CENTRO UNIVERSITÁRIO FEI

DANILO NUNES DOS SANTOS

REPRESENTAÇÃO E RECUPERAÇÃO DE IMAGENS POR MEIO DE RELAÇÕES ESPACIAIS ENTRE OBJETOS.

São Bernardo do Campo
2017

DANILO NUNES DOS SANTOS

REPRESENTAÇÃO E RECUPERAÇÃO DE IMAGENS POR MEIO DE RELAÇÕES ESPACIAIS ENTRE OBJETOS.

Tese apresentada ao Centro Universitário FEI, como parte dos requisitos necessários para obtenção do título de de Doutor em Engenharia Elétrica. Orientado pelo Prof. Dr. Paulo Eduardo Santos.

São Bernardo do Campo
2017

Nunes dos Santos, Danilo .
Representação e recuperação de imagens por meio de relações espaciais entre objetos / Danilo Nunes dos Santos. São Bernardo do Campo, 2017.
115 p. : il.

Tese - Centro Universitário FEI.
Orientador: Prof. Dr. Paulo Eduardo Santos.

1. Segmentação. 2. Recuperação de imagens. 3. Representação formal. I. Eduardo Santos, Paulo, orient. II. Título.

Elaborada pelo sistema de geração automática de ficha catalográfica da FEI com os dados fornecidos pelo(a) autor(a).

centro universitário FEI

Doutorado

APRESENTAÇÃO DE TESE
ATA DA BANCA EXAMINADORA

Programa de Pós-Graduação Stricto Sensu em Engenharia Elétrica

PGE-10

Aluno: Danilo Nunes dos Santos **Matrícula:** 513102-4

Título do Trabalho: Representação e recuperação de imagens por meio de relações espaciais entre objetos.

Área de Concentração: Inteligência Artificial Aplicada à Automação

Orientador: Prof. Dr. Paulo Eduardo Santos

Data da realização da defesa: 27/11/2017 **ORIGINAL ASSINADA**

Avaliação da Banca Examinadora

São Bernardo do Campo, 27 / 11 / 2017.

MEMBROS DA BANCA EXAMINADORA	
Prof. Dr. Paulo Eduardo Santos	Ass.: _____
Prof. Dr. Carlos Eduardo Thomaz	Ass.: _____
Prof. Dr. Guilherme Alberto Wachs Lopes	Ass.: _____
Prof. Dr. João Eduardo Kogler Junior	Ass.: _____
Prof. Dr. Flávio Soares Correa da Silva	Ass.: _____

A Banca Examinadora acima-assinada atribuiu ao aluno o seguinte:

APROVADO ☒ REPROVADO ☐

VERSÃO FINAL DA TESE

ENDOSSO DO ORIENTADOR APÓS A INCLUSÃO DAS RECOMENDAÇÕES DA BANCA EXAMINADORA

_____

Aprovação do Coordenador do Programa de Pós-graduação

_____

Prof. Dr. Carlos Eduardo Thomaz

Dedico aos meus filhos Ana Livia e Raul.

AGRADECIMENTOS

Neste espaço, eu quero agradecer as pessoas que contribuiram para a realização desta tese. Contudo, para prestar os devidos agradecimentos, seria necessário muito mais espaço do que disponho aqui. No entanto, espero sinceramente que a menção aqui seja válida.

Gostaria de agradecer ao meu orientador prof. Dr. Paulo Eduardo Santos cujos valorosos conselhos foram fundamentais para o desenvolvimento deste trabalho, mas não se restrigem apenas ao mundo acadêmico, assim, sou muito grato pela orientação. Agradeço também aos prof. Dr. Flávio Tonidandel pelo apoio durante o curso. Expresso também meus agradecimentos aos prof. Dr. Carlos Eduardo Thomaz e prof. Dr. Reinaldo Bianchi pelos conselhos sugeridos como membros da banca avaliadora.

Ao sr. Adam Pease expresso meus profundos agradecimentos pelas contribuições fornecidas com relação ao SUMO e elucidação em questões sobre representação de conhecimento.

Agradeço a CAPES (bolsa PROSUP Taxa número 513102-4) e a FEI por financiarem esta tese, tal apoio foi de extrema importância para a efetivação e sem este suporte seria inviável a execução deste trabalho.

Devo agradecer também aos companheiros do curso e das empresas em que trabalhei durante este período, as discussões a respeito dos temas investigados foram de grande valia para a definição das hipóteses e consequente investigação realizada no decorrer desta tese.

Agradeço minha família pelo apoio incondicional, meus filhos Ana Livia e Raul e minha esposa Ana Paula que compreendeu e apoiou minha dedicação o que ocasionou meu afastamento em diversas atividades familiares e sociais.

RESUMO

Com o crescente número de imagens digitais geradas a cada dia, também aumenta a necessidade de recuperar informação contida em grandes bases de imagens de maneira rápida e precisa. As técnicas aplicadas em recuperação de imagens fazem uso, em geral, de atributos extraídos das imagens e neste espaço de atributos fazem comparações através de métricas de similaridade. Como o foco destas técnicas de recuperação de conteúdo é o uso de atributos de baixo nível, existe um problema pois o conteúdo semântico não está relacionado com estas características de baixo nível. Além disto, as relações entre objetos não são consideradas. Este trabalho aborda este problema conhecido como lacuna semântica (do inglês *semantic gap*) e procura aplicar relações espaciais com o intuito de reduzir o impacto da lacuna semântica na análise de cenas.

Neste trabalho, as relações espaciais são expressas em termos de preposições espaciais. Assim, para definir uma preposição à partir de uma imagem é necessário segmentar os objetos que compõem a imagem e para cada par de objetos segmentados estabelecer a topologia en- tre eles. Para realizar a segmentação, é utilizado redes neurais convolucionais. Considerando uma imagem, a rede neural provê um valor probabilístico pada cada *pixel* existente na imagem, consequentemente, regiões compreendidas por valores estimados referentes à mesma classe são consideradas como o objeto segmentado. Posteriormente, estes objetos são combinados em pares para o estabelecimento de relações binárias entre eles.

