficult” Animal Images

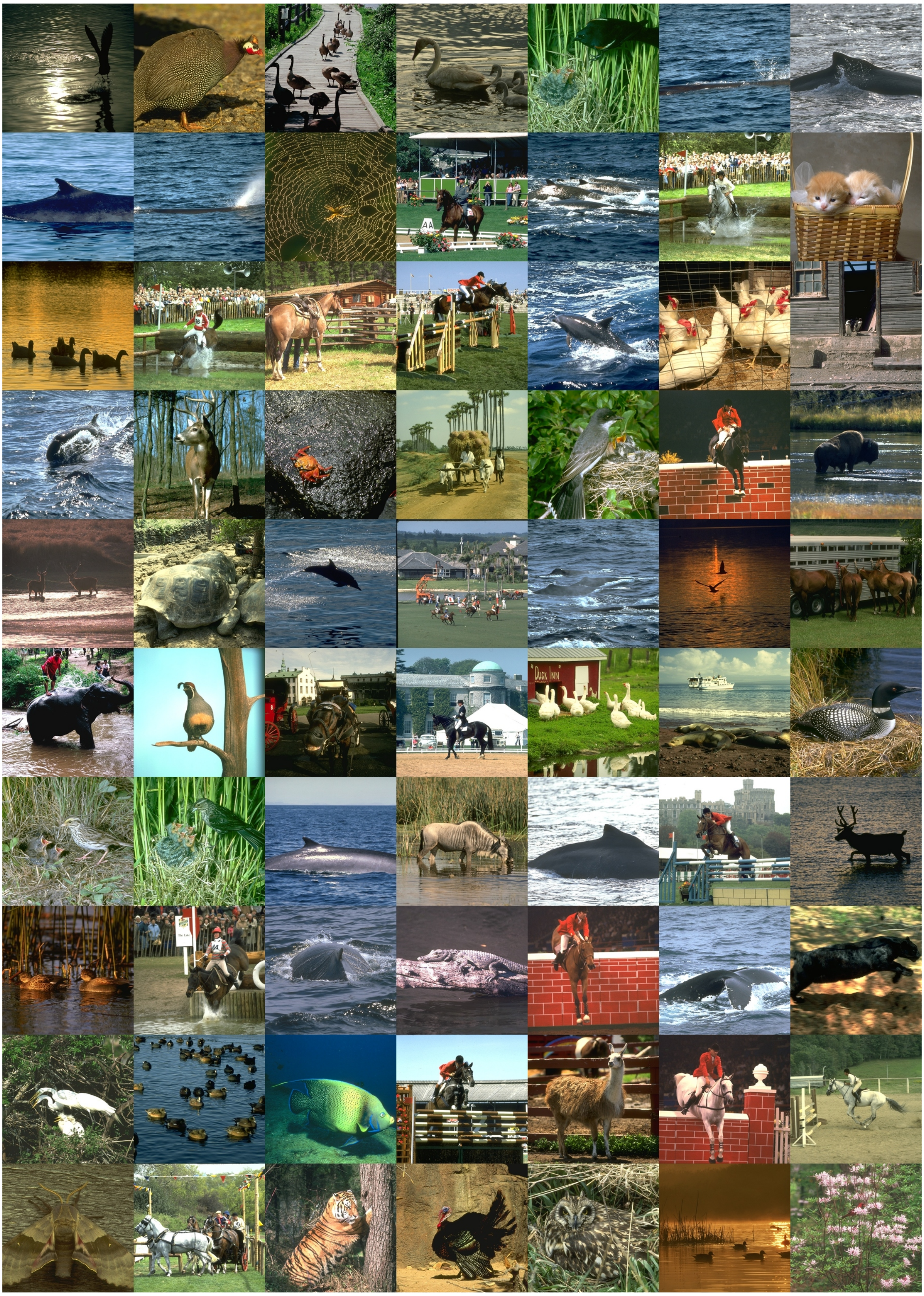

Illustration 48: The 200 "difficult" animal images, no. 1-70

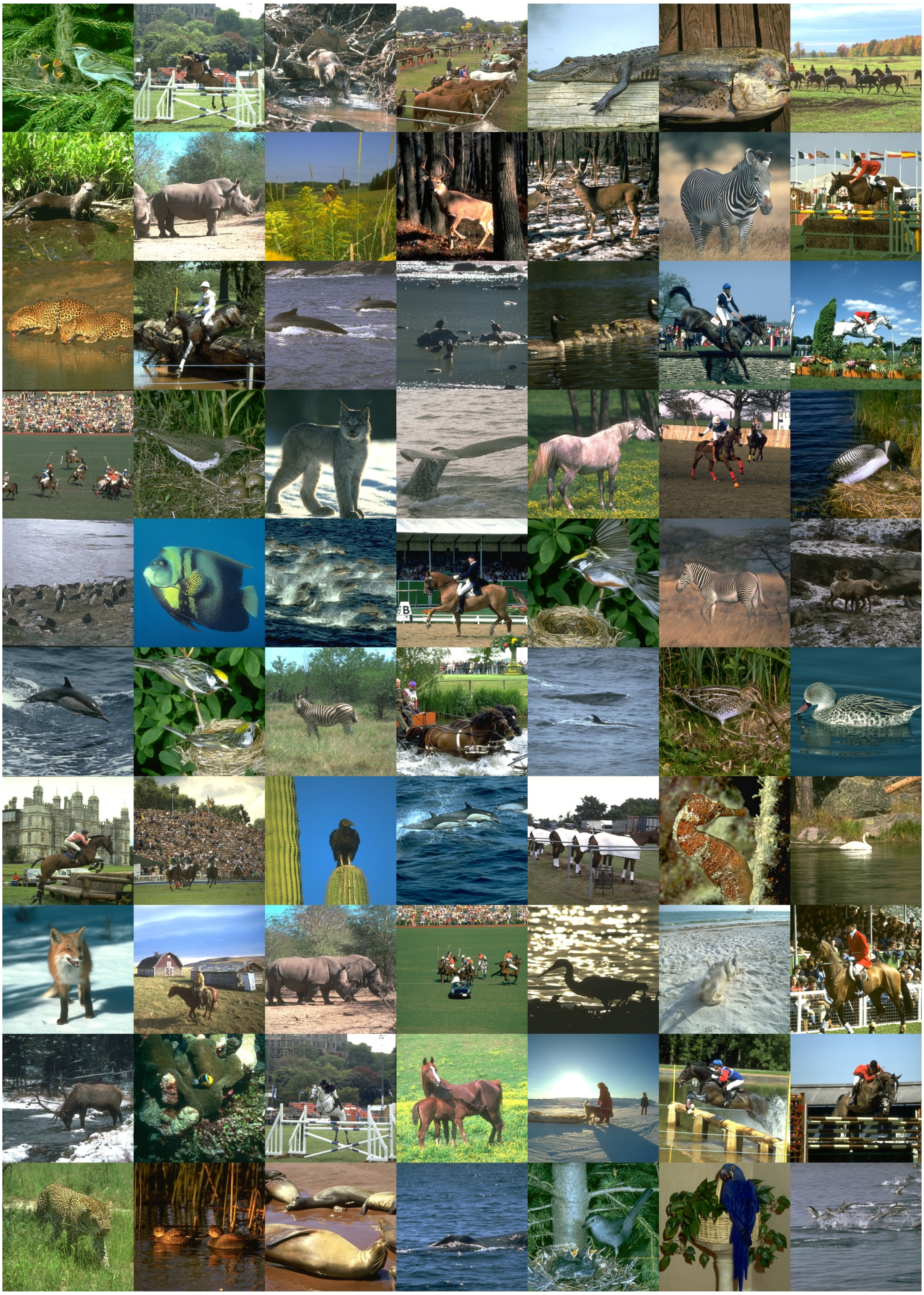

Illustration 49: The 200 "difficult" animal images, no. 71-140

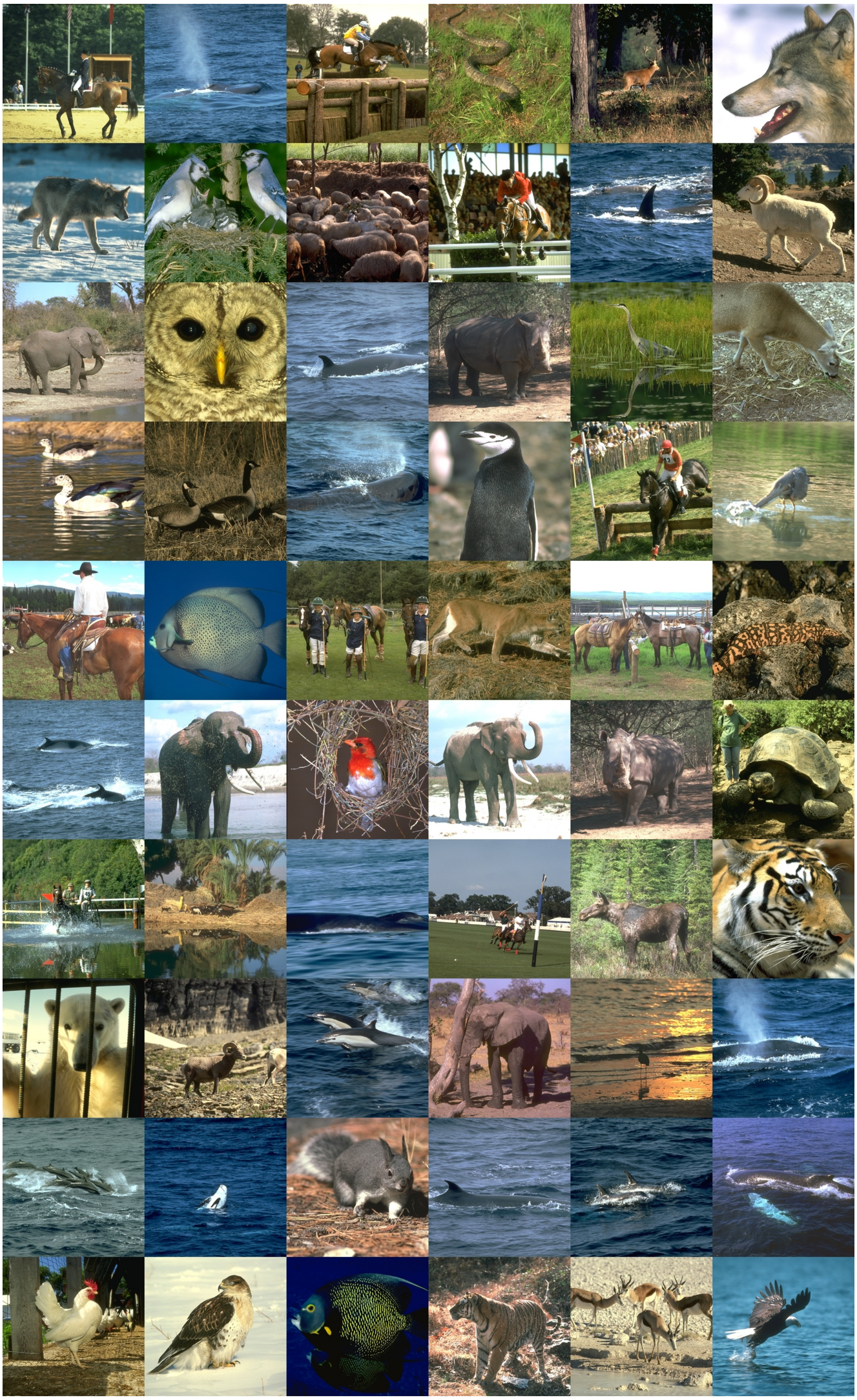

Illustration 50: The 200 "difficult" animal images, no. 141-200

3.4.2.12. The 200 “Easy” Non-Animal Images

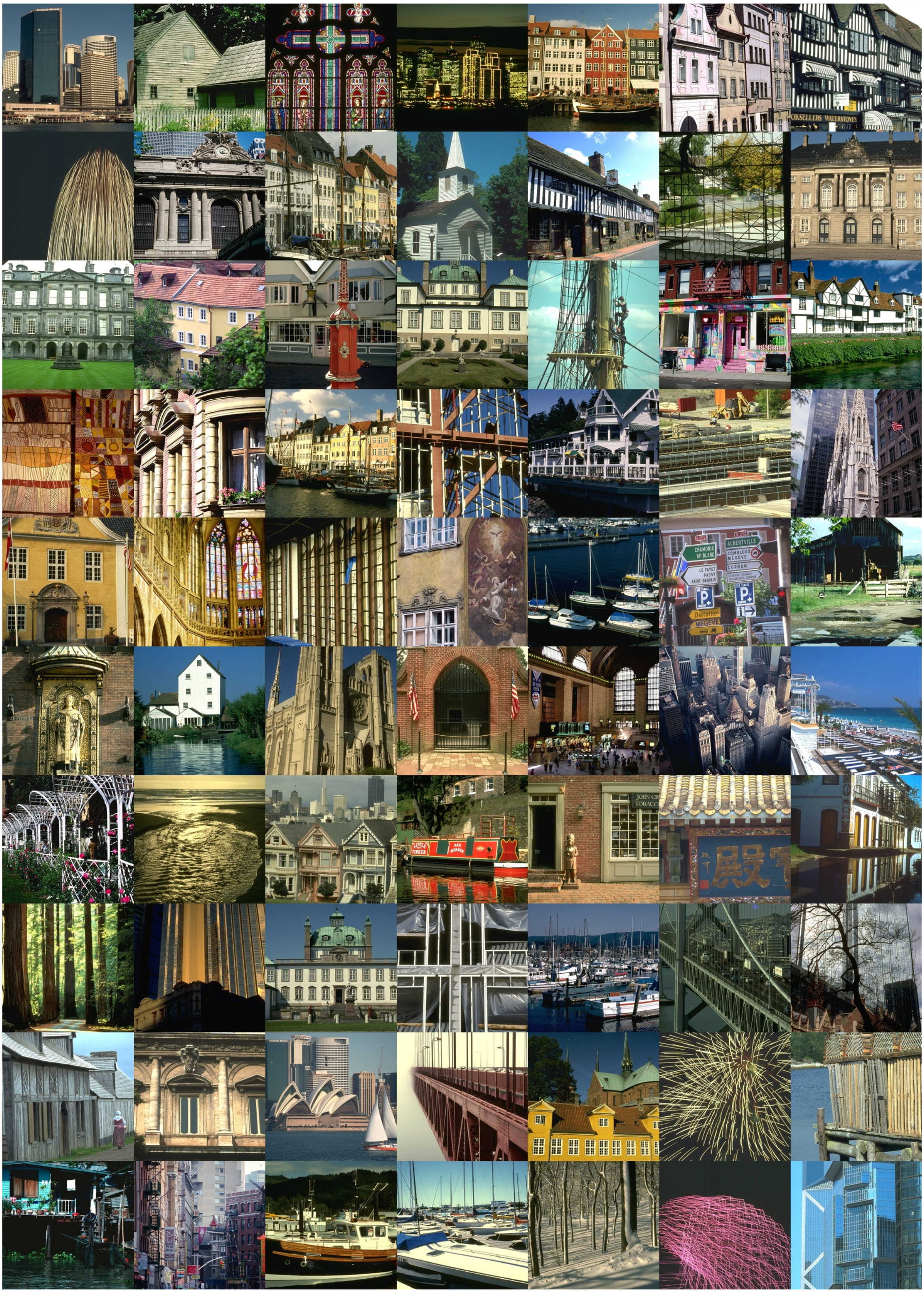

Illustration 51: The 200 "easy" distractor images, no. 1-70

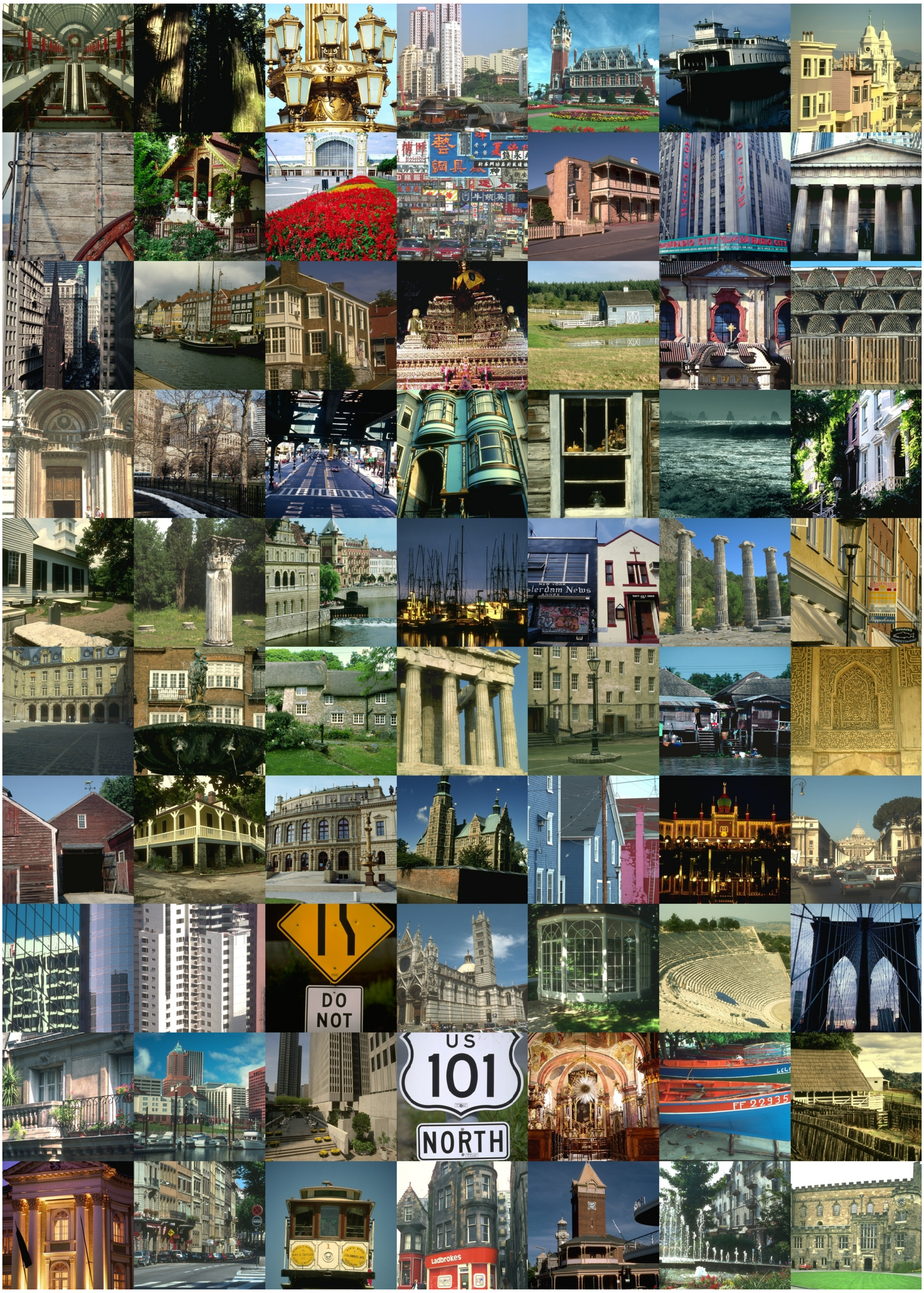

Illustration 52: The 200 "easy" distractor images, no. 71-141

Illustration 53: The 200 "easy" distractor images, no. 141-200

3.4.2.13. The 200 “Difficult” Non-Animal Images

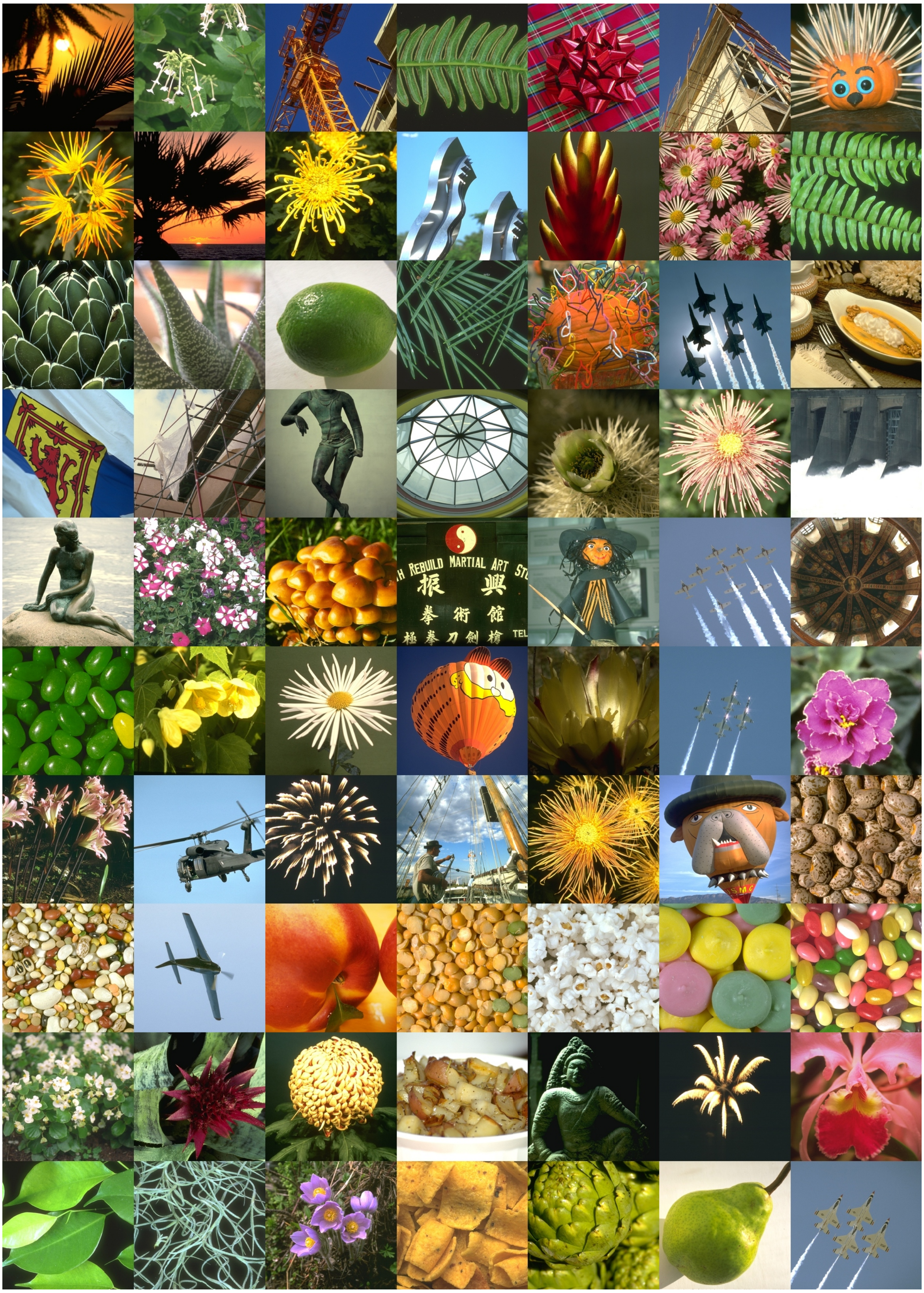

Illustration 54: The 200 "difficult" distractor images, no. 1-70

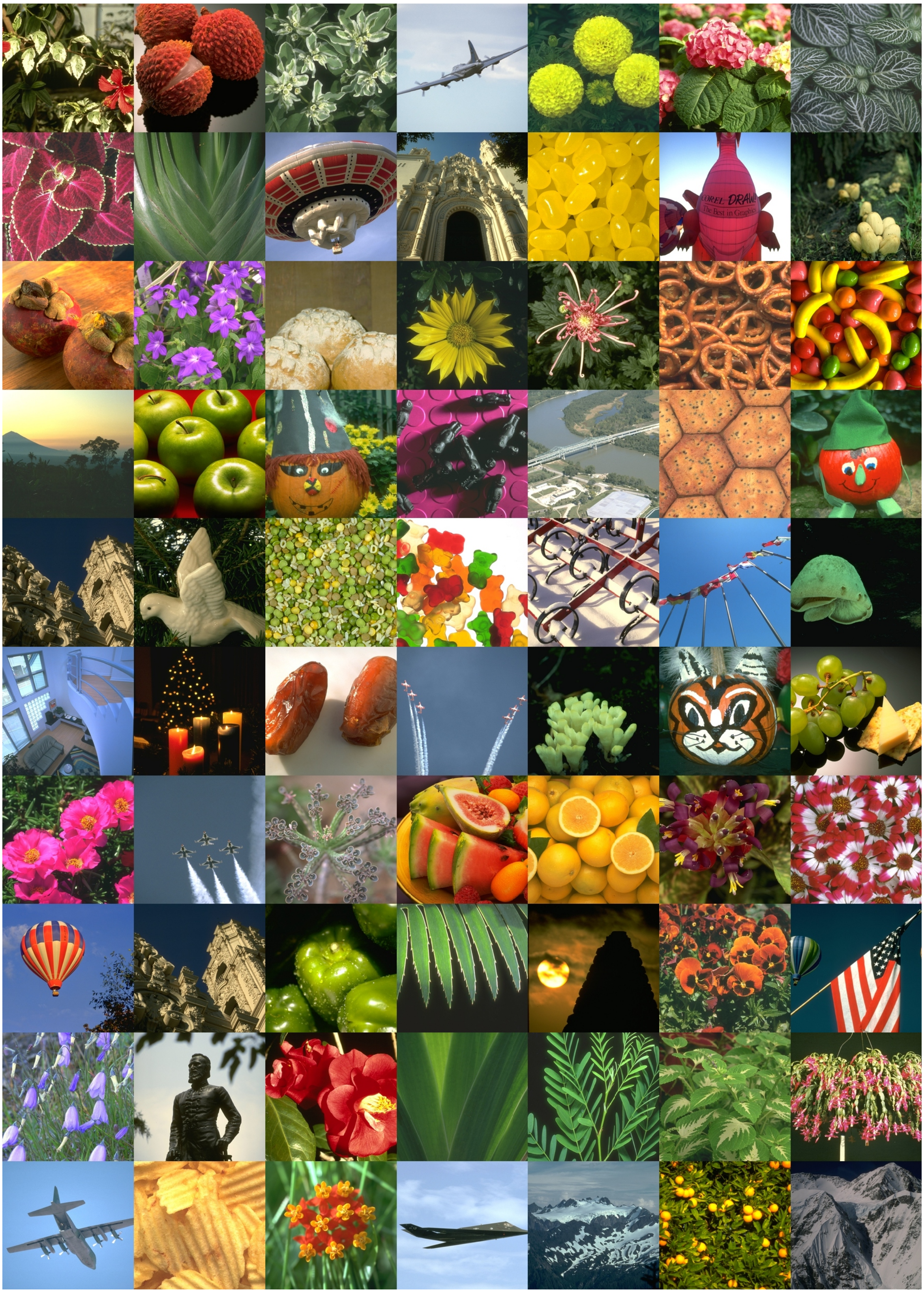

Illustration 55: The 200 "difficult" distractor images, no. 71-140

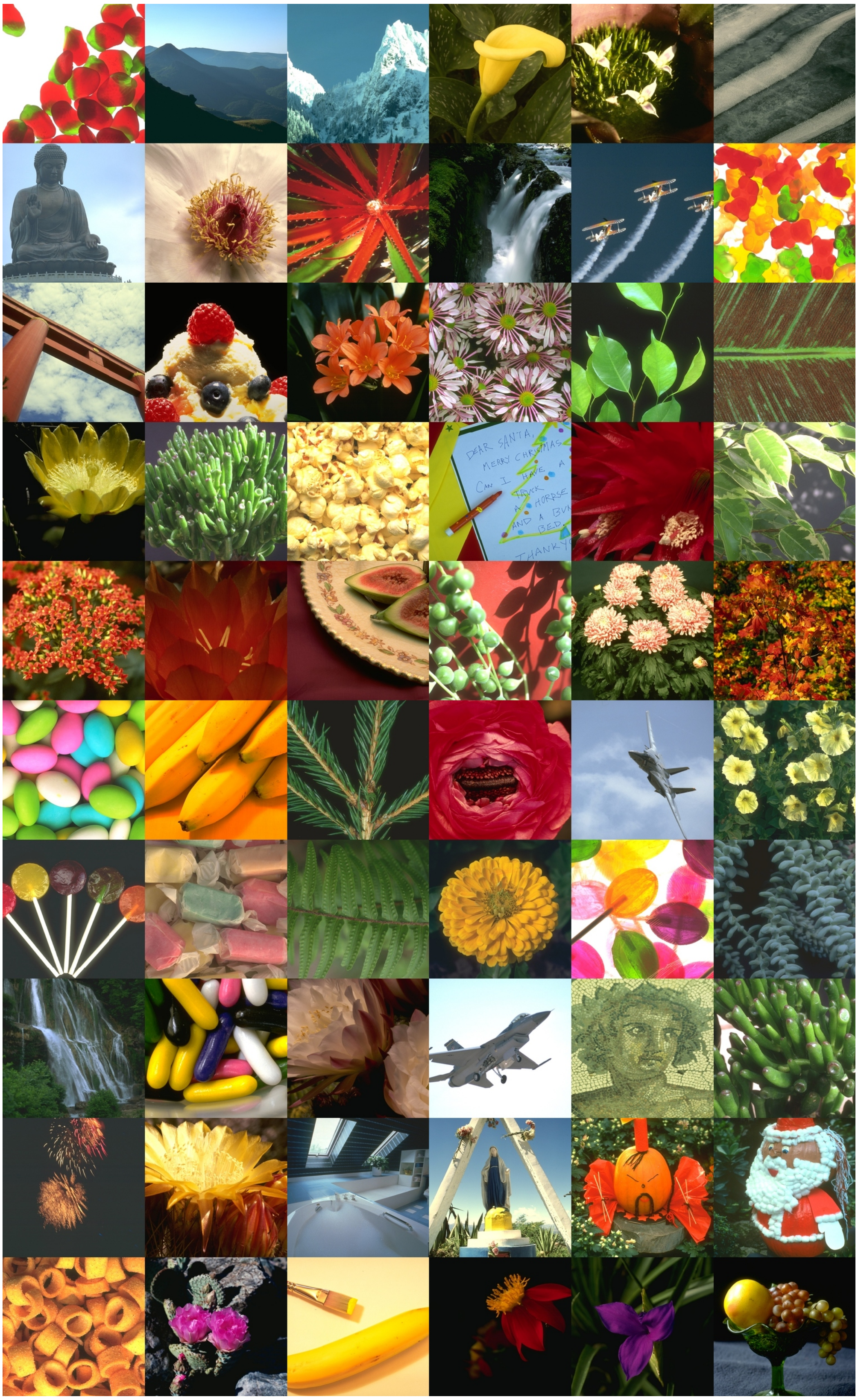

Illustration 56: The 200 "difficult" distractor images, no. 141-200

4. HUMAN CLASSIFICATION: COMPARING MAN AND MACHINE

The results of the computer classification as presented in the previous chapter are rather promising. This is especially true when considering the rather low computational complexity of the image preprocessing applied in the Fourier Fingerprint approach. In general, our results suggest that the global amplitude spectrum may very well be used to classify image content. The Fourier Fingerprint classifier therefore seems to be a suitable emulation function in our black box approach. We will now compare the input-output relation of our Fourier Fingerprint classifier with the one of the human visual system. First, we will compare the effect of image rotation on the computer classifier with the performance of human subjects on rotated images. Second, we will compare the image difficulty rating found in chapter 3.4.2.9 with the performance of human subjects on a special set of images selected by their computed difficulty. Finally, we will test if humans are still able to classify images after their amplitude spectrum has been neutralized as a cue to classification.

4.1. Experimental Equipment

All “2AFC” experiments were run on Microsoft Windows XP, using customized versions of an original program designed for similar purposes by Prof. Dirk Kerzel[1] in Microsoft Visual C++, making extensive use of the open-source “SDL” library. All presentation timings were synchronized to the screen refresh, which was set to 100Hz or 10ms per frame.

All eye-movement measurements were performed using SR Research's EyeLink II eyetracking system. Images were presented on an Iiyama VisonMaster 513 (MA203DT) 21” CRT screen. The viewing area was 40,3cm wide and 30,1 high, with a chin rest 45cm from the screen to stabilize head position, making for a usable visual field of 48,2deg by 36,9 deg. A resolution of 1280x960 pixels was chosen to best accommodate the 4:3 aspect ratio of the screen, delivering approximately 26,6 pixels per degree. Ambient lighting was not calibrated to a particular level, but great care was taken to ensure that the same lighting conditions were in effect for all subjects. A warm-up period of at least 1h allowed the CRT to stabilize its mean luminance. For the “Test for effects of image rotation in human subjects” (chapter 4.2) the screen was otherwise uncalibrated, for the “Test for effects of absence of the amplitude spectrum in humans” the screen was gamma corrected with a separate LUT for each color channel to linearize the luminance function.

Illustration 57: SR Research EyeLink II Eyetracker

Illustration 58: Experimental setup

[1] Prof. Dr. Dirk Kerzel, Université de Genève, Uni Mail, FaPSE, 40 bd du Pont d'Arve, CH-1205 Genève
Email: dirk.kerzel@pse.unige.ch, Fax: 41 (0) 22 – 3799229, Tel: 41 (0) 22 – 3799132, Bureau: 4134

4.2. Experiment 1: Test for Effects of Image Rotation in Humans

4.2.1. Motivation

In chapter 3.4.2 we discovered a rating of the bins of the Fourier fingerprint (Illustration 43, page 77). When taking a close look at this rating, we can see that the distribution of classification relevance of the bins is not rotation invariant (not all of the bins of a given frequency band are rated equally, the horizontal and vertical orientations outrank all others). Any classification mechanism based on a similar approach would therefore have to show a significant effect on image orientation. With our own algorithm, the effect would be so grave that we do not actually need to compute it – we can derive it from the rating of the bins our classifier produced:

The bins most important for the linear classifier are the ones representing the highest spatial frequencies in both the vertical (0°) and horizontal (90°) orientations, while the bins representing the oblique angles (45° and 135°) hardly add anything to the classification performance. If one were to test the presented classifier with rotated images whilst the classifier was trained on upright images, an effect on classification performance can be expected depending on the angle of rotation.

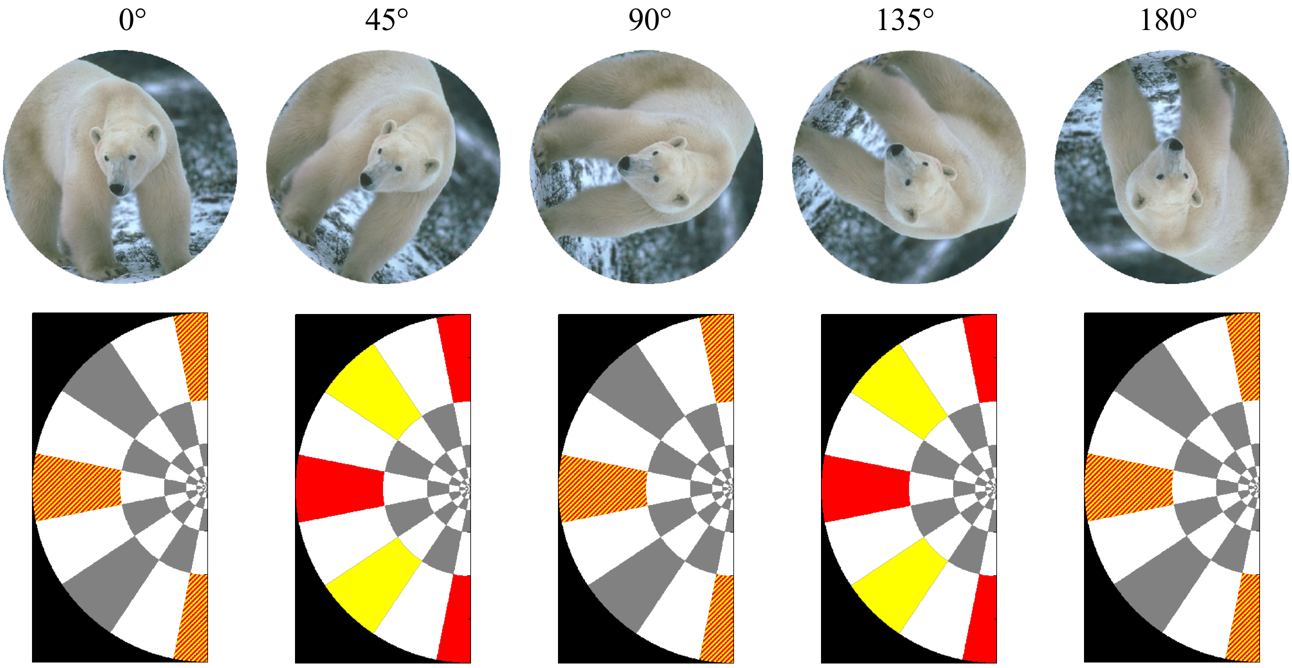

Illustration 59: Rotation of images and their expected effect on the linear classifier

Illustration 59 shows the effects of several different angles of rotation on the information content of the bins. The red color indicates the bins in which the (computer) classifier expects the information relevant for classification. The yellow color indicates the actual location of the expected information. In the 0° case, both are located in the same place as this, of course, is the very case the classifier was trained on. Good classification performance can be expected, as was reported previously.

In the 45° case, however, the relevant information has rotated out of its original location just as far as the image was rotated, now residing in the oblique angles. The classifier will thus classify on the information that used to go into the oblique angles, and in all likelihood no significant classification performance will be achieved.

