(Preprint)
Yu Jiajie and Christopher J. Henry, "Descriptive Topological Spaces for Performing Visual Search." In: Peters J., Skowron A. (eds.), Transactions on Rough Sets XXI. Lecture Notes in Computer Science, vol. 10810 (Berlin & Heidelberg: Springer, 2019). DOI: 10.1007/978-3-662-58768-3_2.

Descriptive Topological Spaces For Performing Visual Search

Jiajie Yu * and Christopher J. Henry

University of Winnipeg
515 Portage Avenue Winnipeg, MB, Canada
{yu-j83@webmail.uwinnipeg.ca,ch.henry@uwinnipeg.ca}
http://www.acs.uwinnipeg.ca

Abstract. This article presents an approach to performing the task of visual search in the context of descriptive topological spaces. The presented algorithm forms the basis of a descriptive visual search system (DVSS) that is based on the guided search model (GSM) that is motivated by human visual search. This model, in turn, consists of the bottom-up and top-down attention models and is implemented within the DVSS in three distinct stages. First, the bottom-up activation process is used to generate saliency maps and to identify salient objects. Second, perceptual objects, defined in the context of descriptive topological spaces, are identified and associated with feature vectors obtained from a VGG deep learning convolutional neural network. Lastly, the top-down activation process makes decisions on whether the object of interest is present in a given image through the use of descriptive patterns within the context of a descriptive topological space. The presented approach is tested with images from the ImageNet ILSVRC2012 and SIMPLIcity datasets. The contribution of this article is a descriptive pattern-based visual search algorithm.

Keywords: Human visual search, guided search model, bottom-up attention, top-down attention, salient objects, visual field, descriptive topological space, descriptive proximity, descriptive set intersection, convolutional neural network.

1 Introduction

The problem considered in this article is the automation of visual search motivated by behaviour performed by the human visual system. The problem of visual search is the process of identifying an object in our field-of-view (FOV) amongst many distractor objects, *i.e.* objects that are not the object of interest. This visual search and it is dependent on our ability to direct visual attention.

* This research has been supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant 418413, and the Faculty of Graduate Studies at the University of Winnipeg. Also, special thanks to Keith Massey for developing the code that produced the VGG object descriptions.

This is a complex task that humans perform seamlessly. The aim of visual search systems are to automatically mimic human behaviour when processing the output of optical sensors, whether images or frames in a video sequence. The task of visual search performed by the human visual system is an active area of research in psychology [1]. The solution presented here is based on the definition of visual search as a type of perceptual task that directs attention [2]. In this context, perceiving particular objects is an act of selective attention, where the selective attention mechanism serves to link the processes of perception, action and learning [3].

The attention model used in this work is the guided search model (GSM) [4], which consists of modelling two types of visual search-based selective attention mechanisms: namely, bottom-up and top-down models. With respect to the GSM, the bottom-up approach focuses on modelling salient regions in the FOV. On the other hand, the top-down attention approach models the selection of a desired object from among salient regions identified by the bottom-up method, where salient regions are matched to some representation of the desired object in human memory. From a systems point of view, the top-down approach is a user-guided attention model.

Practical application of the GSM requires a theoretical framework to relate information captured in digital images to perceptual objects in the FOV. Inspired by [5–8], this work implements the GSM within the context of descriptive topological spaces, where the perceptual objects are pixels obtained from digital images. Here, the bottom-up model is implemented by the graph-based visual saliency (GBVS) method [9], and the top-down model is achieved with descriptive topological spaces and a descriptive proximity relation. Results are generated by using a digital image to represent an object of interest, called a query image, which is compared with other images from an image dataset. The solution consists of generating saliency maps (using the approach in [9]) from two images under consideration and using a descriptive proximity relation – defined within a descriptive topological space – in making decisions on the presence of query objects.

This article is based on the work reported in [10], and the contribution of this work is the Descriptive Visual Search System (DVSS) defined in the context of descriptive topological spaces. The DVSS represents the first attempt of using a descriptive pattern-based visual search algorithm, as well as the first use of convolutional neural networks to generate perceptual object descriptions within a perceptual system. Similarly, this article introduces a novel tolerance-based extension of descriptive intersection for use in the DVSS. Finally, the article is organized as follows. Section 2 presents background material on the visual search program and near set theory. Section 3 gives the theoretical framework implemented in the DVSS. Section 4 describes the implementation details of the DVSS, and Section 5 presents results and discussions. Finally, the article is concluded with Section 6.

2 Background

This section provides context for visual search models and descriptive near set theory, both of which are used to produce the DVSS.

2.1 Visual Search Psychological Model

Visual search is the act of finding a visual object[1] of interest in an FOV containing many distractor objects. This problem is considered a perceptual task, where the focus, within the human visual system, is to direct attention. In our case, we aim to mimic the human ability to quickly and effortless find objects of interest in our FOV [2]. The model of human visual attention used in this article has two aspects, namely bottom-up [11] and top-down attention [3]. The bottom-up approach is a type of instinctual, non-guided attention mechanism and is scene-dependent. In contrast, the top-down approach is a user-guided attention mechanism, or task-dependent.

There are four main psychological models that could be a basis for a computational model for visual attention, namely feature integration theory (FIT) [12], biased competition (BC) hypothesis [13], integrated competition hypothesis (IC) [14], and guided search model (GSM) [4]. Based on the application, there are two approaches to modelling the human approach to searching for objects in our FOV. One is object-based [15]. It involves the analysis of parts of an object and it may be used to recognize an object. The other is space-based [15], and it is concerned with the location or position of an object. Feature integration theory [12] is usually used for space-based searching tasks and is typically associated with bottom-up attention models. It assumes that features come first in the perceptual process and are later combined to form objects. Here, the visual scene is coded along multiple feature dimensions, including colour, orientation, texture, and intensity. These features are then fused together when attention is directed at a specific location (hence the *space-based* moniker) in the FOV to form an object. In contrast, the biased competition hypothesis [13] maintains that, regardless of space-based or object-based, the selection according to attention is a biased, competitive process. The competition is among different objects or local area to determine which of them is a reasonable selection according to the relevant task. Further, [13] asserts that the competition is biased toward bottom-up attention in order to benefit local inhomogeneity, *i.e.* locations most distinct from their surroundings are likely to be the winner of the competition for attention. On the other hand, the top-down attention model in the BC hypothesis shifts the bias based on items relative to the current task, such as visual search. Finally, the integrated competition hypothesis [14] is an extension of the BC hypothesis, which posits that any property of the object could be a basis for guiding attention. In the other words, object properties are also treated as task-relevant features.

[1] The term *perceptual object* has specific meaning in descriptive set theory and perceptual sytems. Hence, we will use *visual object* to represent any salient object in an FOV.