De posse dos objetos e seus respectivos segmentos, relações mereotopológicas são estabelecidas de acordo com o *Region Connection Calculus* (RCC). A definição destas relações é efetuada através de um procedimento que visa associar um par de objetos indicados na imagem com uma das oito relações definidas no RCC. Estas relações encontradas, com base neste procedimento, serão traduzidas para preposições linguísticas (existentes na língua inglesa), através de uma outra rede neural.

Por fim, a informação dos objetos e suas respectivas relações é utilizada para estender uma ontologia superior. A ontologia aplicada é o *Suggested Upper Merged Ontology* (SUMO). O procedimento de introduzir uma representação formal permite o estabelecimento de novas relações, desconhecidas *a priori*. Este fato possibilita a recuperação de imagens à partir de relações derivadas através dos axiomas existentes na representação formal aplicada.

Palavras-chave: Representação Formal. Cálculo de regiões conectadas. Redes neurais. Segmentação semântica.

ABSTRACT

Due to the increasing number of digital images generated every day, also increases the need to recover information contained in large databases of images quickly and accurately. The applied techniques in image retrieval, in general, make use of features extracted from images and in this feature space make comparisons through similarity metrics. As the focus of these techniques of content retrieval is the use of low level features, there is a problem because the semantic content is unrelated with these low level features. Besides, the relations between objects are not considered. This work addresses this problem also known as semantic gap and intends to apply spatial relations with the goal of reducing the semantic gap impact in the scene analysis.

In this work, the spatial relations are expressed in terms of spatial prepositions. Thus, in order to define a preposition from an image it is necessary to segment the objects within the image and by each pair of segmented object establish the topology between them. To perform segmentation, neural networks are used. Considering one image, the neural network provides a probabilistic value for every pixel in the image, therefore, the regions covered by similar values belong to the same class and are considered as being the segmented object. Furthermore, these objects are combined in pairs in order to establish binary relations between them.

Relying on the objects and their respective regions, mereotopological relations are established according to *Region Connection Calculus* (RCC). The definition of these relations is performed through a procedure that intends to assign a pair of objects identified in the image with one of the eight relations defined in RCC. These relations established, based on these procedure, are converted to linguistic prepositions (in English language), through another neural network.

Finally, the object information and their respective relations is applied to extend an upper ontology. The applied ontology is the *Suggested Upper Merged Ontology* (SUMO). The procedure of introducing a formal representation allows the establishment of new relations, unknown a priori. This fact enables image retrieval from relations derived through the existing axioms in the applied formal representation.

Keywords: Formal representation. Region Connection Calculus. Neural networks. Semantic segmentation

LISTA DE ILUSTRAÇÕES

Figura 1 – Método proposto 17
Figura 2 – Modelo de rede neural conhecido como *Perceptron* 31
Figura 3 – Modelo de rede neural multicamadas. 32
Figura 4 – Exemplo de convolução em imagens com diferentes núcleos. 33
Figura 5 – Exemplo de anotação para treinamento da rede neural. 35
Figura 6 – Visão geral da etapa de análise quantitativa número do método proposto. 54
Figura 7 – Etapas do processo de segmentação. 57
Figura 8 – Relação topológicas. 59
Figura 9 – Relação topológicas em imagens. 59
Figura 10 – Arquiteturas das redes utilizadas para estimar preposições 62
Figura 11 – Exemplo de imagem segmentada, com três objetos destacados e o fundo negro. 65
Figura 12 – Imagem de exemplo. 69
Figura 13 – Exemplo de imagem segmentada. 69
Figura 14 – Objetos combinados referentes à imagem 13. 70
Figura 15 – Comparação com outras abordagens. 78
Figura 16 – Objetos de referência e alvo identificados. O objeto azul é a referência enquanto o vermelho é o alvo. Objetos destacados conforme anotação manual existente. 84

LISTA DE TABELAS

Tabela 1 – Ontologias divididas pela formalidade 38
Tabela 2 – Relações contextuais ontológicas. 48
Tabela 3 – Arquitetura da rede utilizada. 56
Tabela 4 – Exemplo de relações topológicas de acordo com uma imagem. 64
Tabela 5 – Medidas de precisão e revocação para o modelo estimador de preposição. . . 76
Tabela 6 – Positivos e negativos na tarefa de recuperação de imagens 79
Tabela 7 – Comparação entre os distintos métodos 79
Tabela 8 – Intersecção sobre união (IoU) para as classes de interesse 81
Tabela 9 – IoU com a inclusão do contexto 82
Tabela 10 – Precisão e revocação para classificação de cenas. 83
Tabela 11 – Valores de precisão e revocação da rede neural para detecção de preposições. Ambas as abordagens, implícita e explícita, são ilustradas. 85
Tabela 12 – Matriz de confusão entre as preposições estimadas para o modelo implícito. . 86
Tabela 13 – Métricas da recuperação de imagens 87
Tabela 14 – Consultas por diferentes relações espaciais. 89
Tabela 15 – Distribuição dos objetos no conjunto de treino e teste. Proporção entre os conjuntos de dados, e a medida IoU para cada classe 112