In the 90° case, the information has moved on by another 45°, now residing in the opposite bin it used to be: horizontal and vertical orientations have been exchanged. We recall that in the mean amplitude spectra, the differences between animal and non-animal images is that the non-animal images have more energy in both the horizontal and vertical orientations. This means that both of the bins have the same orientation in our data space: more energy means less animal probability. One can assume that this will still produce classification results significantly above chance performance – or at least much better results than in the 45° case..

The 135° case is similar to the 45° one, just with the bin contents swapped; one can not expect significant classification performance from our classifier.

In the 180° case the symmetry of the amplitude spectrum manifests: the distribution of energy over all the bins is identical to the 0° case, causing classification performance to be identical as well.

In order to test if our linear spectral based classifier at all resembles the functionality of the human visual system, we decided to test a set of rotated images on human subjects. If the part of the human visual system that is used for image classification works on any kind of data similar to our Fourier fingerprint, an effect of rotation on classification accuracy or response time should be measurable.

4.2.2. Methods

To evaluate the performance of human subjects classifying rotated images, we chose to employ one of the established paradigms for such a purpose: the “Two Alternatives Forced Choice” gap paradigm (“2AFC”).

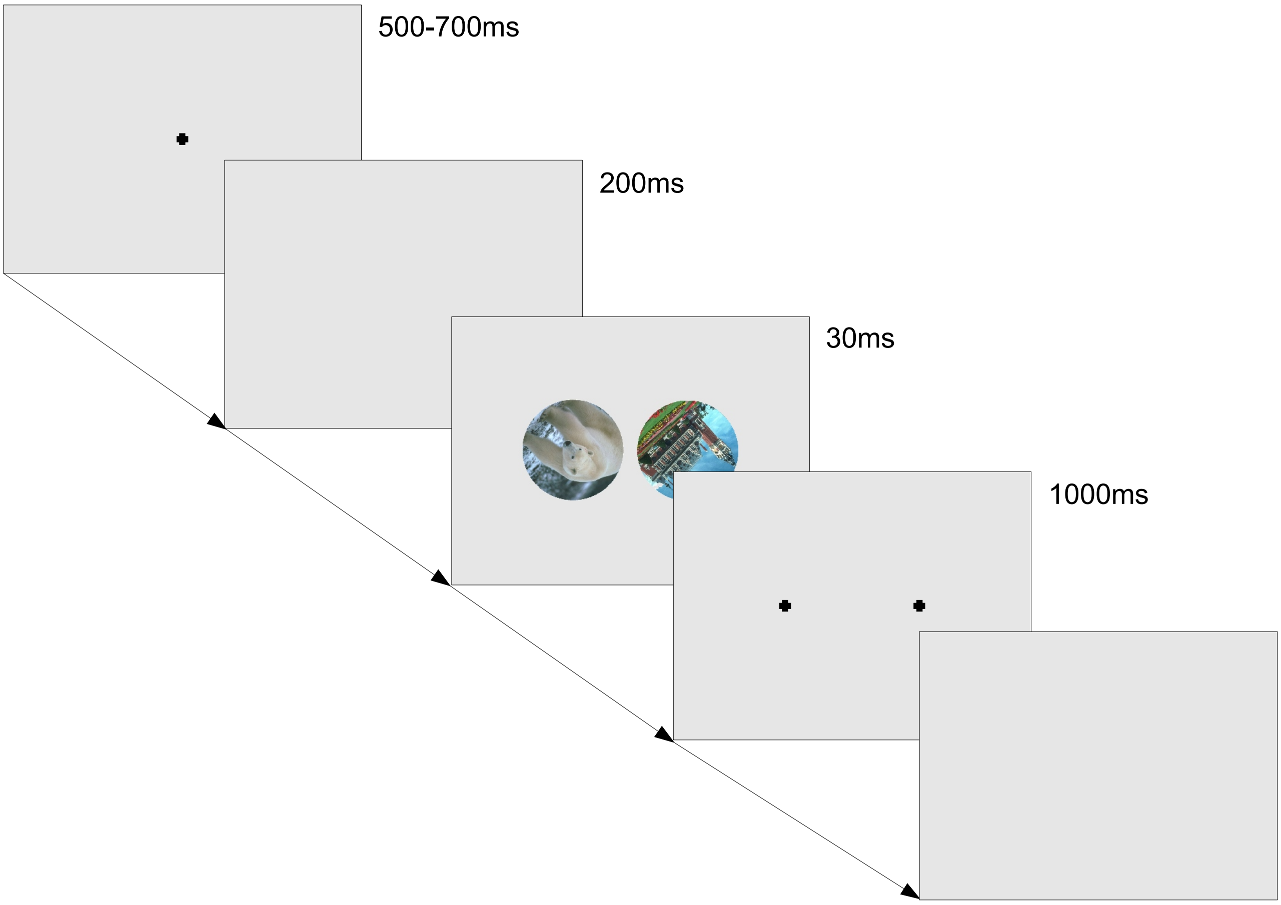

Illustration 60: Design of the 2AFC paradigm with rotated images

In this setup, very trial started with a neutral gray background showing only one fixation marker in the exact center of the screen. The subject then needed to push a button while fixating the marker. Fixation was checked with an EyeLink II Eyetracker (see chapter 4.1) and the trial was only started when good fixation was detected. After the start of a trial, the fixation marker remained on screen for a period of time randomly chosen between 500 and 700ms. After this, the fixation marker vanished, showing an empty screen for 200ms (gap period). Then, 2 images were briefly flashed (30ms), one on either side of the screen center. Images were about 9.5° in diameter (250 pixels), with the eccentricity being about 7° from image center to screen center. Images were vertically centered. After their disappearance, the images were replaced by position markers that served as saccade destinations for the subjects' decisions. After another 1000ms, these markers were replaced with a blank screen an the trial ended. In order to eliminate image borders as a clue to orientation, images were cut to circles. In every trial, exactly one image showed an animal (target), while the other showed a non-animal (distractor). Both images were rotated in the same random direction by the same randomly chosen angle of either 0, 45, 90, 135 or 180 degrees. Target and distractor were randomly distributed across the 2 possible locations so that the probability for the target to appear on either side of the fixation marker was 50%. Subjects were instructed to fixate the marker before initiating a trial by pushing a button, then to keep up fixation until the images appeared, after which they were to make the decision where the animal was as quickly as possible by performing a saccade to the position marker on the side of the screen where the animal image had been. As the next trial would not start without the subject pushing a button again, the overall pace of trials was automatically adapted to the subject's preference.

4.2.3. Results

23 subjects participated in this experiment. All subjects were students of the Justus-Liebig-University Giessen and were payed for their participation. Subjects were between 19 and 33 years of age and had normal or corrected to normal vision. All subjects were able to perform eye movements adequate to achieve calibration accuracy rated as “good” (the best rating) by the EyeLink II system. All subjects were naive, and even though some participated in more than one experiment, they were not informed about the purpose of the individual experiments until after they had completed all the trials they were scheduled for.

Of the 23 participating subjects, 2 were excluded for reasons of exceptionally bad performance, one for barely exceeding chance performance in any of the conditions and the other for producing less than 50% valid trials. The first subject's general behavior suggested a lack of interest in actually conforming with the experimental requirements, the latter reported a long-term medical condition in the tracked eye only after completing the experiment. 500 trials were scheduled for each subject, resulting in 10500 recorded trials from the remaining 21 subjects. Prior to evaluation, the individual results were filtered with the following criteria:

Goodness of fixation:
Fixation between trial start and stimulus onset was not allowed to diverge from the screen center by more than 70 pixels, no saccades were allowed prior to stimulus onset; this made certain that the subjects did not favor one of the locations where the images would be shown and that no saccades were made during the critical phases of the experiment, as this might have hampered visual perception. During this filtering step, 333 Trials were eliminated, leaving 10167 for further processing.

Goodness of saccade direction / destination
Decision saccades were required to end in one of the two smallest possible squares covering the area that was occupied by the target and distractor images. This ensured that the saccades actually were directed into the general area of either of the shown images and eliminated random saccades, e.g. from lack of concentration. Only 17 trials needed to be eliminated to fulfill the requirements, leaving 10150 trials for further processing.

Goodness of response time

Subjects needed to make their decision no later than 700ms after stimulus onset in order to limit conscious thought in the decision process. Responses also had to be slower than 80ms to eliminate random and too early eye movements (reaction times faster than 80ms can safely be assumed not to be based on stimulus content). 325 trials were removed because they were too fast, none were too slow, leaving a total of 9825 valid trials for evaluation. On the valid trials, the mean hit ratio over all conditions was 85,5%, the mean response latency (correct and incorrect) was 298,7ms. Clockwise and counter-clockwise rotations were treated as one.