2.2 Near Sets

In this work, visual objects inherent to digital images are denoted by sets within a descriptive topological space, and the aim is to assess the nearness or apartness between these disjoint sets. In other words, quantifying the nearness between sets is the basis for determining the similarity of visual objects. Inherent to the study of perceptual similarity is the idea of nearness and tolerance, both of which have a rich and rigorous mathematical history [5, 16]. The idea of sets of similar sensations was first introduced by J. H. Poincaré in which he reflects on experiments performed by E. Weber in 1834, and G. T. Fechner's insight in 1850 [17–20]. Poincaré's work was inspired by Fechner, but the key difference is Poincaré's work marked a shift from stimuli and sensations to an abstraction in terms of sets together with an implicit idea of tolerance. Next, the idea of tolerance is formally introduced by E. C. Zeeman [21] with respect to the brain and visual perception. This idea of tolerance is important in mathematical applications, where systems deal with approximate input and results are accepted with a tolerable level of error, an observation made by A. B. Sossinsky [17], who also connected Zeeman's work with that of Poincaré's. In addition to these ideas on tolerance, F. Riesz first published a paper in 1908 on the nearness of sets [22, 23], initiating the mathematical study of proximity spaces and the eventual discovery of descriptively near sets. Specifically, in 2002, Z. Pawlak and J. Peters considered an informal approach to the perception of the nearness of physical objects such as snowflakes that was not limited to spatial nearness [24]. In 2006, a formal approach to the descriptive nearness of objects was considered by J. Peters, A. Skowron and J. Stepaniuk in the context of proximity spaces [23, 25–27]. In 2007, descriptively near sets were introduced by J. Peters [28, 29], followed by the introduction of tolerance near sets [30, 31]. Recently, the study of descriptively near sets has led to algebraic [32, 33], topological and proximity space [6–8] foundations of such sets.

Originally, the notion of nearness between sets, introduced by Riesz, was based on a spatial relationships between sets, called proximity. As has been mentioned, this idea of proximity between sets was recently expanded to include both spatial quantitative interpretation and a non-spatial qualitative interpretation, called descriptive proximity [28, 29, 34, 5, 6]. In this article, the qualitative interpretation of the notion of proximity between sets (*i.e.* description-based) is used. This idea of descriptive proximity between sets is also know as descriptive near set theory.

3 Preliminaries

This section presents descriptive near set theory, which is primarily used in the last stage of the proposed system. However, the ideas presented in this section form a narrative that underlies the entire approach. In particular, visual search is a type of perceptual task that is complementary to the basic inspiration of descriptive near set theory, namely that humans make decisions on nearness of disjoint sets of objects based on perceived features associated with the objects.

In fact, objects in descriptive set theory are labelled *perceptual objects* and the fundamental structure that introduces descriptive near set theory is a perceptual system, which is where this section begins.

3.1 Perceptual System

Sets of perceptual objects and their descriptions form a perceptual system. To begin, a perceptual object [34] is something perceivable that has its origin in the physical world. Thus, perceptual objects are objects which can be perceived in the physical world, using the senses of sight, touch, taste, smell and hearing. However, the focus of this work is visual search and the sense of sight. Thus, in this work the descriptions of the objects are all extracted from digital images. In general, a description is a real-valued tuple representing features of a perceptual object. Each description is a vector of real-valued features associated with each respective object. Continuing on, a perceptual system consists of both perceptual objects and probe functions [34]. Typically, these values are extracted by a series of functions, called a *probe function* [28, 35]. A probe function is a real-valued function representing a feature of a perceptual object [5]. A set of probe functions are used to generate the feature vector that provide descriptions. In this work, probe functions are defined in the context of deep convolutional neural networks [36] and are used to produce feature vectors for each object.

Definition 1 Perceptual System [34]. *A perceptual system* $\langle O, \mathbb{F} \rangle$ *consist of a non-empty set* O *of sample perceptual objects and a non-empty tuple* $\mathbb{F}$ *of real-valued functions* $\phi \in \mathbb{F}$ *such that* $\phi : O \to \mathbb{R}$.

Next, there is a need within a perceptual system to characterize perceptual objects in O. As a result, an object description is given as follows.

Definition 2 Object Description [37, 38]. *Let* $\langle O, \mathbb{F} \rangle$ *be a perceptual system, then the description of a perceptual object* $x \in O$ *is a feature vector given by*

$$\Phi_{\mathbb{F}}(x) = (\phi_1(x), \phi_2(x), ..., \phi_i(x)...\phi_l(x)) \,,$$

where l *is the length of the vector* $\Phi_{\mathbb{F}}$*, and each* $\phi_i(x)$ *in* $\Phi_{\mathbb{F}}(x)$ *is a probe function value that is part of the description of the object* $x \in O$.

Typically, object descriptions are also referred to as feature vectors in other disciplines. Finally, a descriptive neighbourhood of an object is given by the following.

Definition 3 Descriptive Neighbourhood [6] *Let* $x, y \in O$ *be perceptual objects with object descriptions given by* $\Phi(x), \Phi(y)$*, and let* $\varepsilon \in \mathbb{R}$*. Then, a* description-based neighbourhood *is defined as*

$$N_{\Phi(x)} = \{y \in O : |\Phi(x) - \Phi(y)| < \varepsilon\}.$$

A point y *is a member of* $N_{\Phi(x)}$*, if and only if,* $|\Phi(x) - \Phi(y)| < \varepsilon$.

This definition will be used to produce neighbours of a certain point.

3.2 Descriptive Topologies

Descriptive topological spaces are a significant portion of the proposed visual search system, and, since the human FOV is simulated with digital images, this section introduces a descriptive topological framework defined in the context of digital images [6]. Recently, much work has been reported regarding descriptive topological spaces that are defined with respect to the descriptive intersection and union of open sets [5, 6, 39, 7, 40]. Keeping this in mind, a topology is defined as follows.

Definition 4 Topology [40]. *For a given set, X, a topology, τ, on X is a family of subsets of X (called open sets) such that:*

1. *X and $\emptyset$ are in τ,*
2. *unions of members of τ are in τ, and*
3. *finite intersections of members of τ are in τ.*

The pair (X, τ) is called topological space, or, in the other words, a nonempty set X with a topology τ on it is a topological space [40]. Correspondingly, a descriptive topological space is obtained when considering set descriptions and descriptive-based set operators [6], which are defined below.

Definition 5 Set Description [6, 39, 37]. *Let $A \subseteq O$ be a set within a perceptual system $\langle O, \mathbb{F} \rangle$, then the set description of A is defined as*

$$\mathcal{D}(A) = \{\Phi(a) : a \in A\}.$$

A new form of topology – called *descriptive topology* – requires new operators analogous to union and intersection. These are given next, both of which use Defn. 5. Note, the descriptive union and the union operators are equivalent (see, *e.g.*, [39]). Thus, a new definition is not given for set union. On the other hand, the descriptive set intersection operator is defined as follows.