LISTA DE ALGORITMOS

Algoritmo 1 – Algoritmo proposto para a atribuição de uma relação RCC 60

LISTA DE ABREVIATURAS

AKMeans	*Aproximate K-Means*.
AP	*Average Precision*.
BFO	*Basic Formal Ontology*.
BOVW	*Bag of Visual Words.*
BOW	*Bag of Words*.
CBIR	Recuperação de imagens baseada em conteúdo.
CNN	*Convolutional Neural Networks*.
DC**(***x, y***)**	*Disconnected*.
EC**(***x, y***)**	*Externally Connected*.
EQ**(***x, y***)**	*Equal*.
GPU	*Graphics Processor Unit*.
HKMeans	*Hierarchical K-Means*.
HOG	histograma de orientação de gradientes.
IoU	Intersecção sobre união.
IRRA	*Image Retrieval by Region Analysis*.
JEPD	*Jointly Exhaustive and Pairwise Disjoint*.
KIF	*Knowledge Interchange Format.*
LD	Lógica descritiva.
LOG	Laplaciano de Gaussianas.
MAP	*Mean Average Precision*.
NTPP**(***x, y***)**	*Tangential Proper Part Inverse*.
NTPPi**(***x, y***)**	*Non-tangential Proper Part Inverse*.
PCA	Análise de Componentes Principais.
PLN	Processamento de linguagem natural.
PO**(***x, y***)**	*Partially Overlapping*.
Pool	*Pooling*.
QSR	*Qualitative Spatial Reasoning*.
RCC	*Region Connection Calculus*.
SIFT	*Scale Invariant Feature Transform*.
SUMO	*Suggested Upper Merged Ontology*.
SUO-KIF	*Standard Upper Ontology Knowledge Interchange Format*.
SURF	*Speed-up Robust Features*.
SVM	Máquinas de Vetor de Suporte.
TPP**(***x, y***)**	*Tangential Proper Part*.
TPPi**(***x, y***)**	*Non-tangential Proper Part*.
TPTP	*The Thousand of Problems for Theorem Provers*.

SUMÁRIO

1 INTRODUÇÃO 15
1.1 MOTIVAÇÃO 18
1.2 CONTRIBUIÇÃO 18
2 CONCEITOS FUNDAMENTAIS 19
2.1 REPRESENTAÇÃO GLOBAL 19
2.2 REPRESENTAÇÃO LOCAL 21
2.2.1 DETECTORES 21
2.2.1.1 ABORDAGENS BASEADAS NO DETECTOR DE HARRIS 21
2.2.1.2 ABORDAGENS BASEADAS EM LAPLACIANO 22
2.2.1.3 ABORDAGENS BASEADAS NO HESSIANO 23
2.2.2 DESCRITORES 23
2.2.2.1 SIFT 23
2.2.2.2 SURF 24
2.2.2.3 MSER 24
2.2.2.4 ASIFT 24
2.2.2.5 Descritores Binários 25
2.2.2.6 Outras abordagens 25
2.2.3 ASSOCIAÇÃO DE DESCRITORES 25
2.3 APLICAÇÕES UTILIZANDO ATRIBUTOS DE BAIXO NÍVEL 26
2.3.1 DETECÇÃO DE OBJETOS 26
2.3.2 CLASSIFICAÇÃO ATRAVÉS DE BOW 27
2.3.3 ATRIBUTOS EXTRAÍDOS DE MANEIRA NÃO SUPERVISIONADA . . 28
2.4 REDES NEURAIS CONVOLUCIONAIS 30
2.5 VALIDAÇÃO DOS RESULTADOS DA RECUPERAÇÃO 35
2.6 REPRESENTAÇÃO DE CONHECIMENTO 36
2.7 CONCLUSÃO 40
3 TRABALHOS RELACIONADOS 43
3.1 RECUPERAÇÃO DE IMAGENS 43
3.2 SEGMENTAÇÃO SEMÂNTICA 46
3.3 REPRESENTAÇÃO DO CONHECIMENTO 47
3.4 RECUPERAÇÃO DE IMAGENS ATRAVÉS DA SEMÂNTICA 49
3.4.1 ABORDAGENS DE COMPARAÇÃO 50
3.5 CONSIDERAÇÕES FINAIS 50
4 MÉTODO 53
4.1 ANÁLISE QUANTITATIVA 53
4.1.1 SEGMENTAÇÃO DE OBJETOS 53
4.1.2 DETECÇÃO DE RELAÇÕES TOPOLÓGICAS 57

4.1.3 IDENTIFICAÇÃO DE PREPOSIÇÕES ESPACIAIS 61
4.1.4 CRIAÇÃO DE ÍNDICE 63
4.2 ANÁLISE QUALITATIVA 64
4.3 EXEMPLO TRABALHADO 68
4.4 CONCLUSÃO 71
5 TESTES E RESULTADOS 73
5.1 PROCEDIMENTO DE TESTES 73
5.2 TESTES COM CONJUNTO DE DADOS 1 74
5.2.1 SEGMENTAÇÃO DE OBJETOS REFERENTE AO CONJUNTO DE DADOS 1 74
5.2.2 ATRIBUIÇÃO DE PREPOSIÇÕES ESPACIAIS 75
5.2.3 AVALIAÇÃO DA RECUPERAÇÃO DE IMAGENS 77
5.3 TESTES COM CONJUNTO DE DADOS 2 80
5.3.1 IDENTIFICAÇÃO DE OBJETOS 80
5.3.2 IDENTIFICAÇÃO DE CENAS 80
5.3.3 IDENTIFICAÇÃO DE PREPOSIÇÕES 83
5.3.4 RECUPERAÇÃO DE IMAGENS 87
5.4 EXPANSÃO DE CONSULTAS POR UMA REPRESENTAÇÃO FORMAL . . 89
5.5 CONCLUSÃO 90
6 DISCUSSÃO E CONSIDERAÇÕES FINAIS 91
6.1 CONCLUSÃO 95
REFERÊNCIAS 99
ÍNDICE 109
APÊNDICE A – DISTRIBUIÇÃO DOS OBJETOS POR IMAGEM 111
APÊNDICE B – ARTIGO SUBMETIDO 115

1 INTRODUÇÃO

Neste trabalho é investigada a aplicação de relações espacias ao problema de recuperação de imagens. O problema em representar a semântica existente em uma imagem digital tem recebido muito atenção recentemente. Métodos numéricos, definidos como de baixo nível, não são capazes de prover completamente a semântica, definida como de alto nível, devido ao conteúdo semântico ser constituído por conceitos qualitativos. O desafio em integrar informação de baixo nível com conhecimento de alto nível é um problema conhecido em visão computacional, comumente referenciado como lacuna semântica (do inglês *semantic gap*). Neste trabalho, uma abordagem multinível é proposta para recuperar imagens, considerando a semântica, representando o conhecimento de maneira *bottom-up*. Isto é, partindo da informação de baixo nível e combinando-a com informação de alto nível. O método proposto combina pares de objetos para estimar uma relação entre os objetos que compõem este par com respeito a um conjunto, fixo e previamente definido, de preposições espaciais. Este procedimento permite a representação do conteúdo de uma imagem com respeito aos objetos e às relações existentes entre eles.