Angle of rotation	*Number of valid trials*	*Correct trials*	*Incorrect trials*	*Mean latency*	*Mean latency correct*	*Mean latency incorrect*
0°	1962	1705 (86.90%)	257 (13.10%)	296.5ms	299.2ms	285.2ms
45°	1965	1669 (84.94%)	296 (15.06%)	296.0ms	298.0ms	309.0ms
90°	1955	1679 (85.88%)	276 (14.12%)	297.0ms	299.7ms	291.8ms
135°	1977	1675 (83.81%)	320 (16.19%)	301.5ms	302.8ms	305.5ms
180°	1966	1686 (85.76%)	280 (14.24%)	302.2ms	302.8ms	308.1ms

Table 3: General results of the test for rotational effects on classification performance

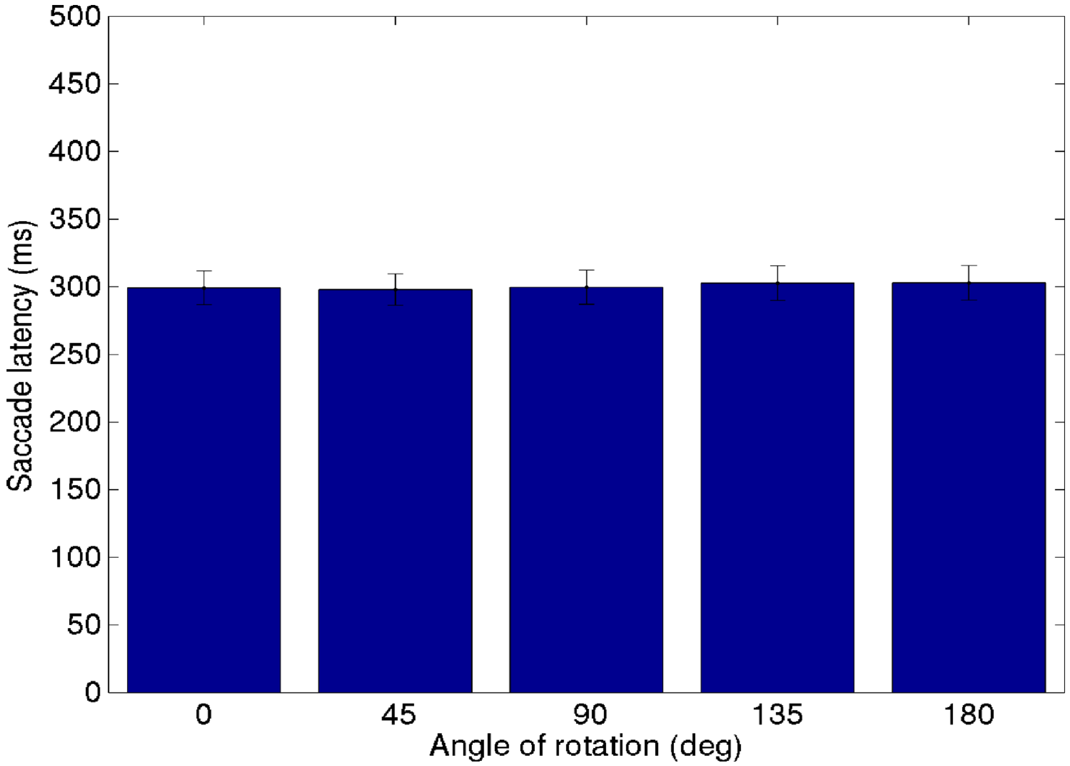

Illustration 61: Response latencies on rotated images (correct trials only)

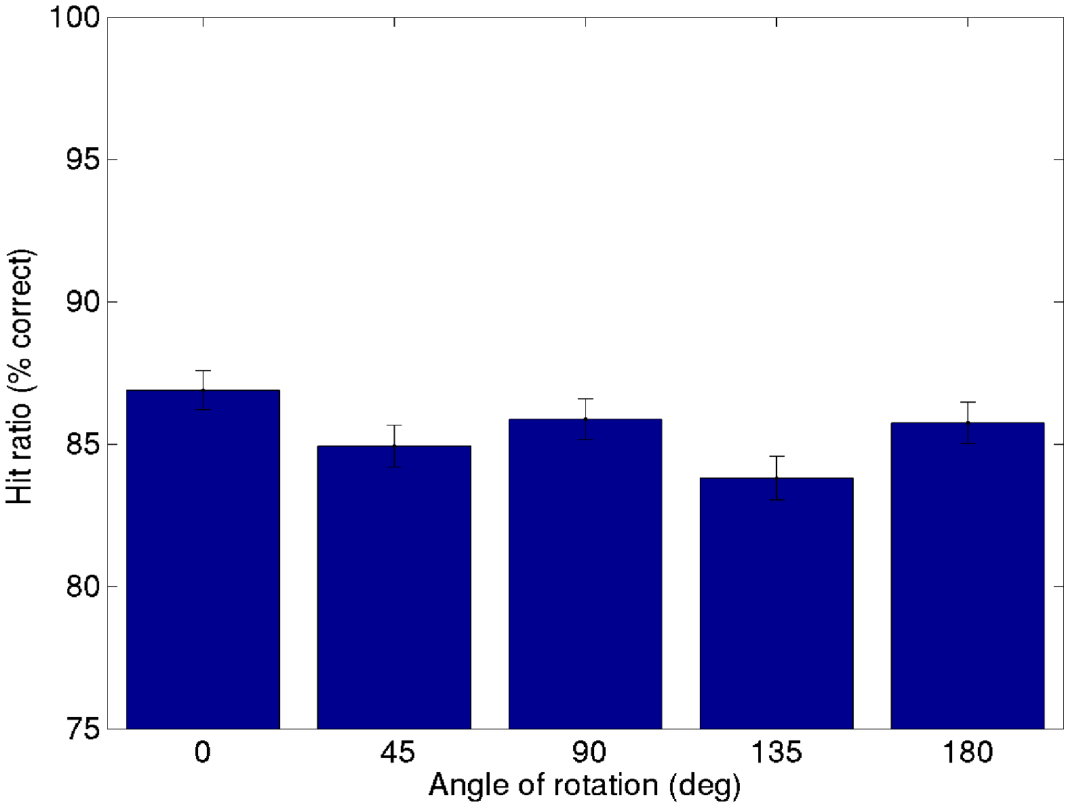

Illustration 62: Classification hit ratio on rotated images

Multiple statistical analysis including ANOVA and paired t-tests of the saccade latencies did not find a significant effect on the response latencies of our subjects (Illustration 61). However, a repeated-measures ANOVA showed a significant effect of the factor rotation on the hit ratios (*F*(4,80)=2.77, *p*=0.033, Illustration 62). Following up on this, we computed a series of correlated t-tests to further analyze the statistical relations in the experimental results.

First group of orientations	***Second group of orientations***	***p-value of correlated t-test***
0°	45°	*0.064*
0°	90°	0.436
0°	135°	**0.001**
0°	180°	0.209
0°,180°	45°, 135°	**0.002**
0°, 90°, 180°	45°, 135°	**0.004**

Table 4: Results of t-tests on grouped orientations

As can be seen from Table 4, the difference between the cardinal and the oblique orientations is significant. This is valid for all of the performed t-tests with an exception in the case of 0° vs. 45°, where the *p*-value of the t-test is 0.064, which is not significant per definition, but might still be considered a weak effect. We conclude that our measurements generally fulfill our expectations; our subjects' classification performance showed a significant dependency on image rotation, even though classification still reached high levels of accuracy even on the oblique rotation angles. These results support the hypothesis that the global amplitude spectrum does matter during the classification of natural scenes in humans.

4.3. Experiment 2: Test for Effects of Image Difficulty in Humans

4.3.1. Motivation

In chapter 3.4.2.9, we presented a subset of the images in our database selected by their supposed “classification difficulty”. The selection process was ultimately based on the expressiveness of the individual amplitude spectra of the images. When assuming that the human visual system utilizes the global amplitude spectrum in tasks of rapid classification, one might find a similar “easy” vs. “difficult” pattern with the classification performance of humans when exposed to these selected images. The following experiment is designed to analyze the hit ratios and response latencies associated with our selected “easy” and “difficult” images for effects of image difficulty in humans.

4.3.2. Method

We chose a Go/NoGo gap paradigm to evaluate human performance on our set of images selected by classification difficulty. Grayscale versions of the images were used for better comparability with later experiments. Subjects were instructed to push and hold a trigger button prior to each trial. When the button was pushed, a fixation dot appeared on the screen and remained there for a random period between 500 and 700 ms, followed by a 200ms gap. A single target (animal) or distractor (non-animal) image was then shown for 30ms without a subsequent masking. Thereafter, a small fixation cross was shown for 1000ms, during which the subjects were to make their decision by either holding the button pressed steadily (“NoGo”-response, signifying a distractor) or releasing the button as quickly as possible (“Go”-response, signifying a target). After a “Go”-response, the next trial would not start again until the subject pressed the button again to signal readiness, allowing the subject to individually pace the trial sequencing. Every of our 800 selected images was shown in random order exactly once to each subject, resulting in a total of 800 trials per subject, taking between 40 and 60 minutes of time depending on the subject's individual pace.

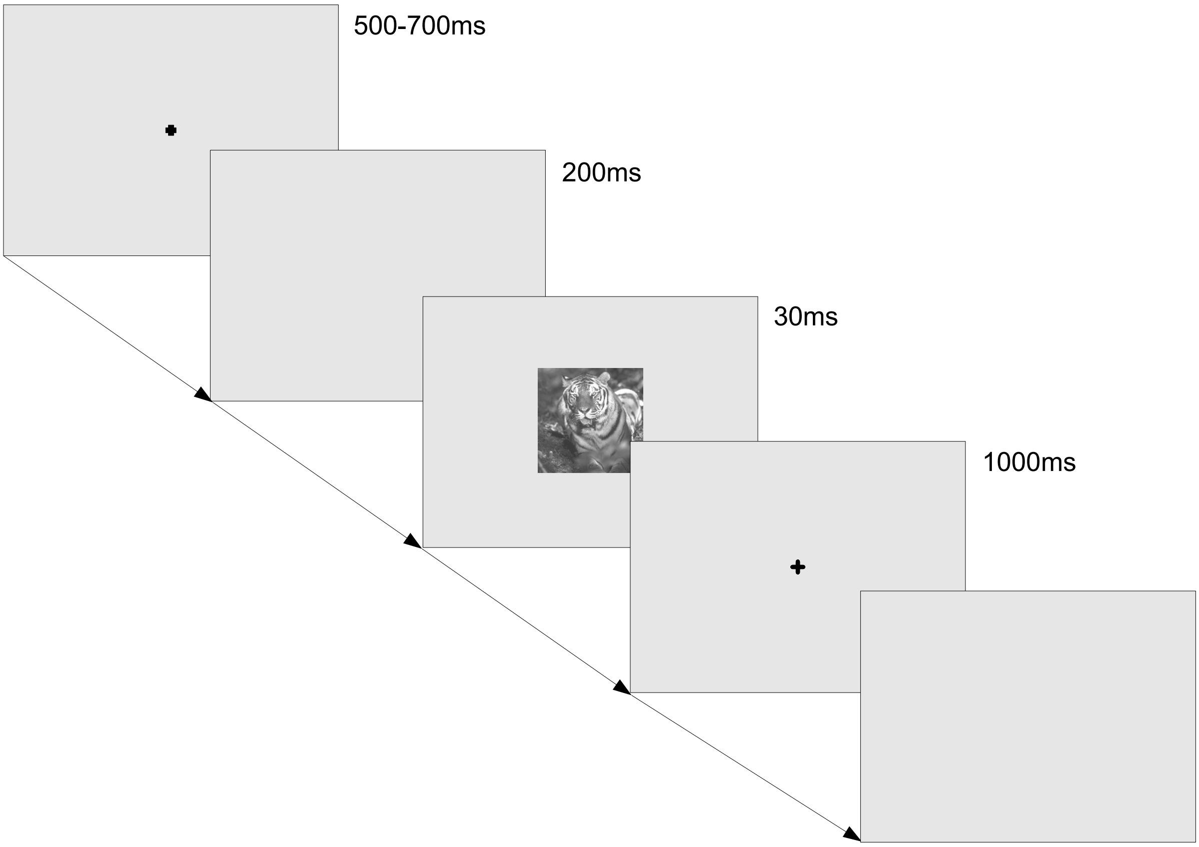

Illustration 63: Design of the "Go/NoGo"-paradigm with images of varying difficulty

4.3.3. Results

10 subjects participated in this experiment. All subjects were students of the Justus-Liebig-University Giessen and were payed for their participation. Subjects were between 19 and 31 years of age and had normal or corrected to normal vision. The means of the measurements taken are reported in Table 5. Overall 8000 trials were recorded, 5 of which were discarded because their response time was faster than 200ms. The remaining ones were considered valid.

Image type	***Hit ratio***	***Response latency***
all	93,9%	413,2ms
“easy”	96,7%	405,5ms
“difficult”	91,2%	421,1ms

Table 5: General results of the test for effects of image difficulty in humans

Generally, the measured response times and hit ratios were within the normally expected range for experiments of this kind. A highly significant effect of image difficulty was found within the subjects' hit ratios: “easy” images were classified more accurately than “difficult” ones (repeated measures one-way ANOVA, *F*(1,10)=68.594, *p*<0.0001). A highly significant effect was also found in the response latencies: “easy” images were classified faster than “difficult” ones (*F*(1,10)= 37.554, *p*<0.001).

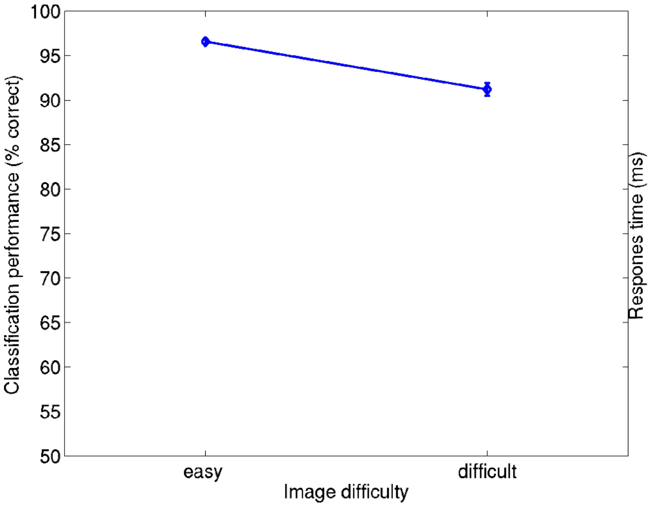

Illustration 64: Classification performance

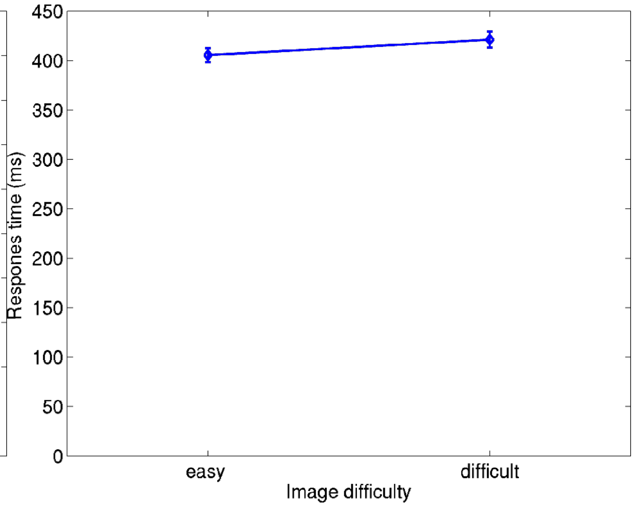

Illustration 65: Response latencies

Summarizing, our human subjects agreed with the difficulty assigned based on the Fourier fingerprint; this strongly supports (but does not prove) the idea that the amplitude spectrum actually matters when humans are performing rapid visual categorization tasks.

4.4. Experiment 3: Test for Effects of Absence of the Amplitude Spectrum in Humans

4.4.1. Motivation

In the previous experiments, we have found reasonable evidence that using the clues given in the amplitude spectrum of our natural scenes can indeed make our computer classifier behave quite similar to what we find in humans (when classifying the same images). What is still missing, however, is some hard evidence that either proves or disproves the use of the amplitude spectrum as a primary source of information for classification tasks within in the human visual system. It is a known fact that some form of a decomposition into spatial frequency and orientation is computed during the early stages of the visual system [Hubel Wiesel 1959] [Field Tolhurst 1986] [Field 1987] [Porat Zeevi 1988]. Also, this can be done in a feed-forward neural network, conforming with the time constraints associated with image classification during rapid serial visual presentation [Thorpe Fize Marlot 1996]. In summary, it is known that the information contained in the amplitude spectrum is *available*; It is not known, however, whether this information is actually *used* in the process of classification at all, or what the extent of it's use to the classification process is. In Experiments 1 and 2, our results support the hypothesis that the global amplitude spectrum matters in human rapid visual classification tasks, as we concluded from the similarities in classification behavior between man and machine. In the following two experiments the use or usefulness of the global amplitude spectrum for fast classification within the human visual system shall be analyzed with the goal of producing hard evidence to either prove or disprove the necessity of the use of the global amplitude spectrum.

4.4.2. Method

It can sometimes be quite difficult to prove that one particular cue is used in a decision process. In the case of the human visual system, the enormous amount of information involved together with the extreme speed of processing makes it even more difficult to examine the different streams of information and their use for individual tasks. In our case, doing the opposite may be simpler and more feasible: If one can isolate a particular informational cue, one can test a subjects' behavior with and without that cue present and compare the results. In other words, it may not be possible to determine whether an information is being used when it is readily available; however, in it's absence, an effect may be observed. We therefore need to eliminate the amplitude spectrum as a possible cue to classification without making the images unrecognizable to our subjects. It has been shown that human vision is much more susceptible to phase noise than to amplitude noise [Wichmann Braun Gegenfurtner 2006], so a modification of the global amplitude spectrum does not necessarily destroy the image content. We proceeded by decomposing our already introduced selection of “easy” and “difficult” images into their individual phase- and amplitude spectra. We completely disregarded the latter, replacing it with the mean amplitude spectrum of all the images in our database, thus removing all cues that the global amplitude spectrum might have provided (see Illustration 66). Any classification mechanism based on the global amplitude spectrum would utterly fail after this procedure. After recomposition of the images from their individual phase spectra and the mean amplitude spectrum, the appearance of the images is somewhat noisy, but the content of the scene is still recognizable in every single case (see Illustration 67). Wanting to evaluate our subjects performance in oder to analyze both for the effect of filtering per se as well as the effect it might have on the previously introduced factor of image difficulty, we designed two experiments. The first (experiment 3a) utilizes a Go/NoGo paradigm very similar to the one from experiment 2, with the only difference being that 50% of all images were shown in their amplitude-normalized version instead of their original, unchanged form (see Illustration 68). The instructions given to our subjects included the information that some of the images might look a bit odd, but otherwise the same task was required. Every image was shown only once and in only one form (amplitude-normalized or not) per subject, alternating between subjects. This setup allows us to examine the effect of the amplitude normalization on single images. For the second experiment (experiment 3b) we employed a version of the 2AFC gap paradigm (see Illustration 71) similar to the one presented in chapter 4.2, replacing rotated images with pairs of (upright) amplitude-normalized and pairs of unchanged images (all images were normalized for RMS contrast and standard deviation). Image pairs were evenly distributed between the 4 possible combinations of “easy” and “difficult” image types, allowing for an analysis of the factor “image difficulty” both in “unchanged” and “amplitude-normalized” condition. Every pair of (randomly chosen) target and distractor images was shown exactly once to each subject, in such way that every image would appear equally often in it's amplitude-normalized or it's unchanged form over the duration of the experiment, but only once and in only one condition per subject. 400 trials were scheduled for each subject, taking about 30-45 minutes of time. This setup allows us to examine the effect of the amplitude normalization on pairs of images; also, the timing of eye-movement based decisions might be more conclusive than button-press responses due to the increased speed of the responses.