Definition 6 Descriptive Set Intersection [5, 6]. *Let A and B be any two sets. The* descriptive (set) intersection *of A and B is defined as*

$$A \underset{\Phi}{\cap} B = \{x \in A \cup B : \Phi(x) \in \mathcal{D}(A) \text{ and } \Phi(x) \in \mathcal{D}(B)\}.$$

Moreover, the formal properties of descriptive intersection depend upon the perceptual system. Based on the definitions given above, a descriptive topology [40] is defined next.

Definition 7 Descriptive Topology [40]. *For a given set X, a descriptive topology, τ_Φ, on X is a family of subsets of X such that:*

1. *X and $\emptyset$ are in τ_Φ,*
2. *descriptive unions of members of τ_Φ are in τ_Φ, and*
3. *finite descriptive intersections of members of τ_Φ are in τ_Φ.*

Here, it is interesting to note that a descriptive topology depends on the underlying perceptual system and it may, in fact, not be a topology. As a result, Defn. 7 can be considered a straightforward translation of the standard definition of topology into the presented descriptive framework and nothing more.

3.3 Descriptive Proximities

A descriptive topology defines a structure that is a collection of sets containing objects with comparable descriptions. The aim in developing the presented visual search system is to make decisions based on the degree of shared descriptions within the topology. As a result, our system relies on the ability to quantify the nearness of members of a descriptive topology. The first step in achieving this goal is the definition of a proximity relation. Proximities are nearness relations among the subsets of X in a topological space (X, τ). In other words, a proximity is a closeness or appartness relation on pairs of subsets of X. In [40], two basic types of proximities are defined, namely traditional spatial proximity and descriptive proximity. In [40], the traditional spatial proximity is considered when nonempty sets that have spatial proximity are close to each other, either asymptotically or with common points. In contrast, [40] defines descriptive proximity as nonempty sets are close provided the sets contain one or more elements that have matching descriptions.

There are a number of well-known proximities [6–8] such as the Čech [41], Efremovič [42], Lodato [43], and Wallman [44] proximities. An example of the simplest proximity, a Čech proximity, δ_c, satisfies the following.

1. $\emptyset \not\delta_c A, \forall A \subset X$,
2. $A \; \delta_c \; B \Leftrightarrow B \; \delta_c \; A$,
3. $A \cap B \neq \emptyset \Rightarrow A \; \delta_c \; B$,
4. $A \; \delta_c \; (B \cup C) \Leftrightarrow A \; \delta_c \; B$ or $A \; \delta_c \; C$.

Finally, a specific descriptive proximity relation is defined below.

Definition 8 Descriptive Proximity Relation [45]. *Given a perceptual system $\langle X, \mathbb{F} \rangle$, with $A, B \in \mathcal{P}(X)$, the descriptive proximity relation is defined by*

$$\delta_\Phi = \{(A, B) \in \mathcal{P}(X) \times \mathcal{P}(X) : A \underset{\Phi}{\cap} B \neq \emptyset\}, \tag{1}$$

where the notation $A \; \delta_\Phi \; B$ reads A is descriptively close to B.

3.4 Descriptive Patterns

Patterns play a pivotal role in the presented approach to measuring the nearness or apartness of visual objects. The descriptive intersection of member sets from two respective patterns is the basis for decisions regarding the presence of visual query objects. These patterns, in turn, are created via pattern generators. Beginning with patterns, the definitions for spatial and descriptive set patterns are given as follows.

Definition 9 Spatial Set Pattern [7]. *A spatial set pattern $\mathcal{P}$ contains sets that are spatially near each other.*

Definition 10 Descriptive Set Pattern [7]. *A descriptive set pattern, $\mathcal{P}_\Phi$, contains sets that are descriptively near a given set and possibly near each other.*

Relying on the definition of a set pattern, a pattern generator is defined as follows.

Definition 11 Pattern Generator [7]. *A pattern generator is a distinguished set that is close to each set in the collection of sets in a set pattern.*

3.5 Tolerance-Based Descriptive Intersection Operator

This section presents a new descriptive set operator based on tolerance spaces and relations [21, 17, 46]. As is discussed below, sets formed in this work are extracted from digital images, where decisions on the similarity of visual objects contained in these images are based on features values extracted from image pixels. The result is that comparison between patterns – generated from sets representing visual objects – is the pivotal step. However, the output of probe functions for two objects perceived to be *the same* is rarely an exact match [47]. As a result, the following operator was defined out of necessity for producing results in presented real-world application.

Definition 12 Tolerance Descriptive Set Intersection [10]. *Let A and B be any two sets. The* tolerance descriptive (set) intersection *of A and B is defined as*

$$A \underset{\Phi,\varepsilon}{\cap} B = \{a \in A, b \in B : \parallel \Phi(a) - \Phi(b) \parallel_2 \leq \varepsilon\},$$

where $\| \cdot \|_2$ is the L^2 norm.

This new definition of descriptive set intersection provides for the introduction of a nuanced version of the descriptive proximity relation. Recall Defn. 3 produced a set of points that are neighbours of a certain point. However, the Defn. 12 gives a similarity measurement between two sets of points.

Definition 13 Descriptive Tolerance Proximity Relation. *Given a perceptual system $\langle X, \mathbb{F} \rangle$, with $A, B \in \mathcal{P}(X)$, the descriptive tolerance proximity relation is defined by*

$$\delta_{\Phi,\varepsilon} = \{(A, B) \in \mathcal{P}(X) \times \mathcal{P}(X) : A \underset{\Phi,\varepsilon}{\cap} B \neq \emptyset\}. \tag{2}$$

4 Descriptive Visual Search System

The proposed descriptive visual search system (DVSS) consists of both the GSM bottom-up and top-down attention models and is implemented in three distinct stages. First, the bottom-up activation process is used to generate saliency maps and to identify salient objects. Second, perceptual objects, defined in the context of descriptive topological spaces, are identified and associated with feature vectors (*i.e.* object descriptions) obtained from a VGG [48] deep learning convolutional neural network. Lastly, the top-down activation process makes decisions on whether the object of interest is present in a given image through the use of descriptive patterns within the context of a descriptive topological space.

4.1 Bottom-Up Attention

The aim of bottom-activation is to guide attention to salient regions of the FOV. These regions can be mapped with respect to the FOV producing *saliency maps*. In other words, a saliency map indicates the degree in which a particular region is unusual or different from its surrounding regions. Within the DVSS, saliency maps are implemented using digital images, which means a region's saliency value is quantified by grey levels. More specifically, each pixel in the saliency map represents a saliency value. The DVSS uses GVBS [9] to translate the RGB pixel values into saliency values using graph theory and ultimately produce these maps. Further, the salient regions of these maps are then used to identify perceptual objects, *i.e.* pixels from the original image, that will subsequently be used to create a pattern generator. It is these generated patterns that are finally used to make decisions on whether or not the (visual) object is contained in a given FOV, *i.e.* a digital image. Finally, a predefined threshold is used to determine which pixels in a saliency map constitute visual objects. Examples of this process are given in Fig. 1.