O intuito deste trabalho é estender os conceitos originais relacionados aos sistemas que fazem detecção de objetos e aplicar uma representação formal que permita realizar consul- tas através de indicadores visuais e obter imagens similares em grandes bases de dados. Este processo segue um paradigma encontrado em inúmeros trabalhos na área de interpretação de imagens, cujo princípio é que para interpretar semanticamente uma imagem é necessário: detectar objetos, definir alguma métrica de comparação e, por fim, representar este conhecimento (WAN et al., 2014).

Assim, este trabalho visa construir uma representação de alto nível de um conjunto de imagens. Esta representação é construída por meio de uma representação formal que descreve as relações espaciais entre os objetos presentes em cada imagem que compõem a base de dados. A construção desta representação formal baseia-se em informações obtidas através do processamento de imagens utilizando atributos de baixo nível, além de conhecimento oriundo de especialistas. Neste trabalho, a definição de conhecimento de especialistas refere-se ao conhecimento espacial obtido através de uma ontologia pública. Em síntese, a hipótese que guia este trabalho é: se a inclusão de conhecimento fornecido por especialistas pode ser utilizada para a assimilação de conhecimento adquirido por agentes autônomos, em imagens digitais, e consequentemente aumentar a percepção destes agentes acerca de um domínio específico.

Para realizar a segmentação de objetos, neste trabalho, aplica-se uma abordagem baseada em redes neurais convolucionais para estimativa da classe à qual cada *pixel* pertence. De acordo com as detecções, distintas regiões são combinadas. É importante mencionar, que o termo região neste momento refere-se à um objeto, visto que, a região é obtida através de um algoritmo de segmentação de objetos, consequentemente neste trabalho ambos os termos região e objeto são intercambiáveis. Tais regiões são combinadas em pares para que se possa definir uma relação binária entre tais regiões. Esta relação binária é definida em termos de preposi-

ções espaciais. Para a definição de relações espaciais é necessário um sistema de referência. Por exemplo, considerando a seguinte relação entre dois objetos x e y, e a relação $R(x, y)$: representando, por exemplo, x à frente de y; três conceitos devem ser definidos: o objeto alvo, o objeto de referência e o sistema de referência (C. HUDELOT; J. ATIF; I. BLOCH, 2008). Neste caso, o objeto alvo é o objeto x e o objeto de referência é y. Por sua vez, o sistema de referência é definido, em geral, do ponto de vista do observador (relativo ou absoluto), ou da maneira na qual tal relação é aplicada: intrínseca, extrínseca ou deítica. Este trabalho foca-se em relações intrínsecas. Adicionalmente, é importante enfatizar que tais relações não são constantes no tempo e consequentemente podem alterar sua situação e deixar de ser intrínsecas de acordo com alterações do ponto de vista (HAYES; BARNES-HOLMES; ROCHE, 2001). A geração destas relações entre objetos em imagens provê meios para que se tenha mais informações à respeito da imagem. Este problema é referenciado na literatura como interpretação de imagens. O problema da interpretação de imagens pode ser definido como a extração de significado de uma imagem. Contudo a semântica em uma imagem, isto é o significado contido, não está disponível de maneira explícita (C. HUDELOT; J. ATIF; I. BLOCH, 2008). Como aplicação prática e, consequentemente, cenário de testes, adotou-se o método proposto na tarefa de recuperação de imagens. Este processo completo compõe a análise quantitativa.

Voltando-se para a análise qualitativa, a representação do conhecimento é construída através de uma ontologia, de acesso livre, denominada SUMO (Adam PEASE; Ian NILES; J. LI, 2002). A SUMO é estendida com dados obtidos na tarefa de análise quantitativa, isto é, distintos objetos segmentados e suas relações expressas através de preposições espaciais para cada imagem existente no conjunto de dados. Por fim, considerando a informação obtida no processo de análise quantitativa, novas relações espaciais foram expressas baseado nas definições existentes originalmente na SUMO.

Em síntese, para representar a semântica existente em uma imagem é necessário extrair atributos de um conjunto de imagens e definir uma métrica de comparação num espaço de menor dimensionalidade para que seja possível estabelecer métricas de similaridades de acordo com um determinado contexto e, por fim, construir uma representação deste conhecimento (WAN et al., 2014). Baseando-se neste paradigma, este trabalho propõe um método que pode ser dividido em duas partes:

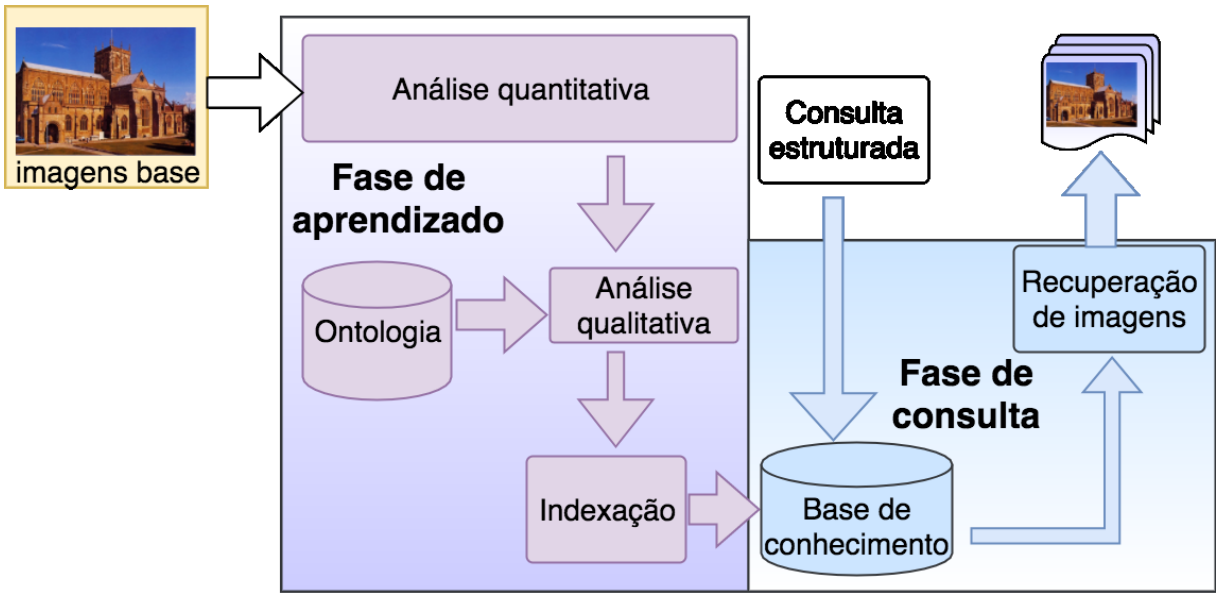
a) **Análise quantitativa** compreende a aplicação de algoritmos e processos computacionais para extrair informação das imagens. É importante ressaltar que a informação extraída oferece meios para possibilitar a interpretação da informação, em geral ruidosa, existente em imagens. De tal maneira que esta informação possa ser entendida por distintos agentes computacionais;

b) **Análise qualitativa** incorpora a informação de baixo nível, oriunda de imagens, extraída na análise quantitativa e a informação fornecida por especialistas em uma base comum, de maneira que seja possível estender o conhecimento obtido pelos

agentes. Além de encontrar possíveis inconsistências oriundas de ruídos existentes em métodos de extração de informação baseados em um processo numérico não determinístico.

A Figura 1 exibe o sistema proposto neste trabalho. A figura 1 ilustra as duas fases que ocorrem no procedimento adotado: aprendizado e consulta. Assim, para uma imagem base, isto é, uma imagem do conjunto de dados de treinamento, aplica-se uma série de procedimentos denominados análise quantitativa. Na sequência o resultado gerado nesta etapa é combinado com uma ontologia (SUMO) e é realizada a análise qualitativa. Por fim na fase de aprendizado tem-se a etapa de indexação, isto é, preparar a informação para que seja possível recuperar imagens que contenham uma determinada relação espacial. Esta informação indexada compõe a base de conhecimento. As consultas estruturadas são processadas de acordo com esta base de conhecimento. Consequentemente, tem-se uma lista de imagens que satisfaçam esta consulta.

Figura 1 – Método proposto



Fonte: Autor.

Em resumo, para a verificação da hipótese apresentada, inicialmente, é necessário extrair informação do conjunto de imagens. Esta informação caracteriza-se por ser de baixo nível, ou seja, informação numérica baseado no valor da intensidade de cada *pixel* existente na imagem, e é independente para cada imagem da base. As imagens são segmentadas de acordo com os objetos que compõem a cena. Estes objetos segmentados são utilizados para estimar preposições espaciais, através de uma relação topológica. Por fim, todas as preposições estimadas de maneira numérica são combinadas em uma base comum alimentada por especialistas, esta base comum é representada por uma ontologia pública existente. Este método proposto é denominado *Image Retrieval by Region Analysis* (IRRA).

1.1 MOTIVAÇÃO

A motivação deste trabalho é reduzir a lacuna semântica existente em aplicações de visão computacional. O intuito é aplicar uma representação formal para viabilizar a interpretação de relações não declaradas explícitamente.

1.2 CONTRIBUIÇÃO

A principal contribuição deste trabalho é a proposta de um arcabouço que contempla a incorporação de atributos de baixo nível, extraído de imagens, com o conhecimento espacial de especialistas representado por uma representação formal.

a) O que este trabalho é:

Este é um trabalho que visa identificar as relações espaciais entre objetos (regiões) existentes nas imagens através de um formalismo previamente definido. A partir destas relações é possível transformar as associações definidas para cada predicado aumentando a percepção dos agentes sobre o domínio.

Este trabalho busca reduzir a lacuna semântica existente atualmente em aplicações que utilizam apenas atributos de baixo nível. A lacuna semântica é o problema, referenciado especialmente na comunidade de recuperação de imagens, como a fundamentação da representação digital em nível perceptual com a interpretação semântica que um usuário associa à imagem em nível linguístico (C. HUDELOT; J. ATIF; I. BLOCH, 2008). Assim, este trabalho visa exprimir relações espaciais entre regiões detectadas na imagem através de um formalismo lógico.

b) O que este trabalho não é:

Este trabalho não pertence à área de reconhecimento de objetos tampouco de classificação. Os avanços recentes obtidos em ambas as áreas são usados apenas como ferramentas para atingir os objetivos propostos.

Este trabalho divide-se da seguinte maneira: no capítulo 2 é realizada uma introdução dos conceitos que serão aplicados; no capítulo 3 é apresentado o estado-da-arte nas áreas em que este trabalho se situa; no 4 o método proposto é descrito em maiores detalhes; no 5 os resultados obtidos são apresentados; e por fim no 6 é realizada uma discussão a cerca do que foi apresentado.

2 CONCEITOS FUNDAMENTAIS

O reconhecimento automático de objetos presentes em uma imagem digital é um campo de pesquisa muito rico. Os primeiros trabalhos na área surgiram em meados da década de 60, contudo apesar do enorme progresso alcançado desde então, ainda hoje tem-se muitos problemas em aberto. Um destes problemas é a lacuna semântica (POPESCU; MILLET; MOELLIC, 2007), existente entre os dados na imagens representados por *pixels* e conceitos semânticos. Conceitos semânticos fazendo referência, neste contexto, a substantivos em linguagem natural representando objetos, e relações entre objetos existentes no mundo real.