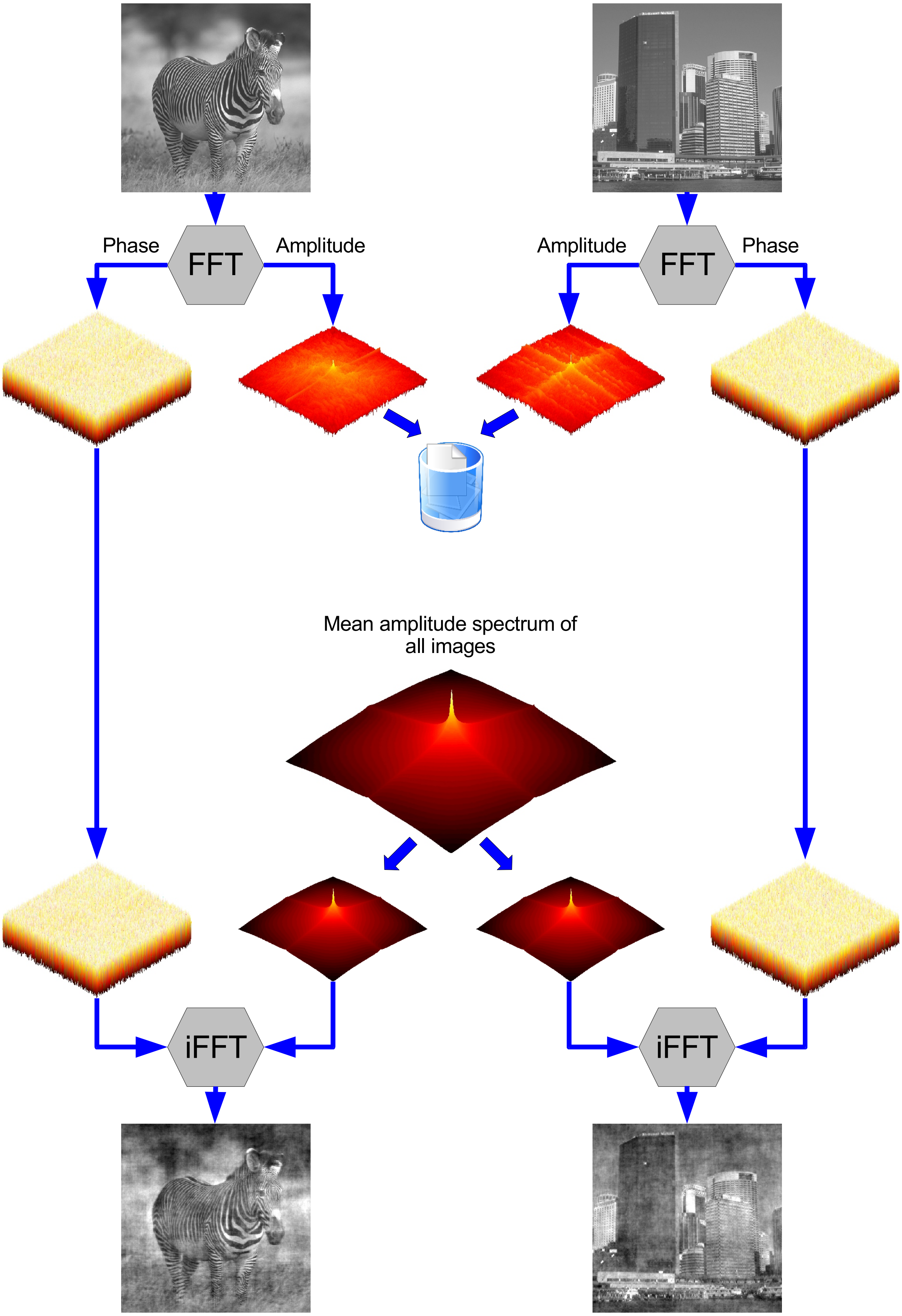

Illustration 66: Schematic of the amplitude spectrum normalization process

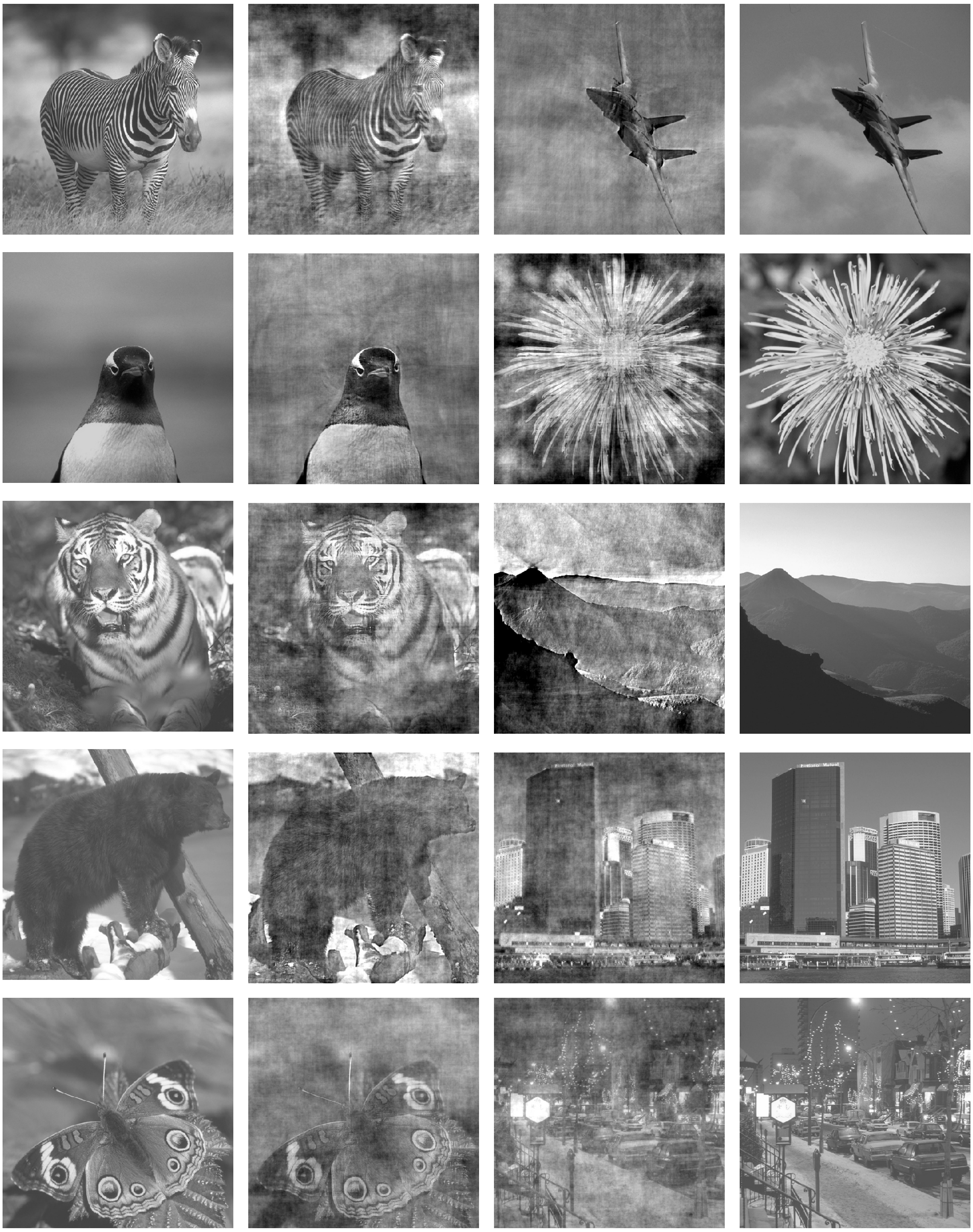

Illustration 67: Samples of images before and after replacement of amplitude spectrum

4.4.3. Results

4.4.3.1. Experiment 3a

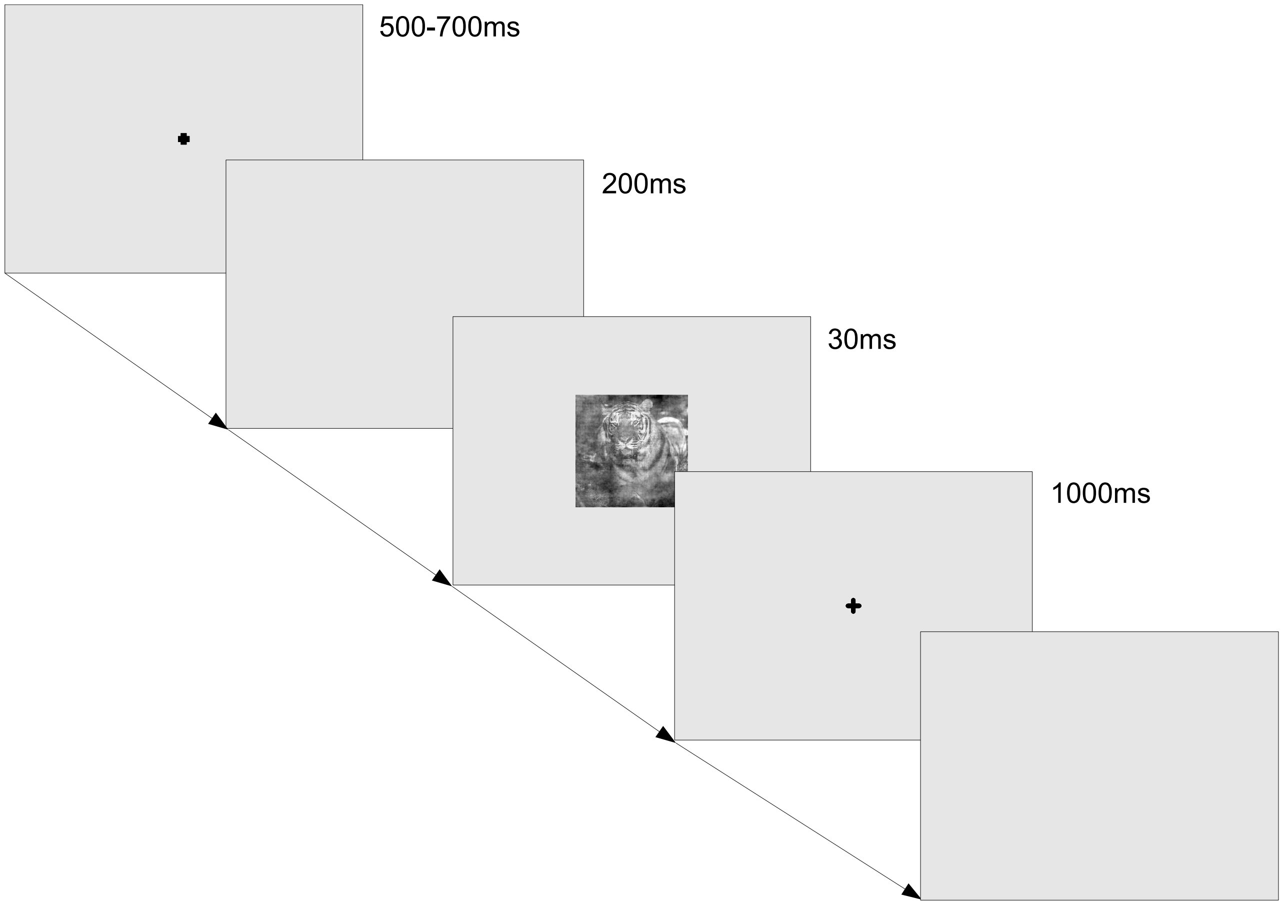

Illustration 68: Design of the Go/NoGo paradigm with amplitude-normalized images (Experiment 3a)

12 subjects participated in this experiment. All subjects were students of the Justus-Liebig-University Giessen and were payed for their participation. Subjects were between 20 and 31 years of age and had normal or corrected to normal vision. The means of the measurements taken are reported in Table 6. Overall 9600 trials were recorded, 53 of which were discarded because their response time was faster than 200ms. The remaining trials were considered valid.

Amplitude spectrum	***Image type***	***Hit ratio***	***Response time***
all	all	89.3%	450.6ms
unchanged	easy	95.2%	437.7ms
	difficult	90.5%	451.1ms
	all	92.7%	443.5ms
amplitude-normalized	easy	90.1%	449.3ms
	difficult	81.6%	464.4ms
	all	86.0%	457.7ms

Table 6: General statistics of experiment 3a

In the “unchanged” condition, we find the difference between “easy” and “difficult” images to be highly significant both in terms of hit ratio (repeated measures one-way ANOVA, $F(1,9)=30.621$, $p<0.001$) and response time ($F(1,9)=8.569$, $p=0.014$), reproducing our results from experiment 2. When performing the same analysis on the data recorded with the amplitude-normalized images, we still find a highly significant effect for both hit ratio ($F(1,9)=119.301$, $p<0.0001$) and response time ($F(1,9)=27.584$, $p<0.001$). We also find a very strong effect for the amplitude-normalization in both hit ratio ($F(1,9)=131{,}8567$, $p<0.001$) and response time ($F(1,9)=21{,}2665$, $p<0.001$), as was to be expected from the random-like noise added by the spectral equalization procedure. The seemingly significant effect of the amplitude-normalization on image difficulty appears in our hit rate data only ($F(1,9)=10.1406$, $p=0.0087$); we believe this to be a ceiling effect, as hit rates in the unchanged condition are well above 90%, even above 95% for the “easy” images. At this high performance levels, the difference between the “easy” and the “difficult” images gets compressed – the response times, however, have not close enough to their ceiling (or floor, respectively) and therefore do not show this effect. This shows that human classification performance on images with non-conclusive amplitude spectra does not work as well as it does on unchanged images; however, though statistically significant, the absolute difference between the two conditions is only about 8%, still allowing for a classification accuracy of 86% on average.

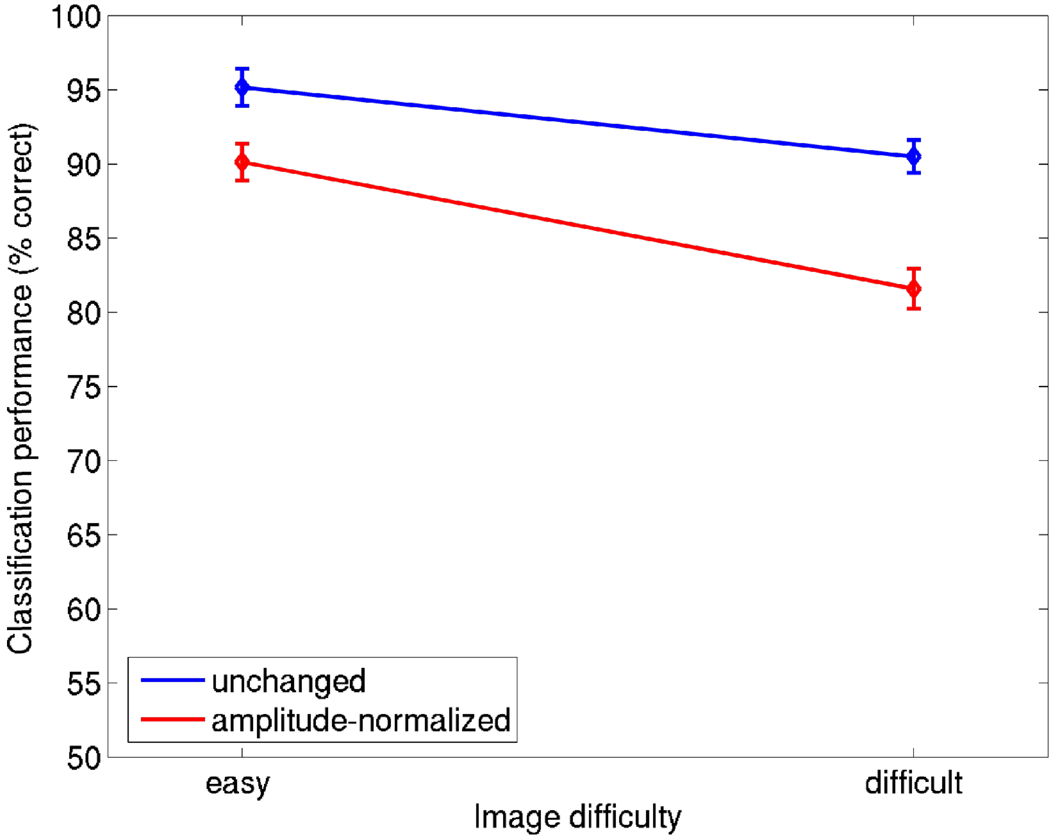

Illustration 69: Classification performance: hit ratio (Experiment 3a)

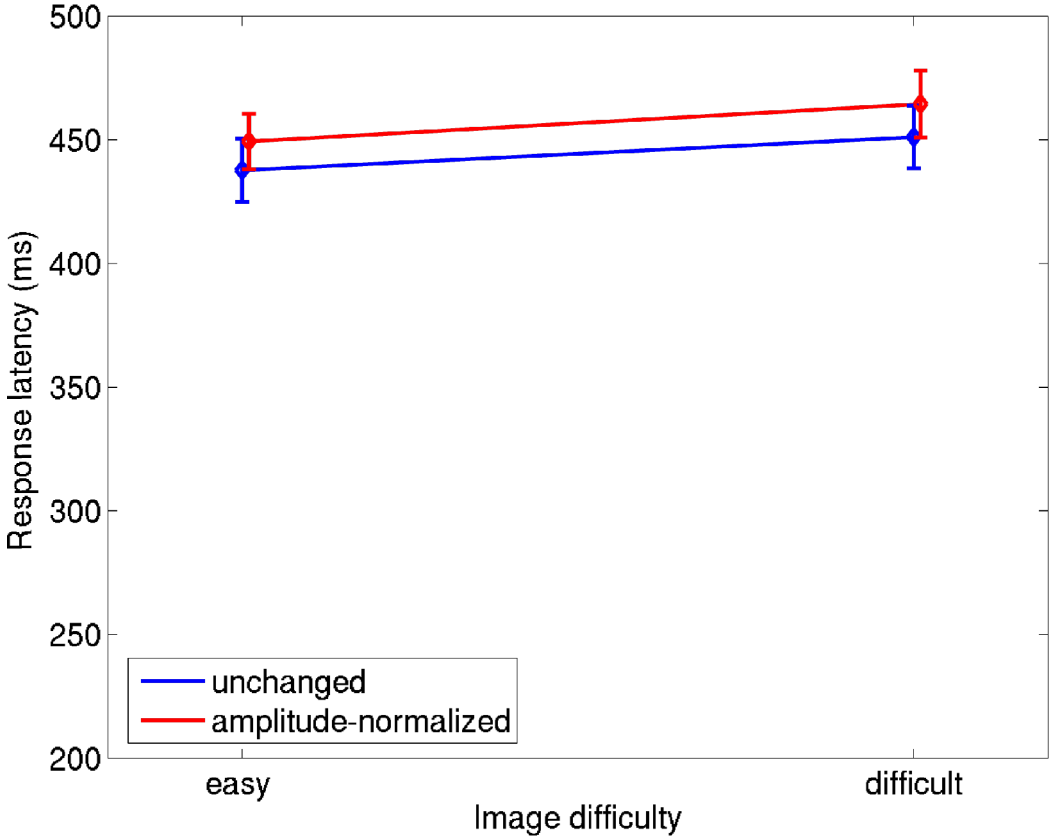

Illustration 70: Classification performance: latencies (Experiment 3a)

4.4.3.2. Experiment 3b

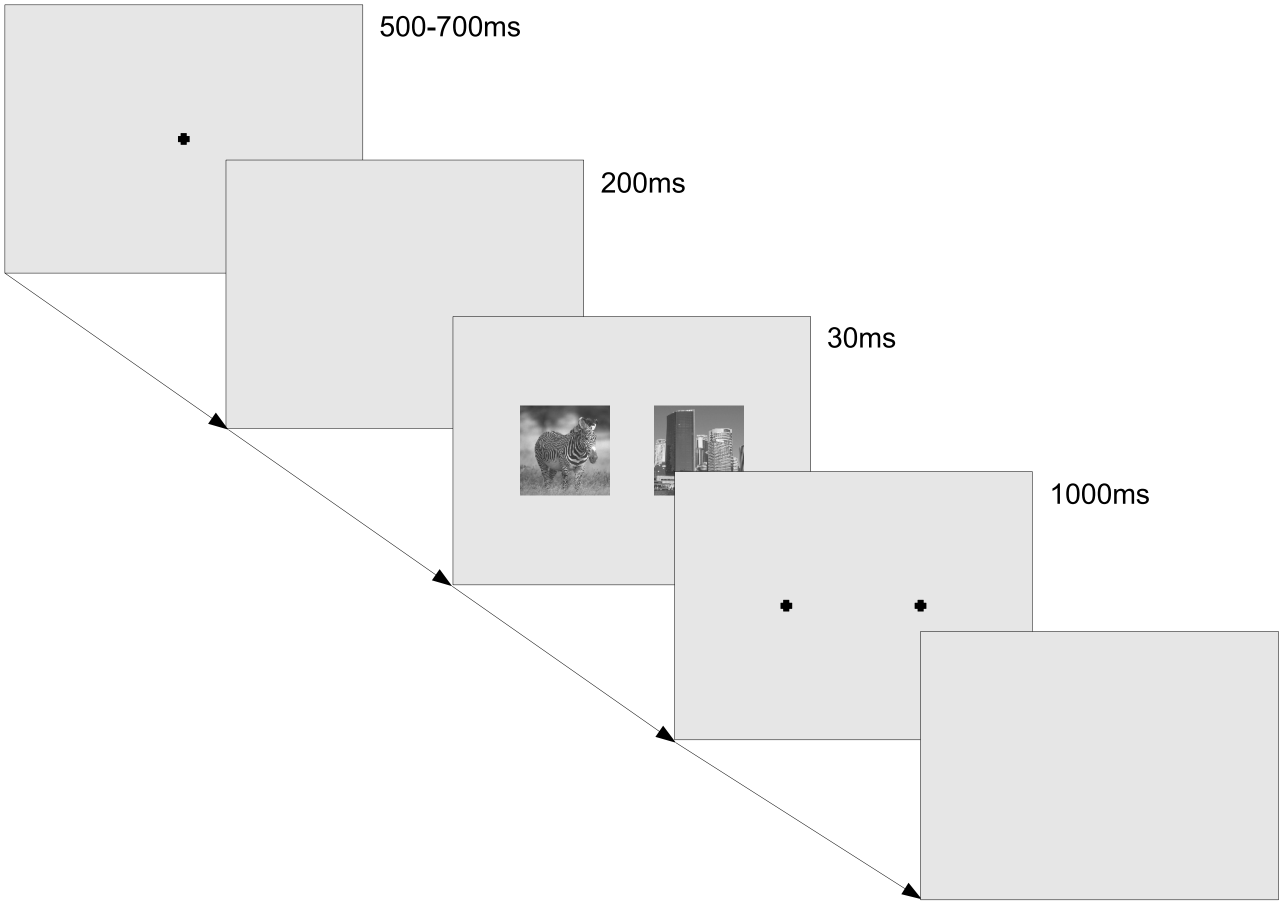

Illustration 71: Design of the 2AFC paradigm with amplitude-normalized images (Experiment 3b)

10 subjects participated in this experiment. All subjects were students of the Justus-Liebig-University Giessen and were payed for their participation. Subjects were between 20 and 31 years of age and had normal or corrected to normal vision. All subjects were able to perform eye movements adequate to achieve calibration accuracy rated as "good" (the best rating) by the EyeLink II system. All subjects were naive, and even though some participated in more than one experiment, they were not informed about the purpose of the individual experiments until after they had completed all the trials they were scheduled for. Prior to evaluation, we applied the same filtering to our measurements as explained chapter 4.2. Of the overall 4000 trials recorded, we discarded 129 for not fulfilling the "goodness of fixation" criterion, another 7 because of the "goodness of saccade direction / destination" criterion, and finally 26 more because of the "goodness of response time" criterion, all of the latter being too fast rather than too slow. In Summary, 3738 trials were considered valid. Measurements are reported in Table 7.

	Type of image pairing	***Hit ratio***	***Response latency***
all	all	81.0%	277.5ms
unchanged	all	85.8%	278.4ms
	easy - easy	95.3%	269.9ms
	easy - difficult	81.9%	281.9ms
	difficult - easy	89.7%	269.5ms
	difficult - difficult	76.3%	288.8ms
amplitude-normalized	all	76.3%	276.8ms
	easy - easy	89.2%	269.3ms
	easy - difficult	74.9%	279.4ms
	difficult - easy	80.6%	276.9ms
	difficult - difficult	60.5%	282.0ms

Table 7: Statistics of Experiment 3b

As we expected from the results of experiment 3a, the effect of image difficulty on the hit ratios was significant; additionally, we can report that the effect of image difficulty is significant both with the target images (animals, $F(1,9)=32.614$, $p<0.001$) and the distractor images (non-animals, $F(1,9)=85.685$, $p<0.001$). The general effect of the amplitude-normalization was also significant ($F(1,9)=58.599$, $p<0.001$), as can be seen from Illustration 72. The also significant effect of the amplitude-normalization on the effect of image difficulty is likely to be a ceiling effect, similar to experiment 3a. With the response times, the results are slightly different: While significant effects can be found with both target images (animals, $F(1,9)=6.033$, $p=0.036$) and distractor images (non-animals, $F(1,9)=42.174$, $p<0.001$), the general effect of amplitude-normalization is not significant ($F(1,9)=0.173$, $p=0.687$, see Illustration 73).