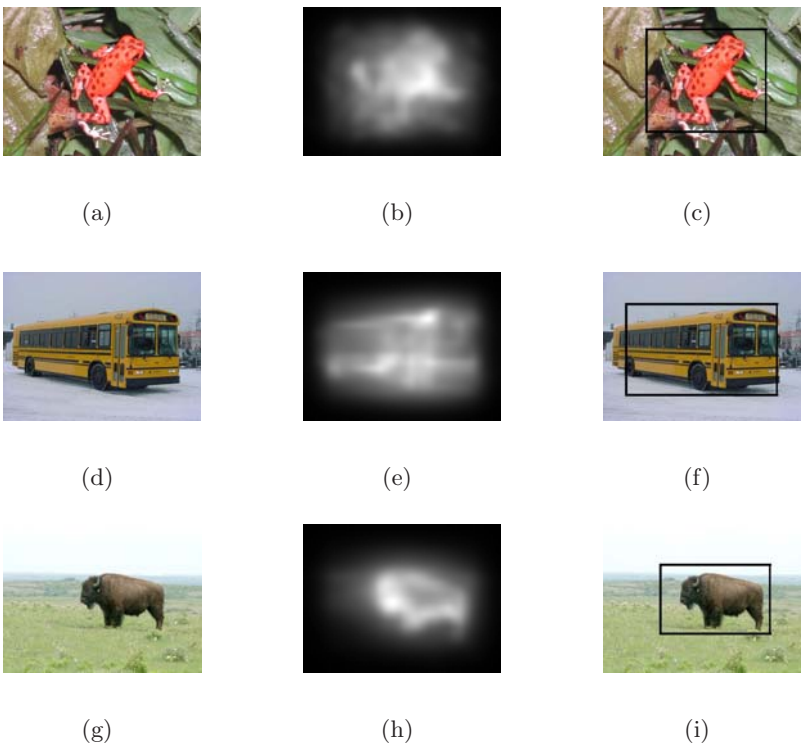

Fig. 1. GVBS [9] saliency map examples. Columns from left to right: original image, saliency map, detected salient object with bounding box.

4.2 Convolutional Neural Network-Based Probe Functions

A pre-trained VGG network [48] was used to extract features in this work. In general, convolutional neural networks (ConvNets) consist of one or more layers, which are labelled as convolutional, fully connected, and pooling layers [49]. Convolutional layers are so named since each neuron develops (*i.e.* learns) a filter during the training process which identifies different types of features within the image. They are named *convolution* since the operation performed by each neuron is analogous to the convolution operation between the filter and the input to the neuron. Pooling layers down-sample the results from the former layer, and neurons in fully connected layers connect all neurons in the previous layer in order to infer classes from the output of the penultimate layer. Fig. 2 present a visual example of a multilayer ConvNet. The VGG ConvNet was developed by the Visual Geometry Group [48], and the VGG network used in this article was pre-trained using the ILSVRC-2012 dataset [50]. This data set contains 1000 categories of images, split into training (1.3 million images), validation (50 thousand images), and testing (100 thousand images with missing class labels) sets.

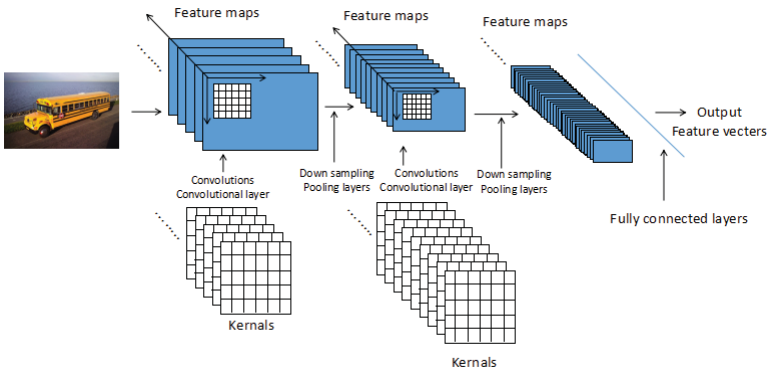

Fig. 2. Example of a standard multilayer convolutional neural network.

In this work, a trained ConvNet was used to extract features (*i.e.* object descriptions) for perceptual objects, where output from the layers were used to produce probe function values (see, *e.g.*, [51]). Any layer in the network could be used as a feature extractor since ConvNets can generate features of visual objects ranging from low-level to high-level. Low-level features are associated with lower levels of a ConvNet and usually describe some low-level digital image characteristic (*e.g.* colour, texture, or edge orientation). They are termed *low-level* features as they are closely related to pixel information and are far removed from global representations of the visual objects in the image. On the other hand, high-level features are associated with higher levels of the network and

are related to the perceptual knowledge of the visual objects within the image. This knowledge is represented by the classes identified by the neural network (*e.g.* people, dog, bus). Further, current deep convolutional neural networks are able to process large numbers of classes with very small differences between categories. For example, the classes of people, dog, and bus can be further expanded to include categories such as seniors, kids, Husky, Chihuahua, school bus, charter bus, *etc.* In all cases, features are produced by presenting an image to the ConvNet input and forwarding outputs through the network until the desired layer is reached. At this point, the features are the outputs of the neurons at that level. In this paper, a 19 layer VGG network consisting of 16 convolutional layers and 3 fully connected layers was used, and the features were extracted from the 13^{th} convolutional layer of the VGG network [48]. Further, layer 13 produces object descriptions of length 512 for each perceptual object. Layer 13 was selected for the DVSS since this level corresponds to higher perceptual representation of visual objects. This is important since the act of visual search does not only rely on low-level features, such as colour or shape, but it also relies on category knowledge and information. In other words, DVSS searchs for visual objects of the same category, but which may have different colours or shapes. In this case, high-level features produce better performance.

4.3 Top-Down Attention

Bottom-up activation guides attention to salient regions, associated with visual objects, that are unusual from their surrounding area, but it does not actively guide attention to visual objects that are the focus of the search. However, visual objects identified by the bottom-up model could be candidate visual objects for further consideration in the visual search process. Thus, in the GSM [4], a top-down attention method is presented to model the act of visual search. This process involves intersections between features associated with salient regions in the FOV and features of the desired visual object, which are, somehow, stored in memory. Thus, attention is based on the result of this comparison. The remainder of this section describes how the top-down model is implemented in the DVSS.