A grande maioria dos sistemas de visão computacional faz uso de alguma espécie de reconhecimento seja de objetos, cenas ou categorias. Para realizar esta tarefa de reconhecimento, em geral, elege-se uma representação para os atributos da imagem e então define-se uma medida de similaridade neste espaço de atributos (WAN et al., 2014). Ambos, atributos e similaridade, sendo dependentes do contexto em que é realizado o reconhecimento.

Estes atributos oriundos da imagem podem ser divididos de acordo com o tipo de informação utilizada, isto é, se é feito o uso de características globais ou características locais. Abordagens que adotam representação global para a detecção de objetos utilizam a imagem por completo ou regiões específicas (JAIN; KASTURI; SCHUNCK, 1995). Em contrapartida as abordagens que utilizam apenas uma vizinhança ao redor do *pixel* são caracterizadas como locais (GRAUMAN; LEIBE, 2011). É importante ressaltar que esta representação de imagens, global ou local, é obtida através de características existentes nas imagens.

Este capítulo apresenta uma revisão das técnicas mais utilizadas para representação de atributos extraídos da imagem aplicadas no reconhecimento de objetos ou classificação de cenas. Visto que há uma linha tênue entre estas duas aplicações e a vasta maioria das técnicas apresentadas neste trabalho podem ser aplicadas em ambos os contextos. Este capítulo divide-se da seguinte maneira: Seção 2.1 apresenta técnicas para descrever imagens usando uma representação global, a Seção 2.2 introduz os conceitos aplicados na representação local, a Seção 2.3 apresenta aplicações que utilizam estas técnicas para reconhecimento de objetos e classificação. A Seção 2.4 elucida as redes neurais, especialmente as redes neurais convolucionais que se tornaram um padrão *de facto* na área de visão computacional. A Seção 2.6 apresenta o SUMO como forma de representar o conhecimento obtido sobre um domínio. Por fim a Seção 2.7 apresenta uma conclusão sobre as diversas técnicas apresentadas e a relação destas técnicas com este trabalho.

2.1 REPRESENTAÇÃO GLOBAL

Métodos que representam imagens através de características globais, em geral, utilizam técnicas como: histogramas, análise de componentes principais, área, perímetro, descritores Fourier e momentos de imagem.

A abordagem global mais simples para o reconhecimento de objetos é através do histograma de cores, onde, diferentes classes tendem a ter uma distribuição do histograma de cores distintas. As instâncias destas classes são representadas por objetos similares. Contudo, a detecção baseada apenas no histograma de cores sofre com variações intra classe muito grandes. Isto é, esta metodologia é extremamente dependente do contexto, o que a torna pouco robusta para aplicações práticas, embora seja facilmente combinada com outras abordagens enriquecendo-a e provendo um maior poder de reconhecimento. Por exemplo, no trabalho apresentado em (ARBELAEZ et al., 2012) é utilizado o histograma de cores no espaço HSV em conjunto com descritores de textura e forma na segmentação de objetos.

Referenciando a detecção de objetos específicos, pode-se destacar a Análise de Componentes Principais (PCA). Desde os trabalhos pioneiros de (TURK; PENTLAND, 1991), o PCA é largamente aplicado em reconhecimento de faces. O intuito desta abordagem é reduzir a dimensão de um conjunto de imagens original, baseando-se nos maiores autovetores. Es- tas abordagens são utilizadas como etapa prévia à tarefas de classificação, por exemplo como o apresentado em (KRIZHEVSKY; SUTSKEVER; G. E. HINTON, 2012), onde as imagens são projetadas no espaço vetorial dos autovetores com o intuito de encontrar atributos que se- jam mais discriminantes. Além disso, a técnica descrita em (KRIZHEVSKY; SUTSKEVER; G. E. HINTON, 2012) permite aumentar o conjunto de atributos, através de informação latente, projetando imagens no espaço dos autovetores, consequentemente o classificador tem mais informação sobre cada classe além de evidenciar as características mais discriminantes em cada amostra.

O histograma de orientação de gradientes (HOG) (DALAL; TRIGGS, 2005) também faz uso da informação global existente em uma imagem. Este método consiste em dividir a imagem em blocos e calcular a orientação dos gradientes nessa janela. Esta técnica, combinada com Máquinas de Vetor de Suporte (SVM) é amplamente utilizada para reconhecimento de pedestres (DOLLAR et al., 2012). Inclusive podendo ser utilizada em aplicações de tempo real (SUDOWE; LEIBE, 2011).

Outra técnica largamente empregada, principalmente em aplicações envolvendo recuperação de informação e classificação de cenas, é o GIST (OLIVA; TORRALBA, 2001; TORRALBA; FERGUS; WEISS, 2008). O GIST tem como intuito descrever o conteúdo espacial de uma imagem dividindo-a em janelas, em geral de tamanho $\mathbf{4 \times 4}$, e calculando o histograma de gradientes em cada uma destas janelas, representando-os de maneira única.

Estas abordagens são adequadas para reconhecimento de estruturas globais. Contudo, tem seu desempenho degradado ao se comparar imagens que tenham sofrido algum transformação mais intensa, como mudança de escala ou oclusão (GRAUMAN; LEIBE, 2011; DOUZE et al., 2009). Como alternativa a estes problemas de escala e oclusão, descritores baseados em informação local tem desempenho superior em tarefas específicas de reconhecimento (GRAUMAN; LEIBE, 2011).

Estes atributos globais são utilizados especialmente para construir uma representação compacta, utilizando pouca memória, que possibilite a recuperação especialmente em grandes bases de imagens (DOUZE et al., 2009).

2.2 REPRESENTAÇÃO LOCAL

Métodos que utilizam representação local fazem o reconhecimento através de três etapas sucessivas (GRAUMAN; LEIBE, 2011):

a) Inicialmente, é necessário encontrar estruturas discriminantes na imagem para identificar os pontos de interesse.

b) Adicionalmente, cria-se uma descrição dos pontos encontrados com o intuito de preservar suas características, inclusive após variações de escala, rotação ou transformações afins e de perspectiva.

c) Por fim, a descrição referente aos pontos de interesse é associada com descritores similares existentes em uma imagem alvo, através de alguma medida de distância.