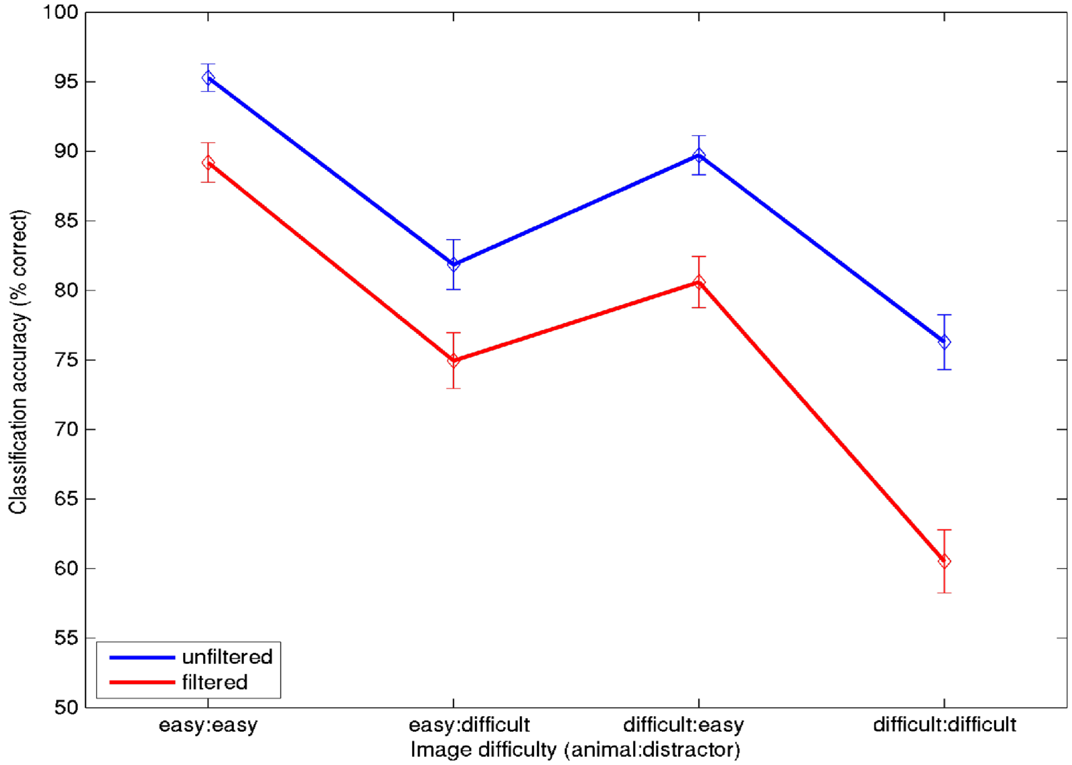

Illustration 72: Hit ratios on pairs of images (Experiment 3b)

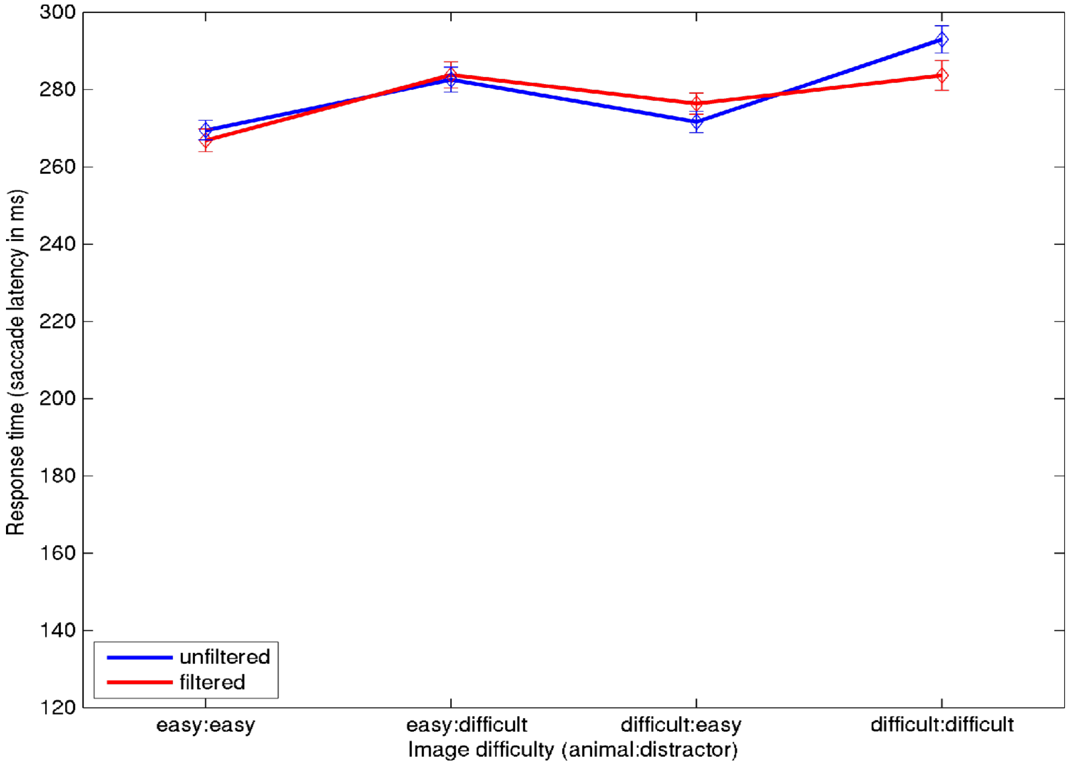

Illustration 73: Response latency on pairs of images (Experiment 3b)

5. CLASSIFICATION USING LOCAL IMAGE STATISTICS

In the previous chapters it was shown that the human visual system does not depend on the global amplitude spectrum for rapid visual classification. The information the human visual system uses must therefore be localized. However, a classifier based solely on the global amplitude spectrum can behave remarkable similar to the human visual system, a fact that suggests that the key elements of the amplitude spectrum – frequency, orientation and amplitude – may still be an appropriate way to represent an image in a computer classification task. Our next step will be to apply a filter to our images, extracting information about the amount of energy in a number of frequency and orientation bands (in this aspect similar to chapter 3.4.2), but retaining information about the spatial location within the image. A suitable way of doing this is the “Steerable Pyramid”, as introduced by Simoncelli and Freeman [Simoncelli Freeman 1995].

5.1. The Steerable Pyramid

The Steerable Pyramid has many possible applications in image processing. In our context, the most important aspects are frequency and orientation separation while still retaining local information. This is achieved through the use of a set of localized filters designed to optimally represent different orientation and frequency bands at each pixel of the image, while keeping both data over-representation and filter aliasing at a minimum. Notably, the resulting values are represented at each pixel of the image for every orientation and frequency band, resulting in a much more elaborate and very high-dimensional representation of the treated images. A side effect of the filters used is that there will be a high-pass and a low-pass version of the treated image that are not orientation-filtered. The details about the mathematical principles used in the Simoncelli Pyramid can be found in [Simoncelli Freeman 1995]. The pyramid images as used in this work were computed in Matlab, using a toolbox provided by E. Simoncelli[1]. After our standard preprocessing (see chapter 2.1), images were scaled to 576x576 to allow for 6 frequency bands, realized through the image sizes of 576, 288, 144, 72, 26 and 18 pixels. The high-pass image is of size 576 as well, but does not contain enough spectral bandwidth for another set of filters; the low-pass image is 9 pixels and therefore too small for the filter matrix. Illustration 74 shows the output of the Simoncelli Pyramid computed on a simple geometrical sample image to illustrate the filter characteristics, while Illustration 75 shows the same computation on a sample animal image.

[1] E. Simoncelli, “The Steerable Pyramid”, http://www.cns.nyu.edu/~eero/steerpyr/

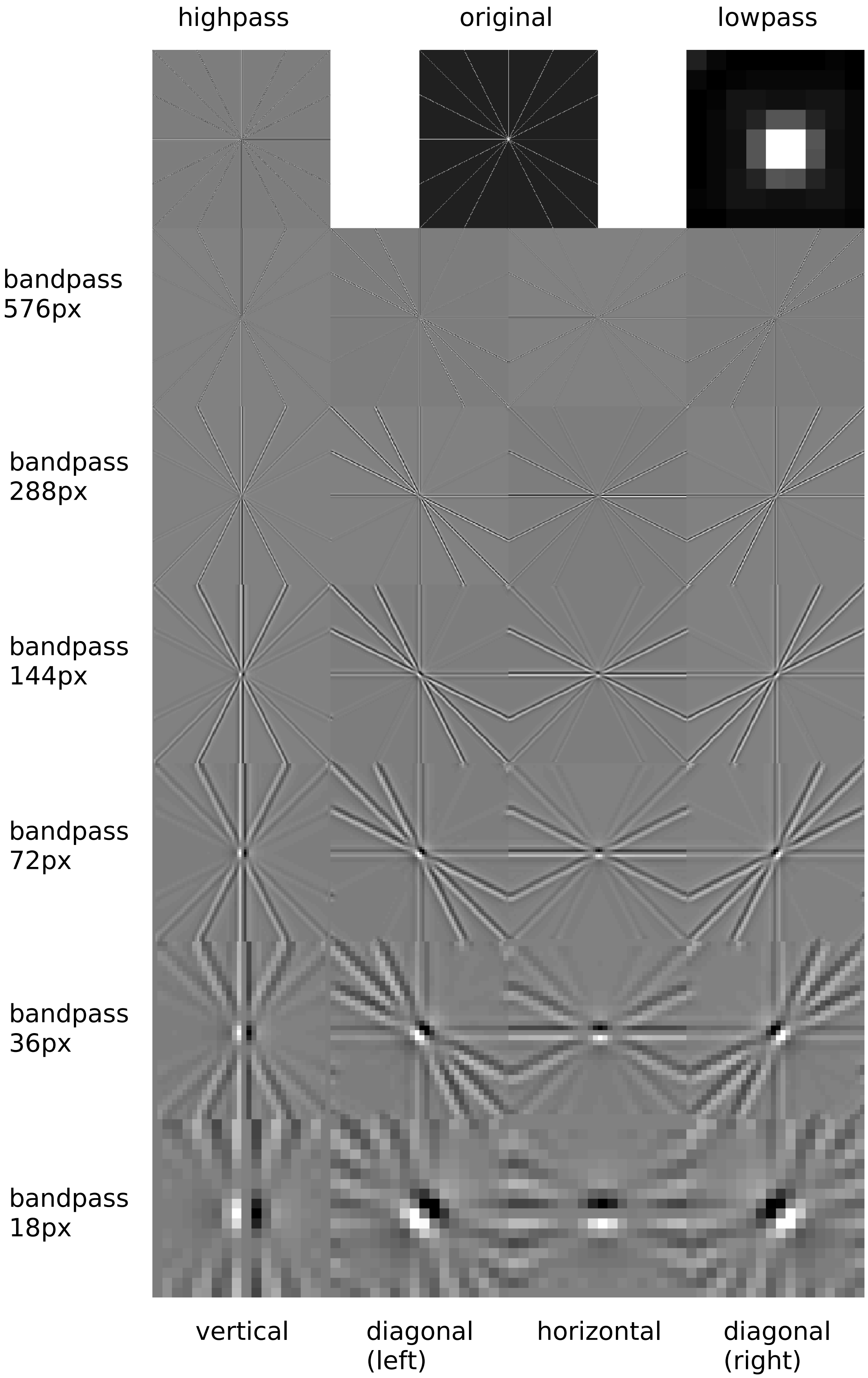

Illustration 74: The Simoncelli Pyramid, computed on an 8-directional star image

Illustration 74:

A high-contrast, synthetic image of a star has been processed through the Simoncelli Pyramid. The original image is located at the center of the top row, the resulting highpass image is at the left of the top row, the lowpass image is at the right. The bandpass images are ordered in rows according to their spatial size in pixels, then in columns according to the preferred direction of the corresponding filter. The original star can be recognized in all of the bandpass images, though only with some difficulty in the lowest bandpass (18 pixels). In the lowpass image, a single blob in the center of the image remains.

When looking at the individual columns, the behavior of the orientation-sensitive filters can be seen. In the preferred direction of any of the filters, e. g. the vertical direction in the first column, the components of each bandpass image are represented at their full strength, while at the orthogonal orientation no components remain. In the area between preferred and orthogonal orientation, the strength or energy of the components gradually declines.

A star image as used here has white lines on black background, causing highly complex spectral patterns around the edges. This causes frequency energy to be present at all over the spectral range of the image and in all orientations, and thus, the star is visible in all bandpass images.

Illustration 75:

In this case, the image of a zebra has been processed through the Simoncelli Pyramid. The general layout is the same as in Illustration 74. Again, elements of the zebra can be recognized in all of the bandpass images except for the lowest bandpass; the lowpass image is too coarse to be recognizable. Due to the mostly vertical orientation of the zebras stripes, they appear most clearly in the first column, where the vertical orientation is preferred. Also, the stripes almost disappear when looking at bandpass images smaller than 144 pixels, an indication that their spectral bandwidth is located closer to the higher end of the available spectrum.

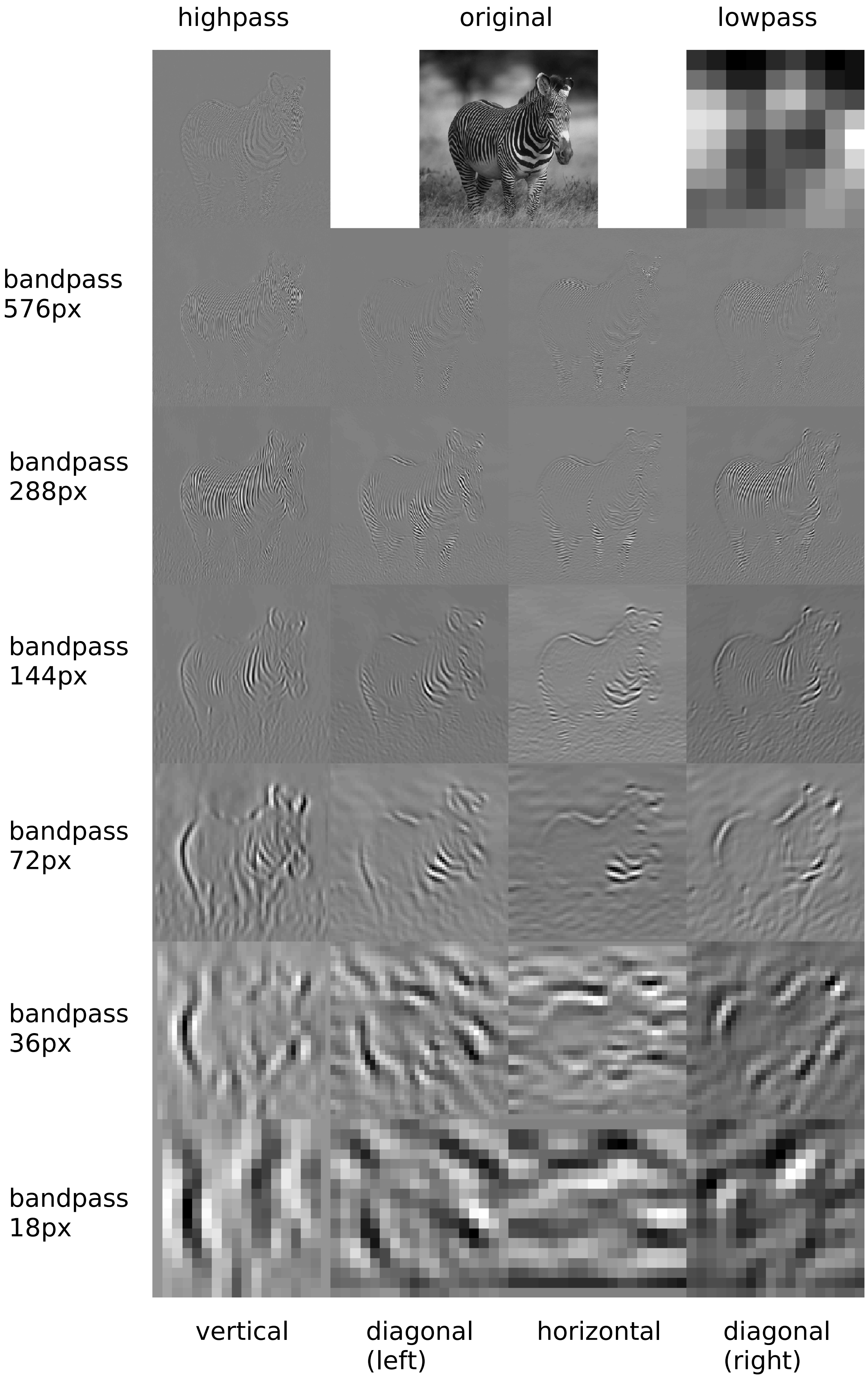

Illustration 75: The Simoncelli Pyramid computed on the image of a zebra

5.2. Dimensionality Reduction

The filtered images computed as described above are not well-suited for direct classification. During the computation of the Simoncelli-Pyramid, the dimensionality of the image is increased by a factor of 6.33, resulting in an enormous 2100897 dimensions:

Level	*Size* *(one edge in pixels)*	*Dimensions* *per image (pixels)*	*Number of* *images*	*Dimensions*
Highpass	576	331776	1	331776
Bandpass 1	576	331776	4	1327104
Bandpass 2	288	82944	4	331776
Bandpass 3	144	20736	4	82944
Bandpass 4	72	5184	4	20736
Bandpass 5	36	1296	4	5184
Bandpass 6	18	324	4	1296
Lowpass	9	81	1	81
			Sum	**2100897**

Table 8: Dimensionality of the Simoncelli Pyramid

The need for a low-dimensional representation of our data has been explained in chapter 3. In its current form, our pyramid offers 8 frequency bands and 4 orientation bands at up to $576x576 = 331776$ discrete positions. The fact that information is available at discrete locations across the image is the most important aspect of our current approach; still, frequency and orientation resolution are already fairly low, so further reduction does not seem practical here. Consequently, we will need to reduce the spatial resolution of our data to a computationally feasible level. The smallest sub-images in our Simoncelli-Pyramid aside from the low-pass image are of size 18x18 pixels; we therefore choose to resize all the higher frequency bands to the same size, once the Simoncelli-Pyramid has been computed. 18 by 18 discrete locations with 6 frequency bands and 4 orientations each result in $18x18x6x4 = 7776$ dimensions plus the low-pass and high-pass images, totaling 8181 dimensions. This is a significant reduction, even though by itself it will not be sufficient to achieve optimum classification performance. It is, however, enough for a first look at our data.

5.3. Spatial Distribution of Classification Performance

As a means to analyze how the information relevant for classification is spatially distributed, we classified on our entire APG dataset, employing the linear classifier to use every single dimension only by itself. An approach like this will in all likelihood lead to disappointing overall classification performance; however, achieving optimum performance is not yet our goal. The intended result of this procedure is to discover the relative differences in classification performance at the discrete spatial positions of our data. Through this, we hope to discover a localized concentration of information, allowing us to further reduce the dimensionality of our dataset without destroying information relevant to the classification process. The result (see Illustration 76) shows a general increase in classification performance in the higher frequency bands, in accordance with our findings in the previous chapters. The classification performance reaches about 63%, located both in the highest frequency band and the high-pass image. We also see that classification performance in the low-pass image as well as the lowest frequency band is very close to chance performance. When taking a closer look at the localized distribution of the classification performance, we can see that classification performance is better at the top and side of the highest-frequency subimages; in the center and towards the bottom the classification performance is lower. The results can be seen in Illustration 76.

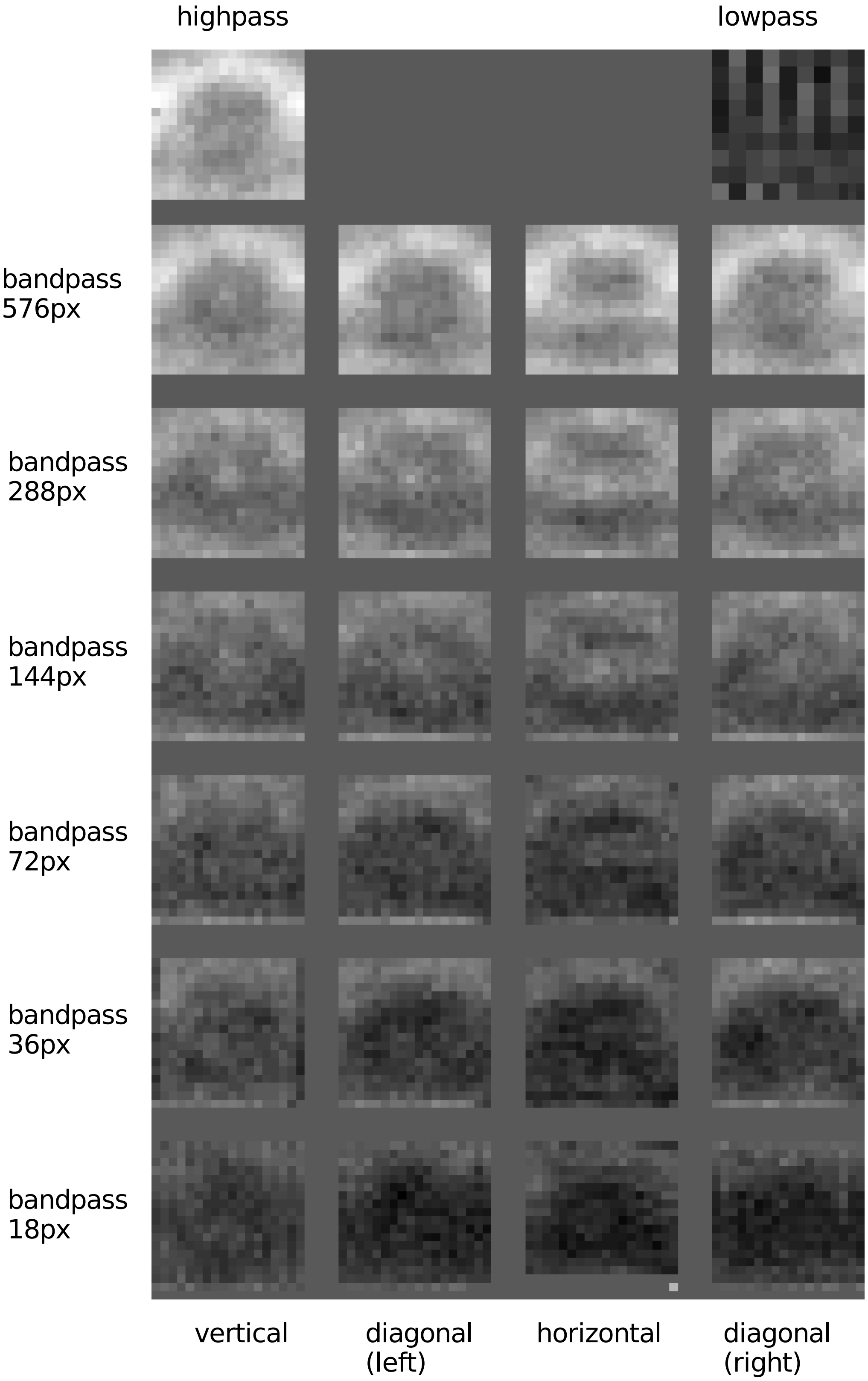

Illustration 76: Evaluation of the spatial distribution of classification-relevant information The bandpass images are ordered in rows, according to their spatial size in pixels. The orientations are ordered in columns, according to the preferred orientation of the corresponding filter. Brighter color represents better classification performance.

5.4. Information Where It Should Not Be

Illustration 77 shows the average distribution of classification performance among the highest-frequency bandpass images, together with the high-pass image. Clearly, the area of the lowest classification performance is localized in the center. The areas of highest classification performance are located at the top and at the sides of the image. When we compare this classification image with the mean animal image introduced in chapter 2.3, we find a striking similarity: The brownish-orange blob, originating from the fact that most animals are centered in the scene, is of almost identical size and position as the center-depression in the classification performance image.

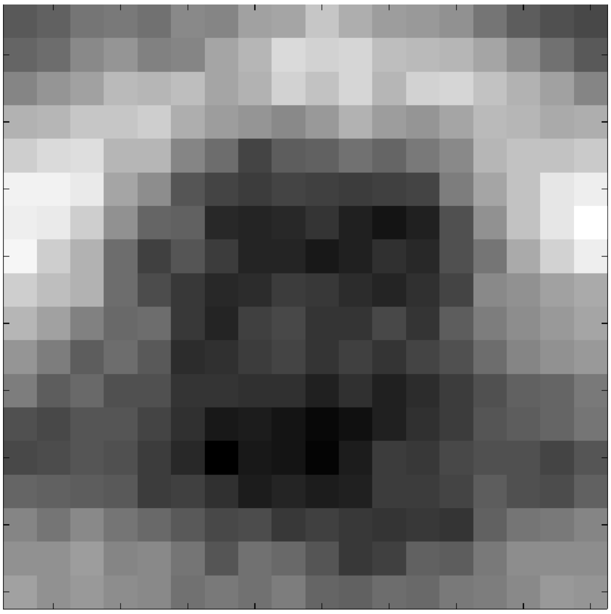

Illustration 77: Average classification performance in the highest frequency band (incl. the high-pass image)

Illustration 78: Average animal image (contrast maximized for easier comparison)

This points to the fact that in the algorithmic animal detection task, the most useful information is not actually the animal itself – it is the area around it! At first, this conclusion seems completely unreasonable: if the animal is not the location of the most important information, then it should be possible to classify images for their content (animal or no animal) even when the animal is hidden from the “view” of the algorithm! However, there may be a very reasonable explanation for this effect: all of the pictures in our image database are photographies taken by professional photographers, and have been taken with professional equipment. This professionalism in the image capturing process might have influenced the final appearance of the images, especially those with animals on them.