Beginning from an overall perspective, the DVSS starts with two input images, where one is the query image, denoted Q, and the other is a candidate image C. The query image is a representation of the visual object that is the focus of the search process, and it may contain several visual objects. However, to start a searching task, only one query object is selected by the user. Typically, query objects are identified by some domain expert, and this process is analogous to the long term memory discussed in [3]. The candidate image represents the FOV, containing one or more visual objects, all of which are considered as candidate objects for the searching task.

Secondly, GBVS[9] is applied to both query and candidate images, and the output is two saliency maps [11]. The saliency maps are denoted as QS and CS for the query image and candidate image, respectively. As was mentioned, saliency maps are represented by greylevel images, and the pixels associated with salient visual objects are determined by defining a greylevel threshold on

the saliency maps. Any saliency map pixels with values above this threshold are considered part of a salient visual object. Keeping this in mind, only one visual object is stored in memory for the query image, while all the candidate visual objects are stored into memory. In terms of storing the visual objects, the coordinates of pixels associated with each object are stored in the memory.

Continuing on, the top-down activation model is simulated and implemented as follows. Let $X = Q \cup S$, and let $K \subset X$ be a set of perceptual objects (*i.e.* pixels) representing a salient visual object identified using QS or CS. Then, a spatial pattern generator, G, is formed by selecting every l^{th} member of K, where l is some application dependent quantity. Next, descriptive neighbourhoods are found for each member of G to produce a descriptive pattern $\mathcal{P}_{\Phi} = \{N_{\Phi(x)} : x \in G\}$. In other words, all the descriptive neighbourhoods are generated by members of G to form a descriptive pattern, $\mathcal{P}_{\Phi}$. In this light, G_q and G_c represent pattern generators formed from query and candidate visual objects, respectively, and they generate descriptive set patterns for query and candidate objects that are denoted by $\mathcal{P}^q_{\Phi}$ and $\mathcal{P}^c_{\Phi}$, respectively.

Recall, the GSM relates bottom-up and top-down attention models through the intersection of features of salient regions identified from bottom-up mechanisms with knowledge or memory of the visual object that is the focus of the search. Thus, the DVSS operates in the same manner by searching for non-empty intersections between members of the query and candidate patterns. In particular, each member of $\mathcal{P}^q_{\Phi}$ is compared to each member of $\mathcal{P}^c_{\Phi}$ using Def. 12. The result of these intersections (*i.e.* the cardinalities) are then accumulated using a nearness measure inspired by [39]. This process is formalized in Algorithm 1.

Here, it is important to make several observations regarding Algorithm 1. First, as stated in line 14, the nearness of members of the respective patterns is determined by the fraction of objects present in the tolerance descriptive set intersection versus the number of objects in the union of the two sets. Moreover, any value of $s > 0$ implies that the two sets satisfy the descriptive tolerance proximity relation given in Eq. 2. Next, each member of the query object pattern, $\mathcal{P}^q_{\Phi}$, is compared with all the members of the candidate pattern, $\mathcal{P}^c_{\Phi}$ in order to find the most descriptively close (*near*) part of the candidate object, and only the maximum value of s will be used to represent this relationship. Therefore, the final value S is the sum of all the maximum values which are normalized by the total number of members in $\mathcal{P}^q_{\Phi}$. This value S is the final basis for determining whether the query visual object is present in the FOV represented by the candidate image.

5 Results and Discussion

The DVSS was tested by performing image retrieval. In this setup, the query image acts as the object that is the focus of a visual search task and the other images in the dataset represent different visual scenes presented to the FOV. Particularly, images from categories other than the query image represent distractor objects and images from the same category as the query represent the

Algorithm 1: Pattern Based Visual Search

Input : A query image QI, a threshold T
Output: A candidate image CI with objects in bounding box

1 Bottom up;
2 $QS \leftarrow GBVS(Q)$ (Section 4.1);
3 $CS \leftarrow GBVS(C)$;
4 $Q_k \leftarrow$ Selected query object in QS (*i.e.* $Q_k \subseteq QS$);
5 $\mathcal{C} \leftarrow$ All visual objects in CS;
6 Top down; $G_q \leftarrow Generator(Q_k)$ (Def 11);
7 **for** $C \in \mathcal{C}$ **do**
8 $\quad G_c \leftarrow Generator(C)$;
9 $\quad \mathcal{P}^q_\Phi \leftarrow PatternGenerator(G_q)$ (Section 4.3);
10 $\quad \mathcal{P}^c_\Phi \leftarrow PatternGenerator(G_c)$;
11 $\quad$ **for** $PQ \in \mathcal{P}^q_\Phi$ **do**
12 $\quad\quad max_s \leftarrow 0$;
13 $\quad\quad$ **for** $PC \in \mathcal{P}^c_\Phi$ **do**
14 $\quad\quad\quad s = \frac{\left|PQ \bigcap\limits_{\Phi,\varepsilon} PC\right|}{|PQ \bigcup PC|}$;
15 $\quad\quad\quad$ **if** $s > max_s$ **then**
16 $\quad\quad\quad\quad max_s \leftarrow s$;
17 $\quad\quad S \leftarrow S + max_s$;
18 $\quad S \leftarrow \frac{S}{|\mathcal{P}^q_\Phi|}$;
19 $\quad$ **if** $S > T$ **then**
20 $\quad\quad CI \leftarrow Boundingbox(C)$

objective of the search. Image retrieval of this nature, *i.e.* based on the content contained in the images, is called content-based image retrieval (CBIR) [52]. Further, CBIR systems are typically evaluated using precision vs. recall plots [53], which is also the case here. Finally, two data sets were used to generate the reported results, namely the ImageNet ILSVRC2012 [54, 55] and SIMPLIcity [56] datasets.

5.1 Experimental Setup

The ImageNet dataset experiment consisted of 10 categories from the ILSVRC2012 training set, where each category contains 1300 images. Moreover, 10 images from each category were randomly selected as query images. These query images were compared to the remaining 12,900 images, where each comparison produced an S value from line 18 of Algorithm 1. The SIMPLIcity dataset experiment also consisted of 10 categories where each category contains 100 images. In this experiment, each image, in turn, is considered a query image and compared to all

the other images in the dataset for a total of 500,500 comparisons. The SIMPLIcity dataset was also used since the resolution is lower than the ImageNet dataset, which allowed for the larger number of comparisons in a realistic timespan. Again, the images are ranked based on line 18 of Algorithm 1. These ranked S values are sorted in descending order, where the largest value represents the results of the first query, the second value the results of the second query, *etc.* Precision/recall plots are then created based on these values. In the ideal case, all images from the same category as the query are retrieved before any images from other categories. In this case, precision is 100% until recall reaches 100%.