Esta seção apresenta em maiores detalhes as abordagens atuais para a detecção de pontos de interesse, (Seção 2.2.1); para a representação destes pontos (Seção 2.2.2) e associação de descritores (Seção 2.2.3).

2.2.1 DETECTORES

As inúmeras abordagens para detecção de pontos de interesse em imagens utilizando atributos locais, na maioria dos casos, detectam estes pontos de interesse através de abordagens baseadas em Harris Detector, Hessiano e Laplaciano (GRAUMAN; LEIBE, 2011).

2.2.1.1 ABORDAGENS BASEADAS NO DETECTOR DE HARRIS

Os métodos baseados em Harris Detector (HARRIS; STEPHENS, 1998) tem como objetivo procurar por estruturas locais invariantes à translação, o que na prática ocorrerá em pontos que sejam cantos (*corners*). Este método utiliza uma janela w, em geral circular ou elíptica, deslocando-a através de uma imagem I em escala de cinza, visando maximizar diferenças encontradas em cada posição de w. Assim, para uma janela (u,v), tem se:

$$E(u,v) = \sum_{x,y} w(x,y)[I(x+u,y+v) - I(x,y)]^2. \quad (1)$$

Considerando pequenos deslocamentos é possível uma aproximação por:

$$E(u,v) \approx [u \quad v] M \begin{bmatrix} u \\ v \end{bmatrix}, \quad (2)$$

em que M é a matriz de auto correlação, calculada pela derivada da imagem, suavizada por um *kernel* gaussiano $G(x,\tilde{\sigma})$, assim tem-se:

$$M = G(x,\tilde{\sigma}) * \begin{bmatrix} I_x^2(x,\sigma) & I_x I_y(x,\sigma) \\ I_x I_y(x,\sigma) & I_y^2(x,\sigma) \end{bmatrix}. \quad (3)$$

A matriz M é definida como um tensor de estrutura e o conjunto de autovetores correspondem à distribuição do gradiente da janela w centralizada em um ponto p. Baseado no conjunto de autovetores, tem-se que:

a) **2** autovalores mais discriminantes indicam um canto;

b) **1** autovalor mais discriminante indica uma borda;

c) **0** autovalores indica que é uma região homogênea.

A obtenção dos autovalores para as janelas w é um processo computacionalmente custoso. A abordagem proposta em (SHI; TOMASI, 1994) apresenta uma melhora do método inicial utilizando apenas um conjunto mínimo de autovalores. Esta melhoria proporciona aumentos na velocidade de execução sem impactos significantes na detecção.

2.2.1.2 ABORDAGENS BASEADAS EM LAPLACIANO

O Laplaciano ou Laplaciano de Gaussianas (LOG) é um método para detecção de pontos de interesse invariante à escala (LINDENBERG, 1998). Também é conhecido como *blob detector*. Esta abordagem consiste na aplicação um núcleo cujo centro contenha valores positivos e a área ao redor, valores negativos. Um exemplo de tal *kernel* pode ser:

$$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix}. \quad (4)$$

Esta operação é sujeita a ruído, então uma suavização Gaussiana é aplicada para contornar este aspecto. Porém, pelo fato desta operação ser associativa, é possível fazer a convolução Gaussiana com o *kernel* do Laplaciano, gerando um filtro híbrido que é aplicado na imagem. O cálculo do Laplaciano é realizado da seguinte maneira:

$$L(x,\sigma) = \sigma^2(I_{xx}(x,\sigma) + I_{yy}(x,\sigma)).$$

Os pontos identificados na imagem serão maximizados caso a segunda derivada da vizinhança tenha um comportamento similar ao núcleo aplicado. Consequentemente, estruturas homogêneas podem ser combinadas ao se utilizar o mesmo núcleo. Então, este método consegue ser invariante à escala, ao se repetir este procedimento em amostras variando o núcleo Gaussiano e procurando por estruturas que se mantenham constantes em escala e localização.

2.2.1.3 ABORDAGENS BASEADAS NO HESSIANO

Uma outra abordagem para este problema é baseada no operador Hessiano (GRAUMAN; LEIBE, 2011), cujo intuito é detectar pontos de interesse que sejam pontos de máximo e mínimo do gradiente de uma imagem. Assim, é utilizada a matriz Hessiano *H* que contém as segundas derivadas da imagem original calculadas pelo Laplaciano *L*. Para uma melhor definição destes pontos, e visando a redução do ruído, é aplicado um *kernel* gaussiano para suavização da imagem. Essa suavização deixa as regiões mais homogêneas, ao mesmo tempo em que aumenta o contraste entre regiões distintas. Em síntese, a aplicação da suavização busca reduzir o ruído (SZELISKI, 2010). Assim, tem-se a seguinte matriz resultante:

$$H(x,\sigma) = \begin{matrix} L_{xx}(x,\sigma) & L_{xy}(x,\sigma) \\ L_{xy} & L_{yy} \end{matrix}. \quad (5)$$

Para identificação dos pontos de interesse é calculado o determinante da matriz *H*:

$$det(H) = L_{xx}L_{yy} - L^2_{xy}. \quad (6)$$

Após esta etapa é realizada a supressão dos valores não máximos, aplicando na imagem uma matriz 3×3, armazenando apenas o pixel cujo valor seja maior do que os valores dos oito vizinhos imediatos. Assim, esta abordagem irá detectar, em sua maioria, pontos que representam cantos.

2.2.2 DESCRITORES

Após a detecção dos pontos de interesse é necessário representá-los de maneira que seja possível a sua associação com pontos similares sob as mais diversas variações nos pontos de vista. Isto é, deseja-se realizar o alinhamento de um ponto em uma imagem, com o ponto mais semelhante em uma outra imagem da mesma cena, apesar das diversas transformações possíveis. Na sequência são apresentados alguns dos principais descritores.