5.5. Preprocessing by Professional Photographers?

When a professional photographer takes pictures, this will usually result in several distinct types of images. For this context, we will differentiate only between scenic views with no particular object (landscapes, city scenes etc.) and views with animals as the center object of the scene. In a typical landscape view, the image will be taken using a wide-angle lens, with the focus set so that most of the view is well in focus (see Illustration 79, and the manually estimated distribution of focus in Illustration 80).

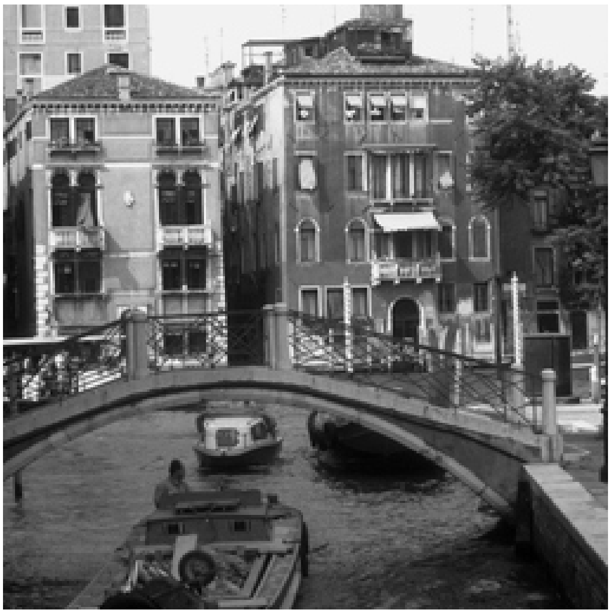

Illustration 79: A city scene, as found in the APG database

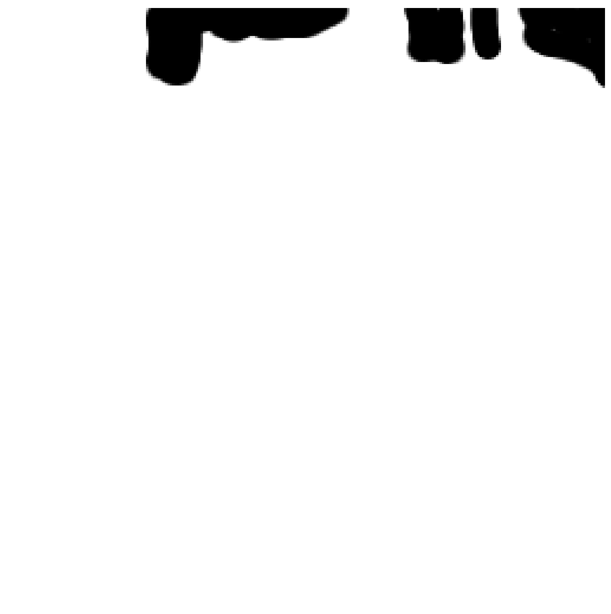

Illustration 80: Approximate distribution of focus: areas in focus are shown in white.

When taking pictures of animals in their natural surrounding, there are a number of factors to be considered. First, most animals have a certain radius of proximity, which, when intruded, will cause them to flee. The photographer will therefore need to stay far enough away from the animal, and will therefore have to use a rather long lens. Second, many animals will usually move around a lot, making it difficult to take a sharp picture without motion blur. This requires the photographer to set his camera for a short exposure duration, which, especially with a long lens, will usually require a rather large aperture. A large aperture is also considered desirable because it will help to produce a rather slim focal plane, segmenting the (focal) target object from the background. This segmentation is commonly perceived as aesthetically pleasing; due to the lack of high spatial frequency the background will be perceived as “quiet” and therefore not distracting.

5.6. Aperture Size and Depth of Field

A point within the scene depicted by a photography is considered “sharp”, in-focus, when its circle of confusion does not exceed a certain diameter. The circle of confusion is the disk (or blob) resulting from the projection of a (theoretical) point of light through the lens (or optical array) onto the film / sensor plane. The acceptable diameter of the circle of confusion depends on a number of factors such as the post-capture magnification (desired print size) of the captured image data and, important but variable, the optical acuity of the observer. When using 36mm film, commonly used in the most popular consumer cameras, the acceptable diameter of the circle of confusion is usually specified to be no more than 0.03mm, based on a desired printing size of 30cm. The connection between the acceptable circle of confusion and depth of field can be illustrated by following the rays of light from the focal plane to the sensor plane: rays originating from a single point on the focal plane will converge in a single point at the sensor plane (assuming a perfect optical array), their circle of confusion will be zero. Extending from this distance of perfect focus, objects that are not too far away from the focal plane will appear focused as well: While they are not as sharp as objects on the focal plane, their circle of confusion is still below the assumed visibility threshold, and so one projected point will still be perceived as one point. However, objects far away from the focal plane appear blurry, as their circle of confusion is of much larger diameter. The maximum distance from the focal plane within which an object still appears focused depends not only on the size of the acceptable circle of confusion, but also on the maximum angle at which a ray of light will still be projected by the optical array onto the sensor plane. This maximum angle depends on the distance from the optical array and on the diameter of the optical array (the aperture). At the time of capture of any given scene, the distance of the target object from the camera can be considered to be fixed, yet the aperture can be modified by the photographer, varying the depth of field. As mentioned before, when taking pictures of animals in nature, usually large apertures will be used, leading to a slim depth of field. The connection between aperture size and depth of field can be seen in Illustrations 81 and 82.

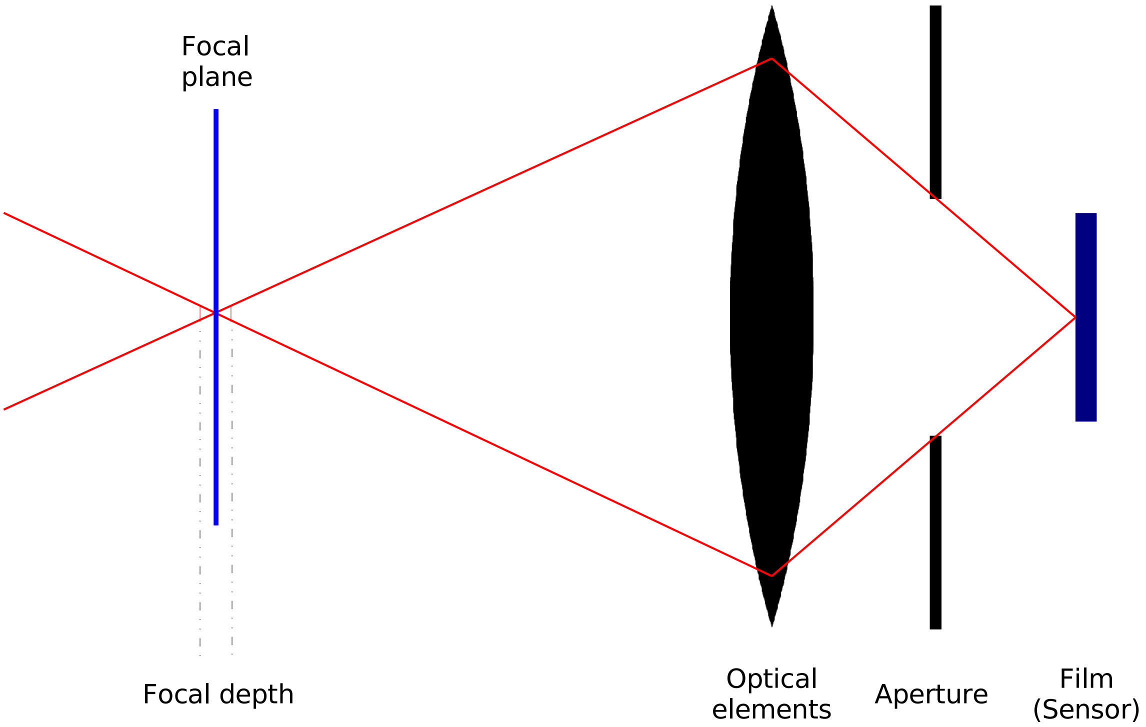

Illustration 81: Shallow focal depth with large aperture setting

With a wide open aperture, the area of good focus is rather slim (shown by the two red markers near the focal point and the dotted black lines near the label “Focal depth”). Symbolic light rays are shown in red.

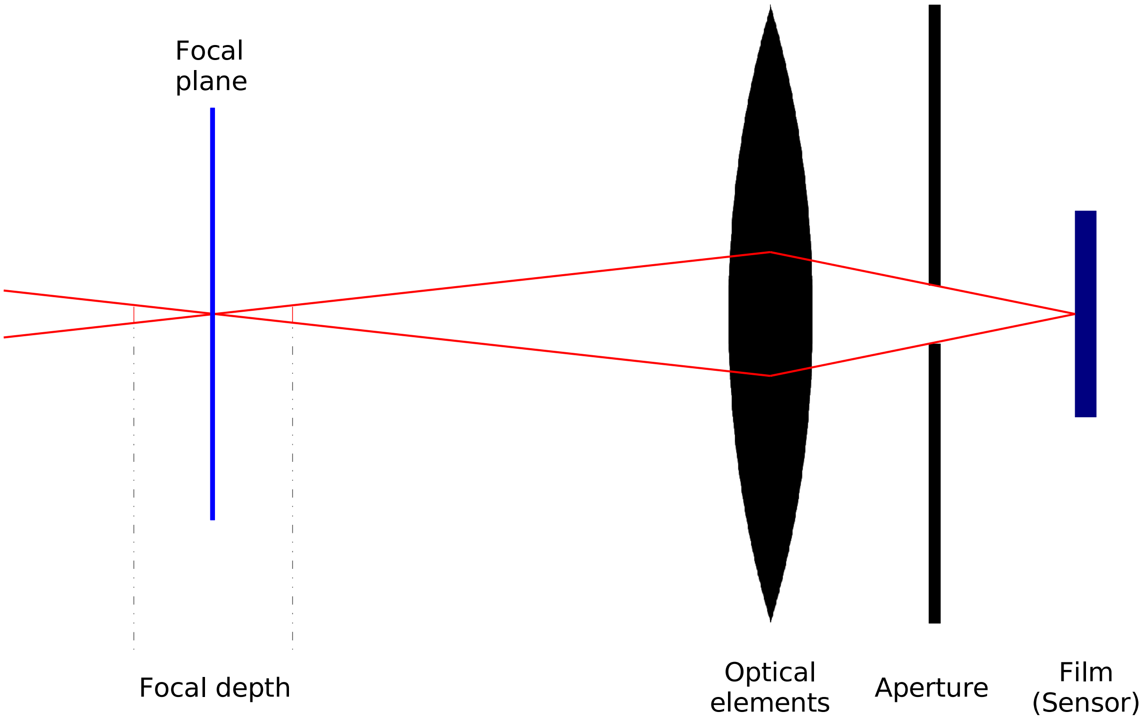

Illustration 82: Extended focal depth with small aperture setting

With a small aperture, the optical diameter of the lens is reduced, and the depth of focus is extended

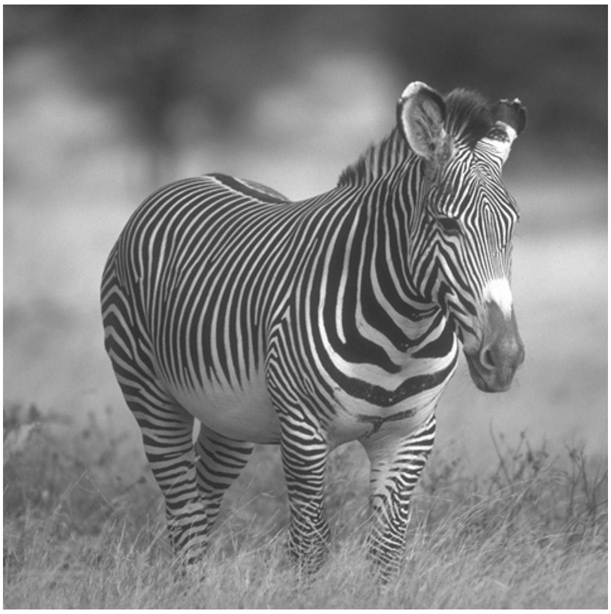

Illustration 83: A zebra, as found in the APG database. The animal is in focus, the surroundings are not

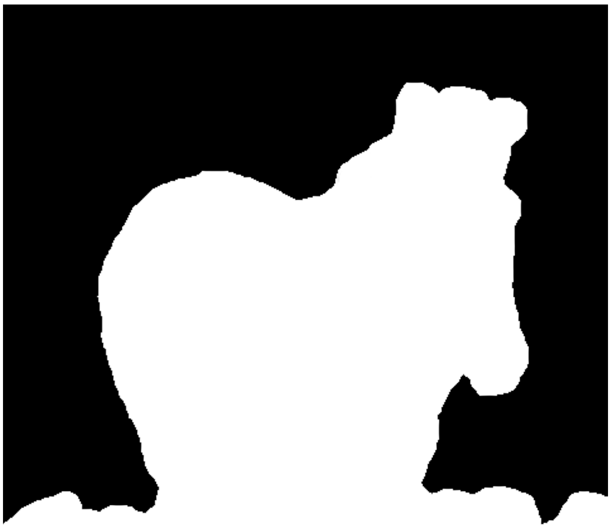

Illustration 84: Approximate distribution of focus: areas in focus are shown in white

The effect of a slim focal plane on a typical animal image can be seen in Illustration 83. The zebra is perfectly in focus, resulting in a clear and sharp representation, with a significant amount of high-frequency energy (see also Illustration 75). The same is valid for the grass around the place where the zebra is standing – the distance from the grass to the camera is about the same as the distance from the zebra to the camera, so the grass is within the area of good focus. The areas of the picture that are to the sides and above the animal, however, are located beyond the focal area, and appear blurry, out of focus. This essentially represents a low-pass filtered image area, with next to no high-frequency components. The manually estimated distribution of focus amongst the image can be seen in Illustration 84; clearly, the shape of the animal is recognizable.

5.7. Localized Computer Classification

We have postulated that the professionalism of photographers effected a local concentration of high-frequency energy around the center of typical animal images. While this may be very convincing on selected images, it is not at all clear whether this effect was intense enough to unintentionally aid computer classification. If it was, it remains uncertain if it affects a number of images within our APG database sufficient to affect average classification accuracy. In order to test this, we need to further reduce the dimensionality of our data space to allow for a higher general classification performance. We will therefore reduce the spatial resolution of our Simoncelli-Pyramid; in this reduction we will account for the shape of the center region found in the above examination. We will design a center region of circular shape, with an equal distribution of area (and thus, dimensionality) between the center region and the surrounding outer region. The subimages of our Simoncelli-Pyramid are of size 18x18, resulting in 324 pixels. The center circle shall therefore cover 162 pixels. Our arrangement of pixels within both the center region and the outer region can be seen in Illustration 85.

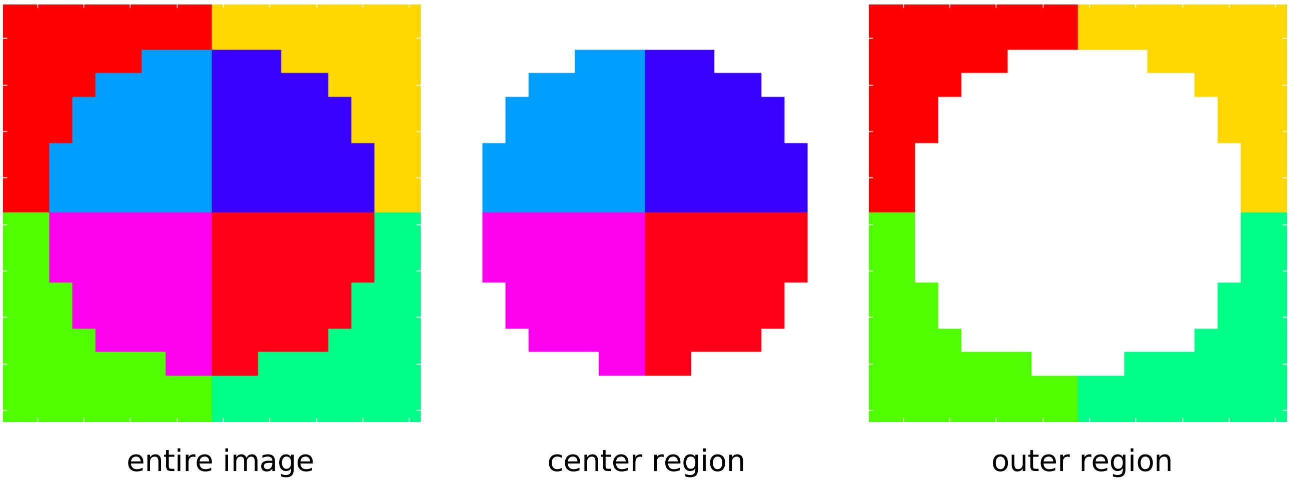

Illustration 85: Bin mask design: center region vs. outer region

The subimages have been divided into quadrants, measuring 81 pixels each. As these 81 pixels cannot be divided evenly into two sectors, 41 pixels have been assigned to the center region from each of the top quadrants, and 40 from each of the bottom quadrants. The result is a center region that approximates a circular shape in a way optimally suited for our purpose. Through this procedure, the dimensionality of each subimage is reduced from 324 to 8. While we went to great lengths to produce data that represents information at discrete locations, we have now reduced these discrete locations to the absolute minimum required by our task. Further, based on our previous results, we do no longer include the lowest frequency bandpass images or the lowpass image. In its new form, the Simoncelli-Pyramid now measures 168 dimensions, or 84 dimensions for either the center region or the outer region. With this kind of dimensionality reduction, it should be possible to perform a highly efficient classification. Also, we are able to split the image evenly into center region and outer region. The number of dimensions involved, especially when classifying using the entire image area, is just within the feasibility limits of our incremental elimination procedure (see chapter 3.4.2.4), which we will use to determine the best possible classification performance. If our hypothesis holds, the classification performance based on the outer regions alone will not be significantly below the performance based on the center regions alone, while both may be slightly below the performance based on the entire images.

5.8. Results of Iterative Bin Elimination

On the newly configured dataset developed in the previous chapter, we employed our iterative bin elimination procedure, with 200 cross-validations (see chapter 3.4.2.4). The elimination was computed separately first for the entire images, then only the center region, then only the outer region. The maximum classification performance reaches 77,9%, achieved on the entire images. The performance on the circular inner region of the images reaches 74%, while the maximum on the outer region reaches 73,7%. All three of the elimination procedures exhibit the plateau typical for this kind of evaluation.

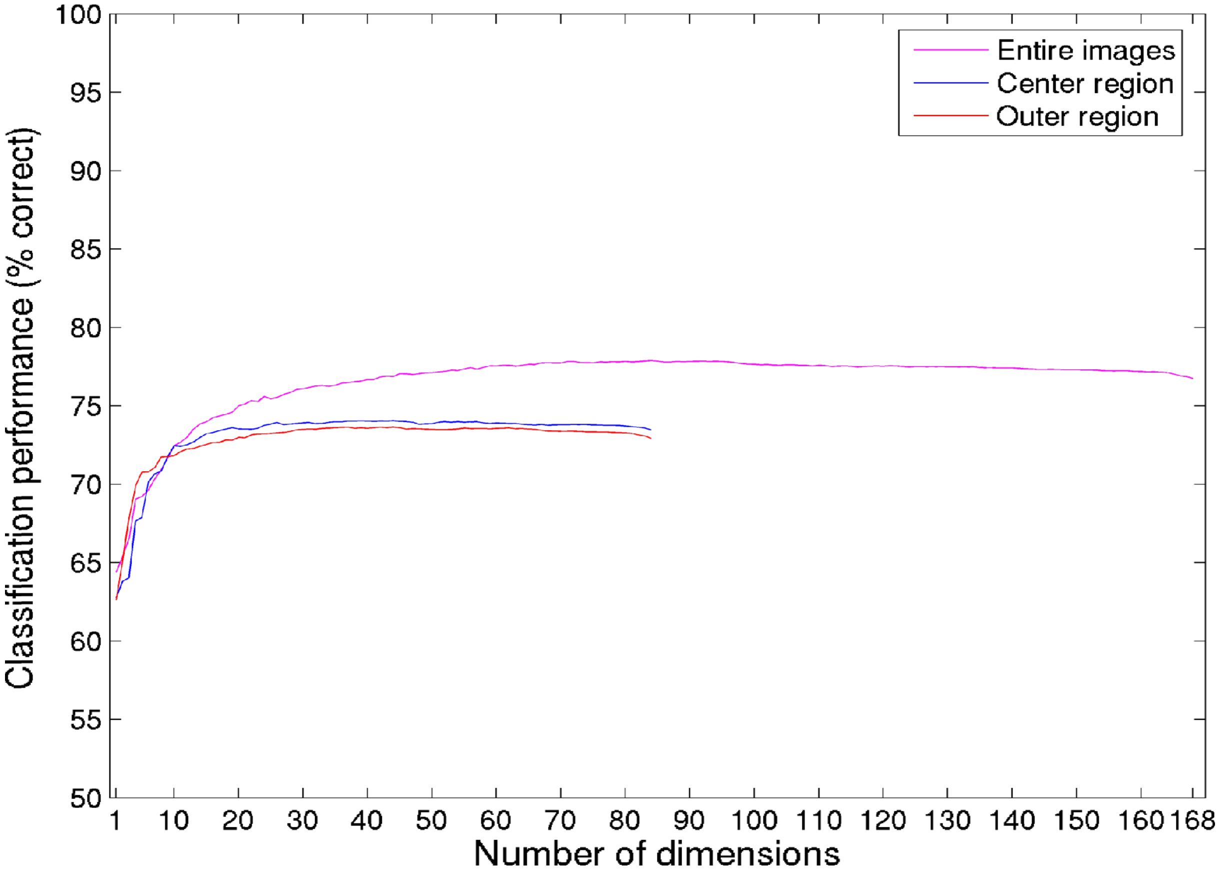

Illustration 86: Result of iterative bin elimination

The difference in classification performance between the inner and outer region is negligible; 0.3% represent a mere 33 images, out of 10864. It is possible to detect the presence of an animal in an image without actually seeing the animal! These results strongly support the hypothesis that the photographers camera setting has an influence on the frequency distribution within the images and that this influence does affect classifiability. As this influence is not truly related to the content of the image, but only to the photographers intended content, we must call this influence an artifact (or bias) of the image capturing process.