5.2 ImageNet Results

Figs. 3 & 4 contain the average precision vs. recall plots for each category of the ImageNet dataset. Notice, the results in Figs. 3(c), 4(a), and 4(e) indicate that categories *bison*, *school bus*, and *pepper* performed quite well, while the remaining categories did not. Specifically, most curves experience a step drop before recall reaches 20%. These plots could be interpreted as poor performance, however 20% recall corresponds to at least 260 images retrieved out of 12,900 images and is more than a typically user would be interested in an image retrieval system. As a result, the average precision of the top 20 queries for each category is presented in Table 1. Observe, the results are much better, and the top categories (*i.e. bison*, *school bus*, and *strawberry*) correspond to the best plots mentioned above.

Table 1. ImageNet precision values for top 20 retrieved images from each category averaged over 10 query images.

Category	**Precision**	**Category**	**Precision**
Frogs	0.655	Socks	0.625
Turtles	0.635	Teapot	0.655
Bison	0.985	Umbrella	0.465
Cellphones	0.610	Bell pepper	0.935
School Bus	0.920	Strawberry	0.760
	Average 0.7245		

5.3 SIMPLIcity Results

The SIMPLIcity dataset was used to further demonstrate the utility of the proposed approach since only 10 query images per category were used in the ImageNet experiment, whereas all images in the SIMPLIcity dataset were used as query images. Additionally, the SIMPLIcity dataset allowed for comparison with four other CBIR methods [57–60]. These methods are briefly summarized in [10], and they all use low features to represent global content. In contrast, the DVSS

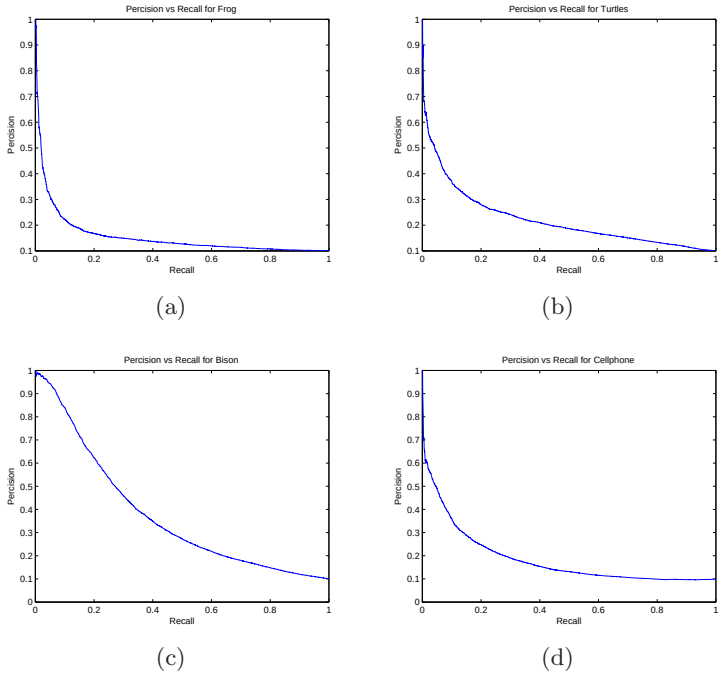

Fig. 3. ImageNet precision vs recall plots for Categories: (a) Frog, (b) turtle, (c) bison, and (d) cellphone.

focuses on salient visual objects and is more localized. Moreover, the features are extracted from higher network levels of the VGG ConvNet that typically produce features associated with visual objects in the images rather than low-level features such as texture or edges, and they are processed within the context of a descriptive topological space. These differences mean the DVSS makes judgements more in line with human perception and understanding of the content within the images. The results of SIMPLIcity experiment are given Table 2.

Notice the proposed approach performs well in the categories *Africans* and *buses*. Markedly, in these categories the DVSS is better than almost all of the other methods, and the DVSS produces the best results with the buses category. On the other hand, the other methods performed very well on the *dinosaurs* category, but the DVSS does not. Upon further investigation, the *Africans* and *buses* categories have analogous categories contained in the training set of the VGG ConvNet[48]. For instance, the ILSVRC-2012 dataset contains images in categories *people*, *school buses*, and *minibuses*. What is more, based on the first experiment, the *school bus* category had very good performance on both precision versus recall plots and top 20 retrieval. Therefore, the DVSS performs very well on the SIMPLIcity *buses* category. On the other hand, the VGG network was not trained with, for example, dinosaur images. Furthermore, the other methods

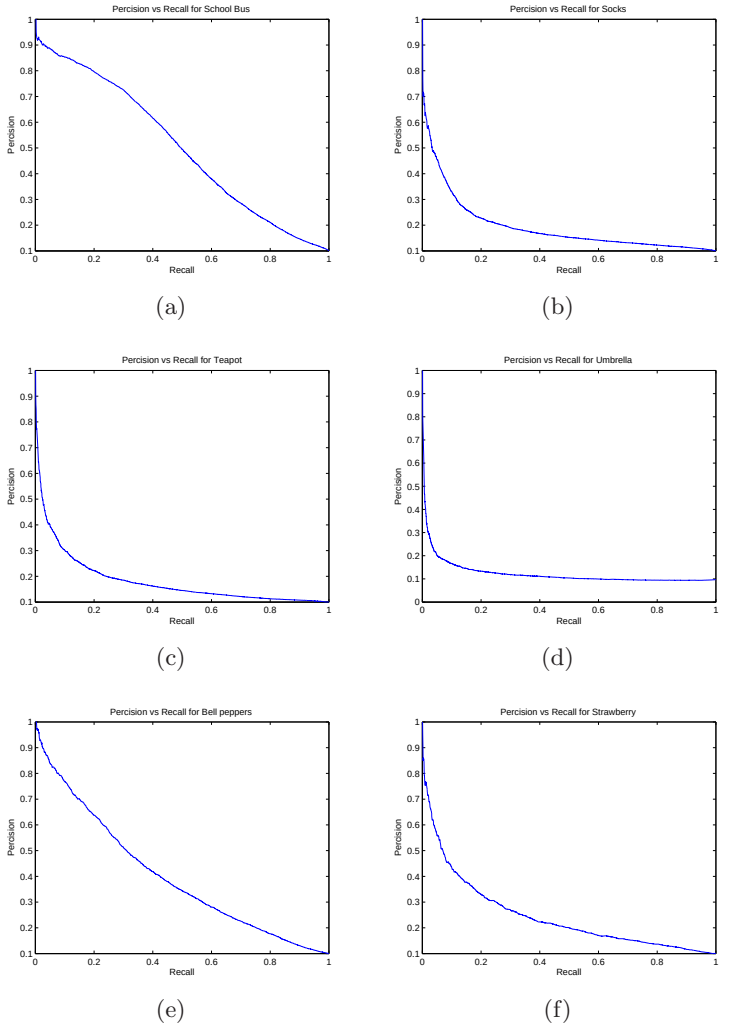

Fig. 4. ImageNet precision vs recall plots for Categories: (a) School bus, (b) socks, (c) teapot, (d) umbrella, (e) bell pepper, and (f) strawberry.

perform quite well on this category because all these images are very similar to each other and are very different from the other categories, thus making for a very clear separation in low-level feature space. Similarly, the other SIMPLIcity categories do not exist in the VGG neural network either [48].