5.9. Experiment 4: Test for localized differences in classification performance

Our computer algorithm found strong evidence for an artifact in the image capturing process. It is therefore theoretically possible that humans tasked with the classification of image content incorporate the same artifact into their decision process. Humans are known to be able to detect a multitude of object classes, many of which would be captured with the same camera settings. However, rapid visual classification is possible when two classes of discrete objects (e. g. vehicles and animals) are to be compared [VanRullen Thorpe 2001], a case in which the photographers settings will be the same for both classes. We therefore strongly doubt that such an artifact could influence human visual classification in a significant way. In order to find out whether humans incorporate such artifact-based information in their classification decision, we need to design an experiment to test human classification performance on the same image regions as our localized computer classifier.

5.9.1. Method

To test human classification performance on image regions analog to those specified in chapter 5.7, we selected a random subset of 800 images from our APG database. Following the same general principle as in the computer classification, we created a circular mask in such way that the resulting inner region consists of equally as many pixels as the outer region. The subimages of our Simoncelli Pyramid are of size 18x18 pixels, and so the circular mask applied was rather coarse. With our human subjects, we used the full image size (256 pixels). Our circular mask hat a diameter of 204 pixels, totaling 32904 pixels or 50,2% of the image surface. The outer region consequently totaled 32632 pixels or 49,8% of the image surface. A perfect distribution of pixels would have required us to slightly alter the shape of the mask, as within the given pixel grid no perfect disk can account for precisely 50,0% of the image surface. We refrained from altering the shape of the mask; instead, to reduce the possible distraction that the sharp border between cut-out image and background might have caused, the sharp edge was transformed into a Gaussian transition, with a total width of 32 pixels between minimum and maximum. The resulting mask (and its inverted form) was then used to cut images into “inner region only” and “outer region only”, blending the eliminated part into neutral gray. The mask and its opacity profile can be seen in Illustrations 87 and 88, the resulting cut-out images can be seen in Illustration 89. As a result of this procedure, most (but not quite all) of the animals were invisible on the “outer region” images. On some images, a few pixels of the animal extended into the outer region and were not eliminated. While these might account for a performance slightly better than random, we would still expect humans to achieve a performance level far below that of the entire images, or even the “inner region” images. To test this, we employed a version of the 2AFC paradigm (see Illustration 90), showing all 800 images in 400 pairs. Every image pair was shown in each of the 3 variations (see Illustration 89), with both images being of the same variation. The resulting 1200 trials for one complete run were split into 3 sets of 400 to ensure that each subject would not spend much longer than 30 minutes in this experiment. The experimental setup and equipment used was the same as described in chapter 4.1, and we used the same timing constraints that we also employed in Experiment 1 and 3b (200ms gap time, 30ms presentation time and 1000ms maximum response time).

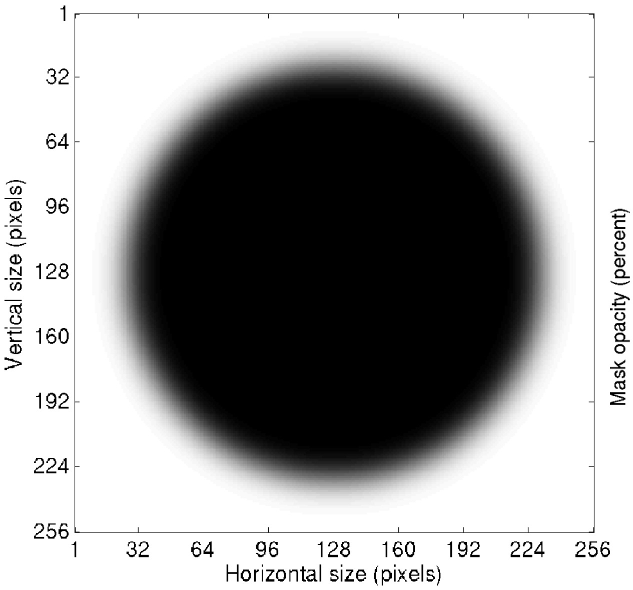

Illustration 87: Mask used to blend between inner and outer region

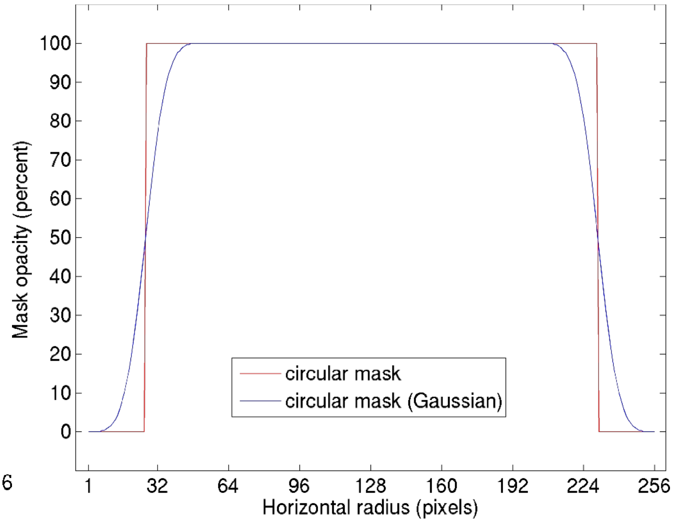

Illustration 88: Opacity/Blending profile of the mask used to blend between inner and outer region (horizontal cut)

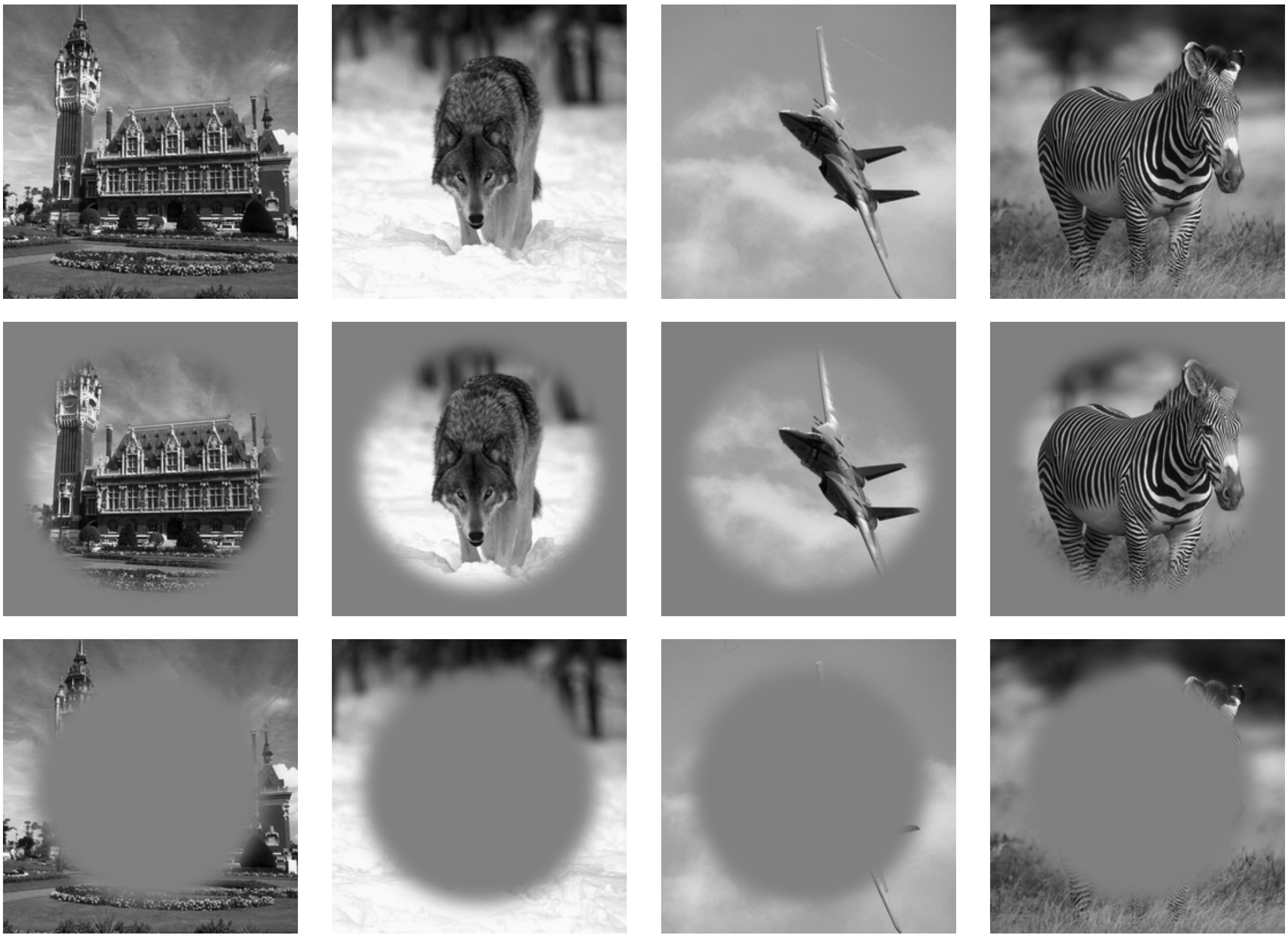

Illustration 89: Sample images from Experiment 4.

Top row: entire images. Second row: inner region only. Third row: outer region only.

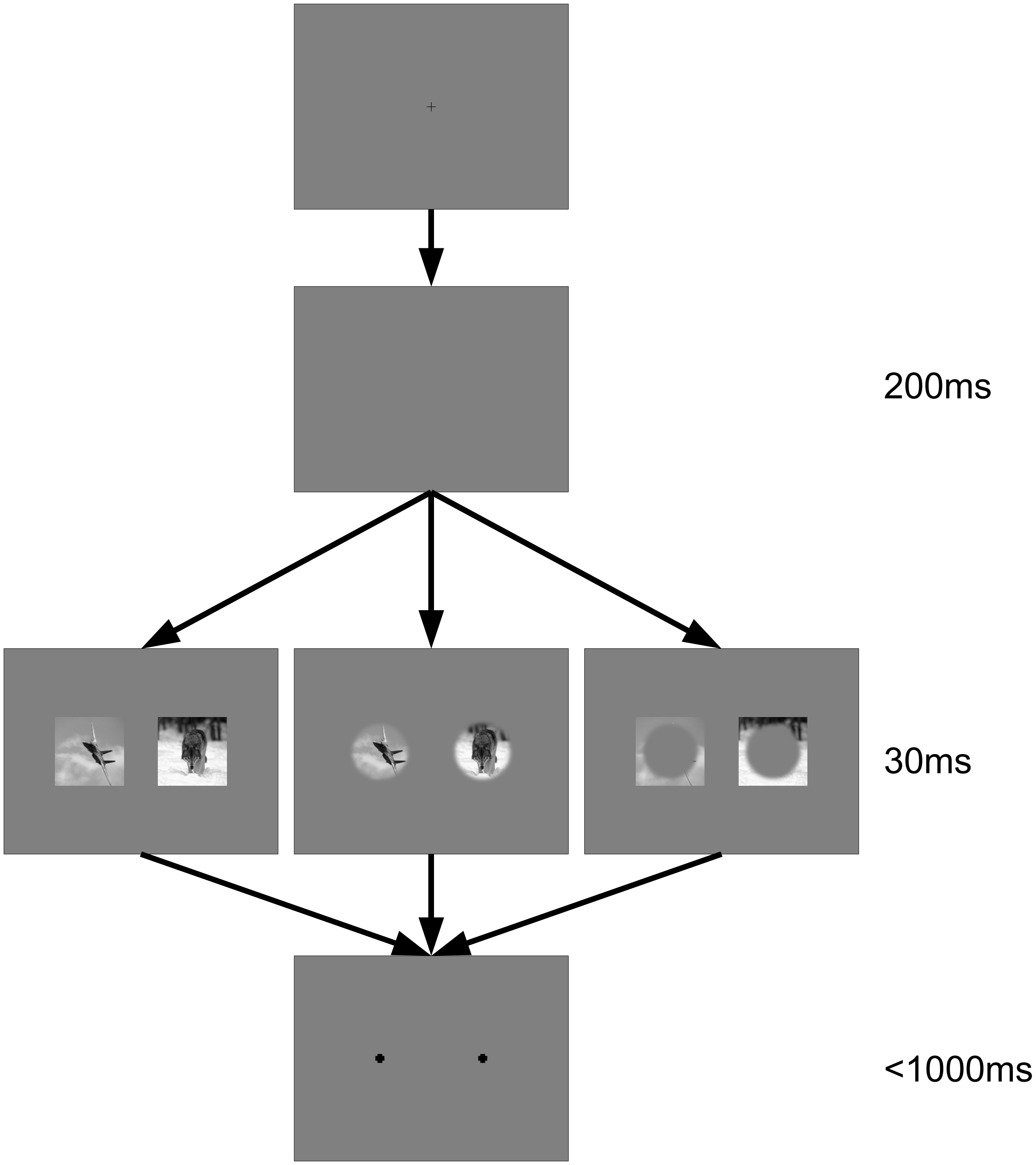

Illustration 90: Experiment 4: 2AFC-paradigm for complete images, circular inner regions and outer regions

5.9.2. Results

12 subjects participated in this experiment, 4 on each of the 3 sets of images. All subjects were students of the Justus-Liebig-University Giessen and were payed for their participation. Subjects were between 19 and 31 years of age and had normal or corrected to normal vision. All subjects were able to perform eye movements adequate to achieve calibration accuracy rated as “good” (the best rating) by the EyeLink II system. All subjects were naive, and even though some participated in more than one experiment, they were not informed about the purpose of the individual experiments until after they had completed all the trials they were scheduled for.

394[1] trials were scheduled for each subject, resulting in 4728 recorded trials from the 12 subjects. Prior to evaluation, the individual results were filtered with the following criteria:

Goodness of fixation:

Fixation between trial start and stimulus onset was not allowed to diverge from the screen center by more than 70 pixels, no saccades were allowed prior to stimulus onset; this made certain that the subjects did not favor one of the locations where the images would be shown and that no saccades were made during the critical phases of the experiment, as this might have hampered visual perception. During this filtering step, 150 Trials were eliminated, leaving 4578 for further processing.

Goodness of saccade direction / destination

Decision saccades were required to end in one of the two smallest possible squares covering the area that was occupied by the target and distractor images. This ensured that the saccades actually were directed into the general area of either of the shown images and eliminated random saccades, e. g. from lack of concentration. Only 9 trials needed to be eliminated to fulfill the requirements, leaving 4569 trials for further processing.

Goodness of response time

Subjects needed to make their decision no later than 700ms after stimulus onset in order to limit conscious thought in the decision process. Responses also had to be slower than 80ms to eliminate random and too early eye movements (reaction times faster than 80ms can safely be assumed not to be based on stimulus content). 278 trials were removed because they were too fast, none were too slow, leaving a total of 4291 valid trials for evaluation.

[1] Due to technical problems, 6 of the original 400 trials had to be discarded.

Image area shown	***Hit ratio***	***Response time***
Entire images	84,4%	294,6ms
Center region	78,4%	294,9ms
Outer region	53,7%	324,8ms

Table 9: General results of the test for localized differences in classification performance

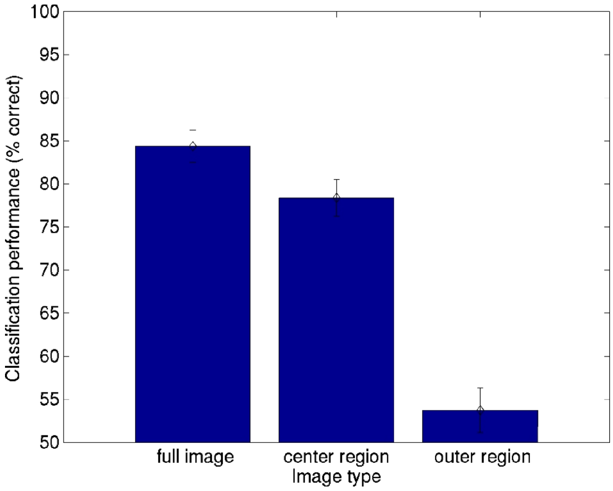

Illustration 91: Experiment 4: Classification accuracy

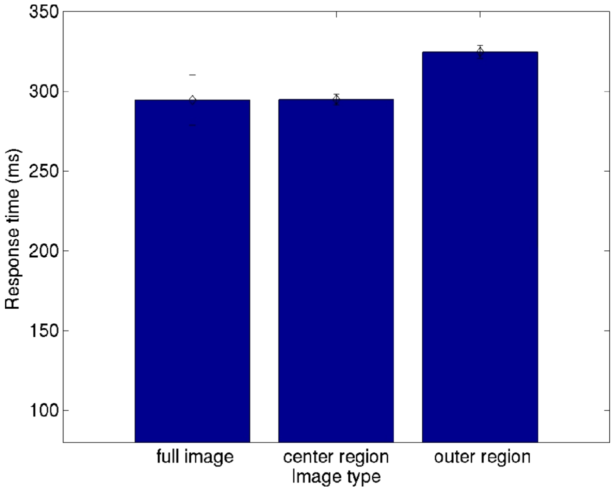

Illustration 92: Experiment 4: Response times

Classification performance on the subset of entire images averaged 84,4%, with a mean response time of 294,6ms. This is in the general range of results to be expected in this kind of experiment (see also chapter 4). When showing only the center region of the images, classification performance averaged 78,4%, a rather modest decline. The mean response time remained almost unchanged (294,9ms). Classification accuracy drops to chance performance (53,7%) when showing only the outer regions of the images, with the mean response time increasing to 324,8ms. A statistical analysis on hit ratio data shows a highly significant effect overall (repeated measures ANOVA, $F(2,22)=183{,}713$, $p<0.001$). Individual t-tests between the 3 conditions show a significant difference between each of the 3 conditions:

Condition 1	***Condition 2***	***p value***
Entire images	Center regions	0.03
Entire images	Outer regions	<0.01
Center region	Outer regions	<0.01

Table 10: T-test results on hit ratio data (Bonferroni corrected)

With the mean response times, the results are a bit different. The ANOVA shows a highly significant effect (repeated measures ANOVA, *F*(2,22)=17.806, *p*=0.001). Due to the very similar means of the response times with entire image regions and with center regions, the degrees of freedom have been corrected as proposed by Huynh and Feldt; we report the original degrees of freedom with the corrected p-value. Individual t-tests show that while the difference between the outer region and each of the two other image areas is highly significant, the difference between entire images and center regions is not.

Condition 1	***Condition 2***	***p-value***
Entire images	Center regions	2.487
Entire images	Outer regions	<0.01
Center regions	Outer regions	<0.01

Table 11: T-test results on response time data (Bonferroni corrected)

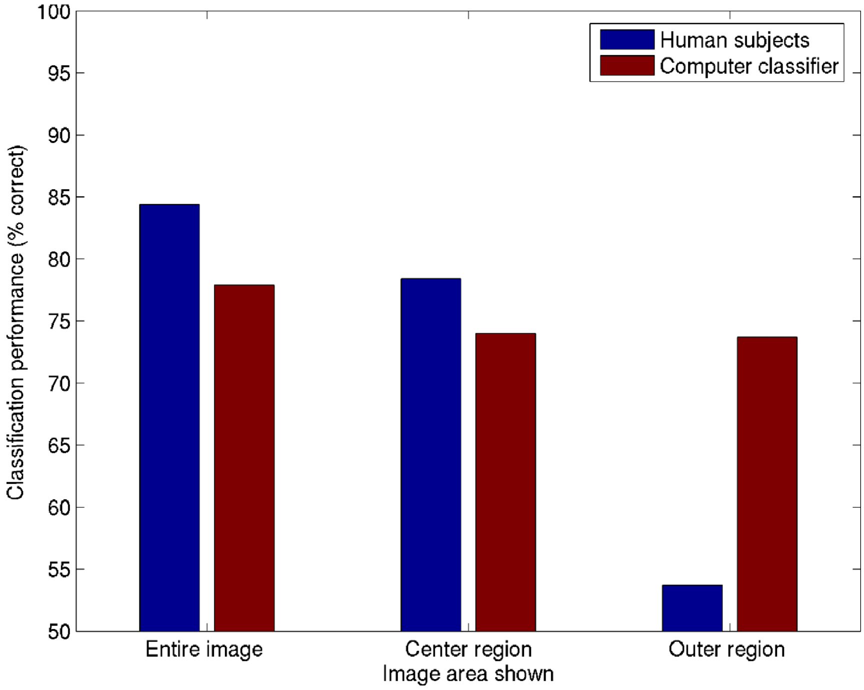

Illustration 93: Comparision of human and machine on image regions

We conclude that while our computer classifier was affected by an artifact in the image capturing process, our human subjects apparently did not incorporate this information in their decision process (see Illustration 93). This may be because of the temporal constraints of the experimental paradigm.

6. SUMMARY AND CONCLUSION

In this thesis, several aspects of human visual classification have been analyzed. We attempted to model the critical properties of the mechanism humans employ in tasks of rapid visual classification. More specifically, we tried to identify the kind of information that might be used by the human visual system to achieve its superb performance. The global amplitude spectrum has repeatedly been suggested as a suitable way of information representation, and so we based our first series of tests and experiments on it. We found that indeed the information contained in the global amplitude spectrum, even at a rather coarse scale, is quite adequate for successful computer classification. We were able to classify almost 75% of our images correctly, despite of the very high degree of variability within our database. With an image collection chosen by more rigid criteria (as the one provided by A. Torralba), higher classification performance can be achieved. We believe, however, that a classification algorithm of true merit will have to stand its ground even (or especially) against difficult image databases. The computational cost of our Fourier Fingerprint classification algorithm is rather low and would allow for an efficient implementation, e. g. to search image databases.

In our black box approach, we identified two very distinct properties of the Fourier Fingerprint classifier that show a very high degree of similarity with the properties of the human visual system:

First, the sensitivity to image rotation, represented by the classification accuracy on rotated images, follows the same profile that we measured in our human subjects. This supports the idea that the global amplitude spectrum is used by the human visual system, as this would explain the rotation/accuracy pattern. This would, however, be a typical property of *any* mechanism utilizing the global amplitude spectrum – it does not tell us in *which way* the global amplitude spectrum would be used.

Second, the order of the images in terms of classification difficulty as computed by our algorithm is very similar to the performance that human subjects exhibit on the same images. If the first experiment supports the idea that the global amplitude spectrum might be used by the human visual system, then the second experiment supports the idea that the individual aspects of the global amplitude spectrum are weighted similarly by both our computer algorithm and the human visual system.

Together, these two findings make a very strong case for the global amplitude spectrum hypothesis. The fact remains that individual humans can identify images at better than 90% correct, while our computer classifiers max out at almost 15% less. When ignoring the findings from our third experiment, however, this difference might be attributed to a deficiency in our preprocessing, some hidden secondary mechanism in the human visual system, or even a limited top-down influence. The performance advantage of our human subjects alone does not diminish the credibility of our findings. Yet still, all these aspects are led ad absurdum by the findings from our third experiment: when the amplitude spectrum is “wiped out”, our computer algorithm is completely blinded, and would be limited to chance performance. Humans, however, can still classify images almost as good as before. From the view of our black box approach, this means that we have found a condition under which the input-output mapping of any classifier based solely on the global amplitude spectrum will differ from the human visual system. The match between the input-output relations of the Fourier Fingerprint classifier and the human visual system is therefore only partial, and not sufficient to claim equivalent functionality. Through this, the question whether the global amplitude spectrum plays a critical role in human visual classification has been answered: it does not.

While our Fourier Fingerprint classifier may not be a perfect solution, it does have some merit. The results presented in experiments 1 and 2 show that representing an image by means of frequency and orientation can be useful indeed. We went along with this knowledge and evaluated a new representation of our image data, also based on frequency and orientation. In our previous approach, the use of the global amplitude spectrum caused us to abandon any and all locality of information. Our Simoncelli-Pyramid approach retained locality in order to find out if this newly added piece of information changed the behavior of our classification algorithm. Surprisingly, we discovered a concentration of classification-relevant information in image regions where we would not have expected it: *around* the animal, instead of *inside* the animal. While we were able to use the new, localized representation of our images to achieve even slightly better classification performance than before (our absolute maximum being almost 78%), we also discovered a strong artifact of the image capturing process: low depth of field hints at the presence of an animal in our scenes. This is truly a stunning discovery: many researchers have used the same original image library in their research, yet until today, no report of such an artifact has come to the authors attention.

Concluding, we have successfully eliminated the hypothesis that the global amplitude spectrum might play a critical role in human visual classification. Also, we have identified a strong potential artifact in the Corel Stock Photo Library, one of the most frequently used image databases in vision research. To discover the inner workings of the human visual system, especially the mechanisms used in rapid visual classification, further research is required; this research should emphasize localized ways of information representation.

7. ADDENDUM: CLASSIFICATION ALGORITHMS

In this context, “classification” is the process of assigning a given data sample of a priori unknown group membership to one of usually at least two groups by means of evaluation of the individual attribute values associated with the sample. All classifications in this context are binary (there exist exactly two groups, exactly one of which a sample belongs to) due to the nature of the analyzed data. Classification is generally done in two steps:

First, a number of samples with known group memberships (“labeled data”) is used to learn two classification equations, one for each group. These equations actually form the “classifier”. Second, a number of samples is used to test the learned classifier. The learned equations are evaluated on each of the testing samples, and the sample is assigned to the group corresponding to the equation that resulted in the highest score. Classification performance is measured in terms of classification accuracy: how many of the samples are being assigned correctly, such that the assigned group is the same as the group the sample actually belongs to. It is important to test a classifier on samples that were not included in the learning set, as it is possible that a (falsely!) learned classifier predicts the learning (or training) dataset very well, even perfectly, but performs very poorly on new data. This phenomenon is called “overlearning” or “overfitting”. The possibility of erroneously mistaking an overlearned classifier for a well-working one can be eliminated by keeping the samples used for training strictly separate from the ones used for testing. This way the learned classifier can not be overfitted to the testing data. In most situations, one will want to keep the training set as large as possible, necessarily reducing the size of the testing dataset (usually, as in this context, the overall number of samples available is quite finite and needs to be spent wisely). This can make for rather small testing samples, which again would produce rather unreliable measurements of classification accuracy. Commonly, a procedure called “cross-validation” is used to circumvent this limitation: for a k -fold cross-validation, the entire available data is arranged in k equally sized subsets. Then, $k-1$ of these are used for training, and the remaining 1 is used for testing. The particular subset used for testing is cycled through exactly k times, so that every subset has been tested once. Classification accuracy is then reported as the mean of all of the classification accuracies on the k testing subsets. Result reliability increases with larger k , but so does the computational complexity due to the number of training / testing cycles. The demand for more reliable results and computational feasibility therefore need to be balanced out, limiting the amount of cross-validations used to values that are usually far below the theoretically possible maximum (which would be N , the number of samples available).

7.1. Linear Classifier

The majority of classification evaluations performed in this context make use of the linear classifier, similar to a discriminant analysis. For a given sample S , the score of the classification function C_j for one of the two possible groups $j=\{1|2\}$ is found by multiplying the values in each dimension of the data space by their associated, learned coefficients c_{jn} :

$$C_j=c_{j0}+\sum_{n=1}^{N} c_{jn}S_n \quad ,$$

with $n=1:N$ iterating the dimensions of the data space. The coefficients c_{jn} are learned from the means and the pooled covariance matrix of the data dimensions.

This classifier can be thought of as a N-dimensional hyperplane separating the dataset into 2 parts. The advantage is a comparably low computational cost and good performance on linearly separable datasets; the disadvantage, however, is that there can be only one flat hyperplane, which can cause this classifier to perform poorly on datasets with several “islands” of data belonging to the same group. A detailed introduction to (not only) linear classification can be found in [Tabachnick Fidell]. For this context, the implementation given in Matlab's “classify” function was used, with auxiliary extensions designed to provide for cross-validation support.

7.2. Support Vector Machine Classifier

The second classification mechanism used in this context is the Support Vector Machine (SVM). Based on work by Vladimir Vapnik and colleagues, the general is to find a hyperplane that separates the data points in our dataset in an optimal fashion. The original algorithm as proposed by Vapnik was a linear classifier, which was modified later using the so-called "kernel trick" [Boser Guyon Vapnik 1992], introducing a transformation of the feature space. This transformation allows the classifier to yield a non-linear result in the original feature space even though it is a hyperplane in the transformed feature space. This allows also for several unconnected planes in the original feature space, which can be a huge advantage over the linear classifier as introduced in the previous chapter. In this work, the only kernel used is the Gaussian radial basis function,

RBF: $k(x,y)=e^{\frac{-\|x-y\|^2}{2\sigma^2}}$.

For the practical requirements of this work, optimum classification performance will be achieved through the optimization of the two relevant parameters of the RBF kernel, the penalty parameter C and a measure of the size of the Gaussian, γ . For practical reasons, the search for an optimal pair of parameters can be reduced to assigning a fixed value to one parameter and then performing a search on the other parameter.

In this context, the computations were done in Matlab, using the OSU SVM[1] toolbox. Our procedure to find an optimal set of parameters was implemented as a two-stage interval search testing C parameters ranging from e^{-20} to e^{20} . The first stage consisted of 200 logarithmically spaced steps s distributed over the entire search interval. This stage will result in a maximum found somewhere in the searched interval at position $C_{max}\alpha$. As the second stage, the interval $\left[C_{max}\alpha-s \;\; C_{max}\alpha+s\right]$ is searched in another 200 steps, resulting in a reasonably high search accuracy. The whole process is then wrapped in a cross-validation, with the result reported being the maximum of the means of the cross-validations of the second stage. For demonstration purposes the following data have been computed with 10 cross-validations.

[1] OSU SVM for MATLAB, by Junshui MA and Yi Zhao. http://svm.sourceforge.net

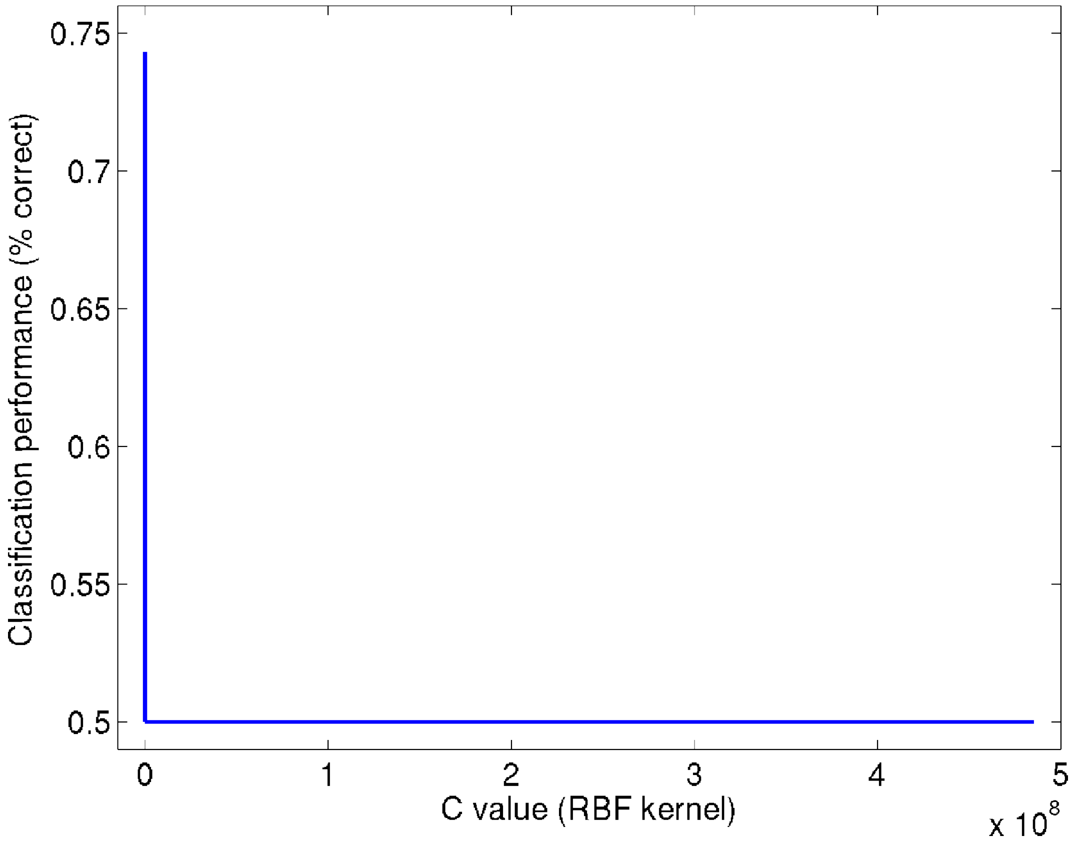

Illustration 94: Exemplary result of a C-value search, first stage. A significant maximum close to zero can be seen

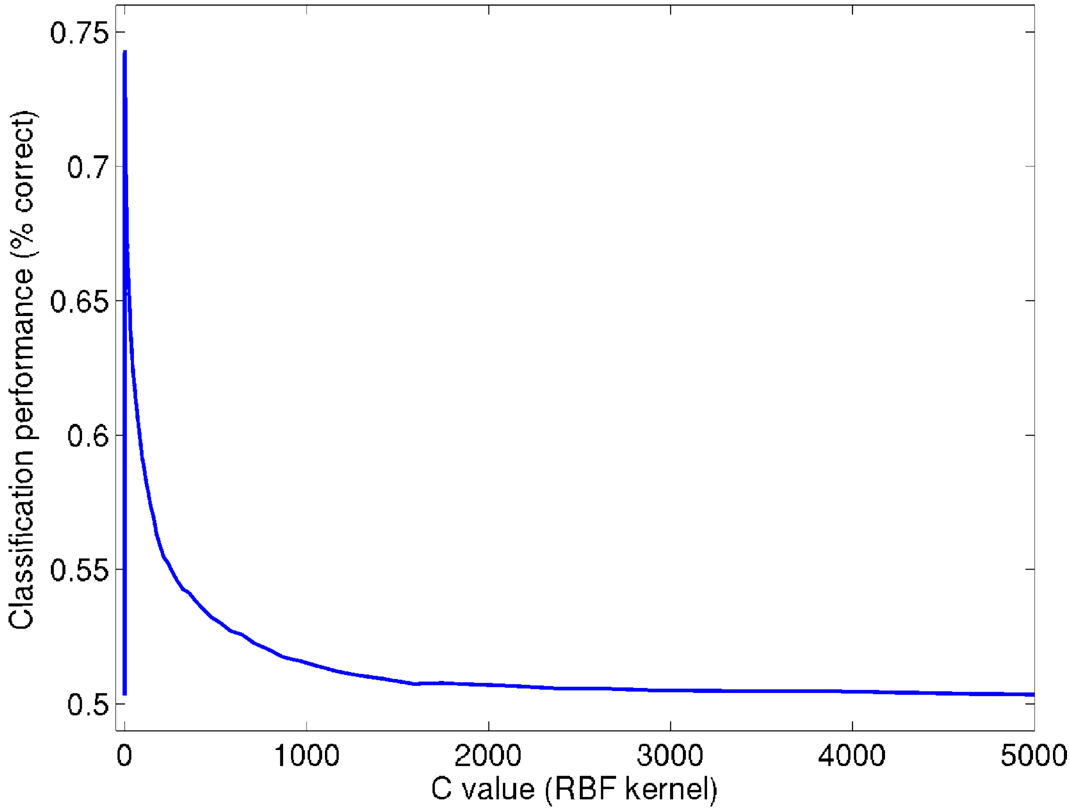

Illustration 95: Magnification of the near-zero end of the previous plot

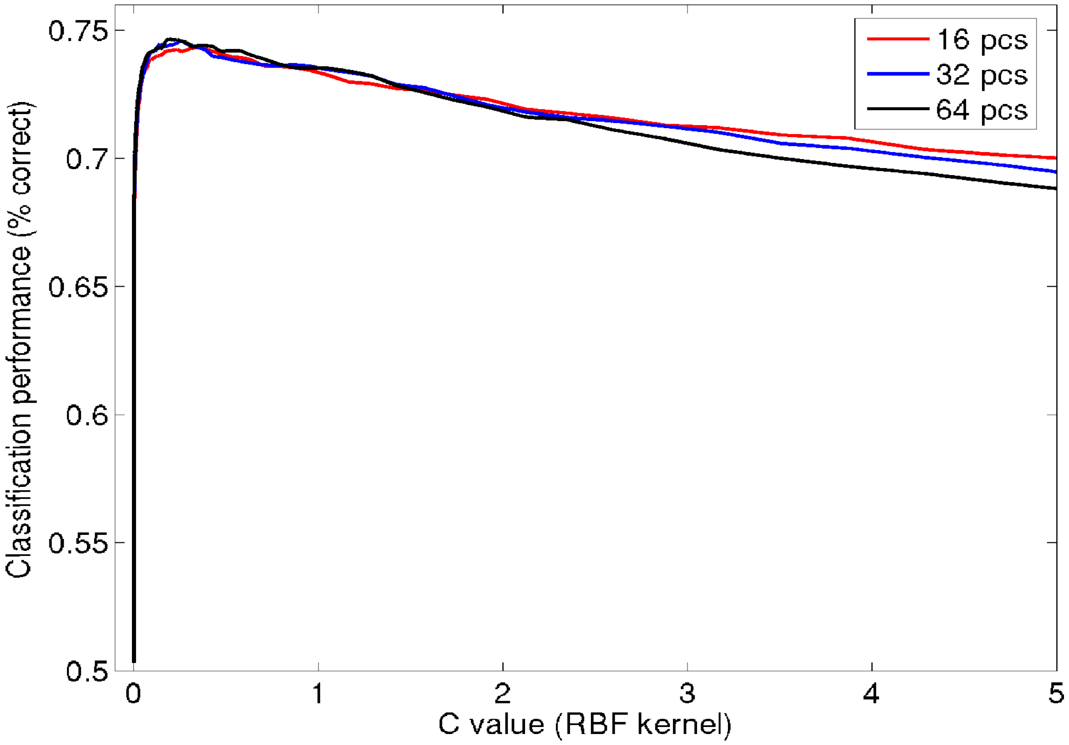

Illustration 96: Further magnification of the near-zero area; differences can be seen between different dimensionalities (see chapter 3.4.1.5 for details)

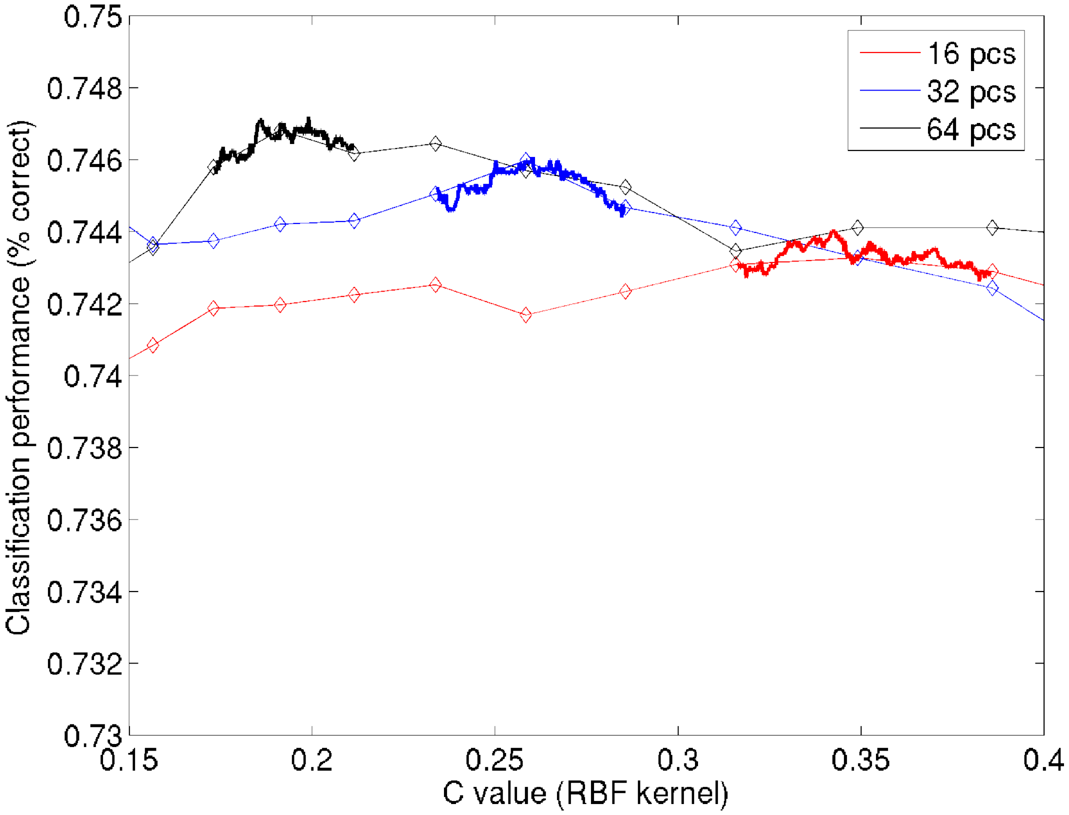

Illustration 97: Magnification of the area around the maximum; dashed lines are from the first stage, solid lines from the second stage of the C-search

8. INDEX OF TABLES

9. INDEX OF ILLUSTRATIONS

10. BIBLIOGRAPHY

[**Aubertin et al 1999**]: Aubertin A., Fabre-Thorpe M., Fabre N, Geraud G., [Fast visual categorization and speed of processing in migraine], *C R Acad Sci III.* 1999, 3228 695-704

[**Bacon-Mace et al 2005**]: Bacon-Mace, N; Mace, M J-M; Fabre-Thorpe, M; Thorpe, S J, The time course of visual processing: Backward masking and natural scene categorisation, *Vision Research* 2005, 45 (2005) 1459-1469

[**Boser Guyon Vapnik 1992**]: Boser BE; Guyon IM; Vapnik VN, A training algorithm for optimal margin classifiers, *5th Annual ACM Workshop on COLT, Pittsburgh, PA* 1992, 144-152

[**Butz 1998**]: Tilman Butz, *Fouriertransformation für Fussgänger*, B. G. Teubner Stuttgart Leipzig, 1998

[**DeValois & DeValois 1988**]: DeValois, R. L.; DeValois, K. K., *Spatial Vision*, Oxford University Press Inc., USA, 1988

[**Fabre-Thorpe et al 1998**]: Fabre-Thorpe, M; Richard, G; Thorpe, SJ, Rapid categorization of natural images by rhesus monkeys, *NeuroReport* 1998, 9 303-308

[**Fabre-Thorpe et al. 2001**]: Fabre-Thorpe, M.; Delorme, A.; Marlot, C.; Thorpe, S., A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes, *J Cogn Neurosci.* 2001, 132 171-180

[**Field 1987**]: Field, D. J., Relations between the statistics of natural images and the response properties of cortical cells, *J Opt Soc Am A.* 1987, 4(12) 2379-94

[**Field Tolhurst 1986**]: Field DJ; Tolhurst DJ, The structure and symmetry of simple-cell receptive-field profiles in the cat's visual cortex, *Proc R Soc Lond B Biol Sci* 1986, 228(1253):379-400

[**Hubel Wiesel 1959**]: Hubel DH; Wiesel TN, Receptive fields of single neurones in the cat's striate cortex, *J Physiol* 1959, 148(3) 574-591

[**Itti Koch Niebur 1998**]: Itti, L; Koch, C; Niebur, E, A Model of Saliency-based Visual Attention for Rapid Scene Analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998

[**Johnson, Ohlshausen 2003**]: Johnson, J. S.; Ohlshausen, B. A., Timecourse of neural signatures of object recognition, *Journal of Vision* 2003, 3 499-512

[**Kalaska Crammond 1992**]: Kalaska, JF; Crammond, DJ, Cerebral Cortical Mechanisms of Reaching Movements, *Science* 1992, 1517-1523

[**Keysers Perrett 2002**]: Keysers, C; Perrett, DI, Visual masking and RSVP reveal neural competition, *TRENDS in Cognitive Sciences* 2002, 6 no 3 120-125

[**Kirchner Thorpe 2006**]: Kirchner, H; Thorpe, SJ, Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited, *Vision Research* 2006, 46 1762-1776

[**Oliva et al. 1999**]: Oliva, A; Torralba, AB; Guerin-Dugue, A; Herault, J, Global Semantic Classification of Scenes using Power Spectrum Templates, *Challenge of Image Retrieval (CIR99), Newcastle*, 1999

[**Porat Zeevi 1988**]: Porat M; Zeevi YY, The Generalized Gabor Scheme of Image Representation in Biological and Machine Vision, *IEEE Transactions on Pattern Analysis and Machine Intelligence*

1988, 10 452-468

[**Rousselet et al 2004**]: Rousselet GA; Thorpe SJ; Fabre-Thorpe M, Processing of one, two or four natural scenes in humans: the limits of parallelism, *Vision Research* 2004, 44 877-894

[**Simoncelli Freeman 1995**]: Simoncelli, EP; Freeman, WT, The Steerable Pyramid: A Flexible Architecture For Multi-Scale Derivative Computation, *Proceedings of the 2nd Annual IEEE International Conference on Image Processing, Washington DC* 1995,

[**Tabachnick Fidell**]: Barbara G. Tabachnick, Linda S. Fidell, *Using Multivariate Statistics - 4th ed.*, Allyn and Bacon, 2000

[**Thorpe Fize Marlot 1996**]: Thorpe SJ; Fize D; Marlot C, Speed of processing in the human visual system, *Nature* 1996, 381

[**Thorpe Gegenfurtner et al 2001**]: Thorpe, SJ; Gegenfurtner, KR; Fabre-Thorpe, M; Bülthoff, HH, Detection of animals in natural images using far peripheral vision, *European Journal of Neuroscience* 2001, 14 869.876

[**Torralba, Oliva 2003**]: Torralba, A; Oliva, A, Statistics of natural image categories, *Network: Computation in Neural Systems* 2003, 391-412

[**VanRullen Thorpe 2001**]: VanRullen, R; Thorpe, SJ, Is it a bird? Is it a plane? Ultra-rapid visual categorisation of natural and artifactual objects, *Perception* 2001, 30 655-668

[**Vogels 1999a**]: Vogels, R, Categorization of complex visual images by rhesus monkeys. Part 1: behavioural study, *European Journal of Neuroscience* 1999, 11 1223-1238

[**Vogels 1999b**]: Vogels, R, Categorization of complex visual images by rhesus monkeys. Part 2: single-cell study, *European Journal of Neuroscience* 1999, 11 1239-1255

[**Wichmann Braun Gegenfurtner 2006**]: Wichmann, FA; Braun, DI; Gegenfurtner, KR, Phase noise and the classification of natural images, *Vision Research* 2006, 46 1520-1529