There are other fundamental reasons explaining why the DVSS did not perform as well as the other methods on the SIMPLIcity dataset. For instance, the *beaches* category is characterized by images primarily consisting of background information representing the beaches, sea, and sky. However, the DVSS

was designed to search for salient regions and visual objects in the FOV. Examples in the *beaches* category include people, or umbrellas. As a result, the DVSS uses these objects for quantifying the nearness or apartness of a candidate image with that of a query image instead of the background information which contains the defining features of the category. Another issue compounding the problem is that 19-layer VGG ConvNet down-samples the original input image many times, and and SIMPLIcity images are much lower resolution than the ILSVRC-2012 dataset. Thus, once the images are down-sampled, the corresponding visual objects have only a very few points, which may further weaken the results for categories such as *elephants* and *horses*. Nevertheless, this SIMPLIcity comparisons was important for two reasons. First, it demonstrated that the DVSS performs very well on visual search tasks for sample visual objects that were represented by the VGG training data set. Secondly, the average precision values from Table 1 are better than three of the four methods used for SIMPLIcity comparison. Of course, the results are from different datasets, this observation places the Table 1 values in context for the problem of CBIR.

Table 2. Comparison between the DVSS and four existing approaches using the top 20 precision values for the SIMPLIcity dataset

Category	**[57]**	**[58]**	**[59]**	**[60]**	**DVSS**
Africans	0.5315	0.6975	0.7825	0.683	0.7115
Beaches	0.4385	0.5425	0.4425	0.540	0.3870
Buildings	0.4870	0.6395	0.5910	0.562	0.2950
Buses	0.8280	0.8965	0.8605	0.888	0.9420
Dinosaurs	0.9500	0.9870	0.9870	0.993	0.2570
Elephants	0.3485	0.4880	0.5900	0.658	0.3180
Flowers	0.8835	0.9230	0.8535	0.891	0.4915
Horses	0.5935	0.8945	0.7495	0.803	0.1310
Mountains	0.3080	0.4730	0.3655	0.522	0.3730
Food	0.5040	0.7090	0.6440	0.733	0.2305
Average	**0.5873**	**0.7211**	**0.6866**	**0.730**	**0.4137**

6 Conclusion

This article presented the DVSS that automates the task of visual search using descriptive topological spaces defined in a perceptual system based on probe functions from a convolutional neural network. The results indicate that the proposed solution works very well when employed on data that was also part of the ConvNet training set. As a result, the proposed approach has potential for widespread use and application as machine learning methods based on convolutional neural networks are becoming extremely popular and prevalent. Future work will consist of improvements in execution runtime as the current

approach demonstrates inherent parallelism, but was implemented serially on a CPU rather than using highly parallel co-processors (such as GPUs). Additionally, more investigation will be performed on improving accuracy through using larger convolutional neural networks and larger training datasets.

References

1. Y. Yu, G. K. I. Mann, and R. G. Gosine, "A goal-directed visual perception system using object-based top-down attention," *IEEE Transactions on Autonomous Mental Development*, vol. 4, no. 1, pp. 87–103, 2012.
2. J. Duncan and G. W. Humphreys, "Visual search and stimulus similarity." *Psychological review*, vol. 96, no. 3, p. 433, 1989.
3. Y. Yu, G. K. I. Mann, and R. G. Gosine, "A Goal-Directed Visual Perception System Using Object-Based Top-Down Attention," *IEEE TRANSACTIONS ON AUTONOMOUSMENTAL DEVELOPMENT.*, vol. 4/1, pp. 87–103, 2012.
4. J. M. Wolfe, "Guided Search 2.0 A revised model of visual search," pp. 202–238, 1994.
5. J. F. Peters and S. A. Naimpally, "Applications of near sets," *Notices of the American Mathematical Society*, vol. 59, no. 4, pp. 536–542, 2012.
6. S. A. Naimpally and J. F. Peters, *Topology with Applications.Topological Spaces via Near and Far.* Singapore: World Scientific, 2013.
7. J. F. Peters, *Topology of Digital Images. Visual Pattern Discovery in Proximity Spaces*, ser. Intelligent Systems Reference Library. Berlin: Springer, 2014, vol. 63.
8. J. Peters, *Computational Proximity: Excursions in the Topology of Digital Images*, ser. Intelligent Systems Reference Library. Berlin: Springer, 2016, vol. 102.
9. J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," *Advances in neural information processing systems*, pp. 545–552, 2006.
10. J. Yu, "A Descriptive Topological Framework for Performing Visual Search," Masters.thesis, University of Winnipeg, 2017.
11. L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.
12. A. M. Treisman and G. Gelade, "A feature-integration theory of attention." *Cognitive psychology*, vol. 12, no. 1, pp. 97–136, 1980.
13. R. Dismone and J. Duncan, "Neural Mechanisms of Selective Visual Attention," *Annual Review of Neuroscience*, vol. 18, pp. 193–222, 1995.
14. J. Duncan, G. Humphreys, and R. Ward, "Competitive brain activity in visual attention." *Current opinion in neurobiology*, vol. 7, no. 2, pp. 255–261, 1997.
15. G. R. Fink, R. J. Dolan, P. W. Halligan, J. C. Marshall, and C. D. Frith, "Space-base and object-based visual attention: Shared and specific neural domains," *Brain*, vol. 120, no. 11, pp. 2013–2028, 1997.
16. J. F. Peters, "Near sets," *Wikipedia, The Free Encyclopaedia*, 2015, Edited by C. J. Henry.
17. A. B. Sossinsky, "Tolerance space theory and some applications," *Acta Applicandae Mathematicae: An International Survey Journal on Applying Mathematics and Mathematical Applications*, vol. 5, no. 2, pp. 137–167, 1986.
18. H. Poincaré, *Science and Hypothesis.* Brock University: The Mead Project, 1905, L. G. Ward's translation.

19. L. T. Benjamin, Jr., *A Brief History of Modern Psychology*. Malden, MA: Blackwell Publishing, 2007.
20. B. R. Hergenhahn, *An Introduction to the History of Psychology*. Belmont, CA: Wadsworth Publishing, 2009.
21. E. C. Zeeman, "The topology of the brain and the visual perception," in *Topoloy of 3-manifolds and selected topices*, K. M. Fort, Ed. New Jersey: Prentice Hall, 1965, pp. 240–256.
22. S. A. Naimpally, "Near and far. A centennial tribute to Frigyes Riesz," *Siberian Electronic Mathematical Reports*, vol. 6, pp. A.1–A.10, 2009.
23. S. A. Naimpally and B. D. Warrack, "Proximity spaces," in *Cambridge Tract in Mathematics No. 59*. Cambridge, UK: Cambridge University Press, 1970.
24. Z. Pawlak and J. F. Peters, "Jak blisko (how near)," *Systemy Wspomagania Decyzji*, vol. I, pp. 57–109, 2002.
25. C. J. Mozzochi and S. A. Naimpally, "Uniformity and proximity," in *Allahabad Mathematical Society Lecture Note Series*, vol. 2. The Allahabad Math. Soc., 2009, p. 153 pp.
26. S. A. Naimpally, *Proximity Approach to Problems in Topology and Analysis*. München: Oldenburg Verlag, 2009, iSBN 978-3-486-58917-7.
27. J. G. Hocking and S. A. Naimpally, "Nearness-a better approach to continuity and limits," in *Allahabad Mathematical Society Lecture Note Series*, vol. 3. The Allahabad Math. Soc., 2009, p. 153 pp.
28. J. F. Peters, "Near sets. General theory about nearness of objects," *Applied Mathematical Sciences*, vol. 1, no. 53, pp. 2609–2029, 2007.
29. ——, "Near sets. Special theory about nearness of objects," *Fundamenta Informaticae*, vol. 75, no. 1-4, pp. 407–433, 2007.
30. ——, "Tolerance near sets and image correspondence," *International Journal of Bio-Inspired Computation*, vol. 1, no. 4, pp. 239–245, 2009.
31. ——, "Corrigenda and addenda: Tolerance near sets and image correspondence," *International Journal of Bio-Inspired Computation*, vol. 2, no. 5, pp. 310–318, 2010.
32. E. İnan and M. A. Öztürk, "Near groups on nearness approximation spaces," *Hacettepe J. of Math. and Statistics*, vol. 41, no. 4, pp. 545–558, 2012.
33. J. F. Peters, E. İnan, and M. A. Öztürk, "Spatial and descriptive isometries in proximity spaces," *General Mathematics Notes*, vol. 21, no. 2, pp. 1–10, 2014.
34. J. F. Peters and P. Wasilewski, "Foundations of near sets," *Information Sciences*, vol. 179, no. 18, pp. 3091–3109, 2009.
35. J. F. Peters, "Classification of perceptual objects by means of features," *International Journal of Information Technology & Intelligent Computing*, vol. 3, no. 2, pp. 1 – 35, 2008.
36. F. Li and A. Karpathy, "CS231n: Convolutional neural networks for visual recognition," 2015, Course Lecture Notes, Standford University.
37. C. J. Henry and G. Smith, "Proximity system: A description-based system for quantifying the nearness or apartness of visual rough sets," in *Transactions on Rough Sets XVII*. Springer, 2014, pp. 48–73.
38. C. J. Henry, "Near Sets: Theory and Applications," Ph.D. thesis, University of Manitoba, 2010.
39. ——, "Metric Free Nearness Measure using Description-based Neighbourhoods," *Mathematics in Computer Science*, vol. 7, no. 1, pp. 51–69, 2013.
40. J. F. Peters, "Computational proximity," in *Computational Proximity*. Springer, 2016, pp. 1–62.
41. E. ˇ Cech, *Topological Spaces*. London: Wiley, 2014, fr seminar, Brno, 1936-1939; rev. ed. Z. Frolik, M. Katětov.

42. V. A. Efremovič, "The geometry of proximity I (in Russian)," *Mat. Sb. (N.S.)*, vol. 31(73), no. 1, pp. 189–200, 1952.
43. M. Lodato, "On topologically induced generalized proximity relations," Ph.D. dissertation, Rutgers University, 1962, supervisor: S. Leader.
44. H. Wallman, "Lattices and topological spaces," *Annals of Math.*, vol. 39, no. 1, pp. 112–126, 1938.
45. J. F. Peters, "Local near sets. pattern discovery in proximity spaces," *Mathematics in Computer Science*, vol. 7, no. 1, pp. 87–106, 2013.
46. J. F. Peters and P. Wasilewski, "Tolerance spaces: Origins, theoretical aspects and applications," *Information Sciences*, vol. 195, no. 0, pp. 211–225, 2012.
47. C. J. Henry, "Perceptually indiscernibility, rough sets, descriptively near sets, and image analysis," *Transactions on Rough Sets*, vol. LNCS 7255, pp. 41–121, 2012.
48. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Iclr*, pp. 1–14, 2015. [Online]. Available: http://arxiv.org/abs/1409.1556
49. A. Karpathy, *CS231n: Convolutional Neural Networks for Visual Recognition*. Stanford University, 2015.
50. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances In Neural Information Processing Systems*, pp. 1–9, 2012.
51. C. Perone, "Deep learning – convolutional neural networks and feature extraction with python," http://blog.christianperone.com/2015/08/convolutional-neural-networks-and-feature-extraction-with-python//, 2015.
52. A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, 2000.
53. R. Yates-Baeza and B. Ribeiro-Neto, *Modern Information Retrieval*. New York: ACM Pres/Pearson Addison Wesley, 1999.
54. J. D. J. Deng, W. D. W. Dong, R. Socher, L.-J. L. L.-J. Li, K. L. K. Li, and L. F.-F. L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2–9, 2009.
55. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.
56. J. Z. Wang, J. Li, and G. Wiederholdy, "SIMPLIcity: Semantics-sensitive integrated matching for picture libraries?" in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 1929, 2000, pp. 360–371.
57. N. Jhanwar, S. Chaudhuri, G. Seetharaman, and B. Zavidovique, "Content based image retrieval using motif cooccurrence matrix," *Image and Vision Computing*, vol. 22, no. 14, pp. 1211–1220, 2004.
58. M. Subrahmanyam, Q. M. Jonathan Wu, R. P. Maheshwari, and R. Balasubramanian, "Modified color motif co-occurrence matrix for image indexing and retrieval," *Computers and Electrical Engineering*, vol. 39, no. 3, pp. 762–774, 2013.
59. A. Vadivel, S. Sural, and A. K. Majumdar, "An Integrated Color and Intensity Co-occurrence Matrix," *Pattern Recognition Letters*, vol. 28, no. 8, pp. 974–983, 2007.
60. C.-H. Lin, R.-T. Chen, and Y.-K. Chan, "A smart content-based image retrieval system based on color and texture feature," *Image and Vision*

Computing, vol. 27, no. 6, pp. 658–665, 2009. [Online]. Available: http://dx.doi.org/10.1016/j.imavis.2008.07.004