2.2.2.1 SIFT

O algoritmo *Scale Invariance Feature Transform* (*Scale Invariant Feature Transform* (SIFT)) proposto em (LOWE, 2004) é um dos métodos mais populares para a obtenção de descritores locais (LARLUS; JURIE, 2009). Para cada ponto de interesse detectado é criada uma máscara 4×4 da vizinhança, e para cada elemento desta máscara é computado o histograma da orientação do gradiente, dividindo o espaço de orientações em oito. Consequentemente, um descritor de $4 \times 4 \times 8 = 128$ dimensões é gerado. Por fim, este vetor é normalizado de acordo com o comprimento de cada unidade.

O SIFT serviu de inspiração para inúmeros métodos, porém mostrou-se mais eficaz em relação ao número de transformações suportadas na etapa de associação do que muitas abordagens desenvolvidas posteriormente (MIKOLAJCZYK; SCHMID, 2005). Isto acontece pois o SIFT possui uma maior complexidade de espaço e de memória do que outras abordagens, dada a magnitude de $\mathbf{128}$ dimensões para cada ponto de interesse. Consequentemente, o SIFT é computacionalmente mais custoso. Assim, em aplicações onde memória ou armazenamento sejam aspectos críticos, outras técnicas podem ser mais efetivas oferecendo um melhor desempenho.

2.2.2.2 SURF

Speed-up Robust Features (SURF) foi proposto em (BAY; TUYTELAARS; GOOL, 2008). O descritor é definido em uma região quadrada na vizinhança do ponto de interesse, esta região é rotacionada conforme a orientação definida na detecção. A área compreendida pela vizinhança do ponto detectado é então subdividida em uma matriz $\mathbf{4 \times 4}$ e então para cada uma das regiões é aplicado *Haar wavelets*, e os valores dx, $|dx|$, dy e $|dy|$ são somados. Este processo gera um descritor de 64 dimensões, invariante em escala e rotação.

2.2.2.3 MSER

MSER é um descritor invariante à transformações afins (MATAS et al., 2002). Este descritor conecta *pixels* cuja intensidade é similar a outros, de acordo com um limiar. Em outras palavras, cada um destes *pixels* é agrupado com outros que preservam características similares, formando regiões homogêneas com propriedades que as tornam invariantes à transformações afins.

2.2.2.4 ASIFT

Affine Scale Invariance Feature Transform (ASIFT) (MOREL; YU, 2009) é uma extensão do SIFT padrão. Em sua forma básica, o SIFT cobre apenas um espectro limitado de transformações afins. Isto é, o SIFT consegue reconhecer atributos com até quatro graus de liberdade considerando-se transformações espaciais. Em contrapartida, o ASIFT consegue cobrir seis graus de liberdade.

O conceito básico do ASIFT é extrapolar o algoritmo SIFT gerando imagens sintéticas da imagem original sob diferentes transformações afins. Assim, o ASIFT é capaz de gerar um número muito maior de atributos os quais são comparados de maneira combinatória e depois estatisticamente verificados para diminuição do número de falsos positivos, utilizando a abordagem proposta em (ESPUNY; MONASSE; MOISAN, 2013). Este processo é mais eficiente na associação de imagens em relação ao SIFT original porém o custo computacional é muito maior, o que pode inviabilizar o seu uso em aplicações cujo o tempo seja fator crucial. No

entanto, em certas situações cujas imagens estão sob um número previamente conhecido de transformações, é possível otimizar sua execução gerando apenas um subconjunto das imagens sintéticas. Por um lado isso é menos robusto, por outro lado o tempo de execução é melhorado.

2.2.2.5 Descritores Binários

O intuito de descritores binários é fornecer uma representação concisa dos atributos o que diminui consideravelmente os custos de processamento e armazenamento. O princípio geral é representar um atributo utilizando um número reduzido de *bits*. Como exemplo pode-se citar: BRIEF (CALONDER et al., 2010) e BRISK (LEUTENEGGER; CHLI; SIEGWART, 2011) e FREAK (ALAHI; ORTIZ; VANDERGHEYNST, 2012).

A utilização de descritores binários proporciona ganhos no cálculo, armazenamento e comparação. Contudo, a representação dos atributos desta maneira pode não ser eficaz ao se realizar associações entre imagens que estejam sob diferentes transformações espaciais, como alterações de escala e rotação (LEUTENEGGER; CHLI; SIEGWART, 2011).

2.2.2.6 Outras abordagens

O campo de descritores locais é uma área de pesquisa muito rica. Consequentemente a lista apresentada neste trabalho não é exaustiva. Novas abordagens são constantemente criadas, assim como a combinação de técnicas existentes, por exemplo o ORB que combina o detector FAST com o descritor BRIEF (RUBLEE et al., 2011).

De acordo com o foco da aplicação tem-se uma vasta gama de descritores a ser investigados, como exemplo os descritores densos que consideram localidades e escalas fixas (C. LIU et al., 2008; TOLA; LEPETIT; FUA, 2010). Outro exemplo é encontrado em (KE; SUKTHANKAR, 2004) que segue o princípio do SIFT, porém aplica PCA no espaço de gradientes para obter uma representação do ponto de interesse que seja mais discriminante.

2.2.3 ASSOCIAÇÃO DE DESCRITORES

A associação entre descritores de duas imagens distintas é feita utilizando-se alguma métrica de distância padrão, por exemplo distância Euclidiana. Adicionalmente, pode-se utilizar a distância de Hellinger para aumentar a acurácia da associação, especialmente ao se utilizar o SIFT (ARANDJELOVIC, 2012).

Em especial no caso dos descritores binários, essa associação pode ser realizada utilizando o operador lógico XOR o que permite ganhos consideráveis na velocidade de processamento (MARIUS MUJA, 2012).

O problema em se realizar a associação entre descritores locais é o elevado número de falsos positivos. Isto acontece pois este processo é numérico e não leva em consideração carac-

terísticas da imagem como a localização dos pontos detectados. Contudo, através da aplicação de geometria epipolar, é possível refinar a associação considerando as transformações entre a imagem origem e a imagem alvo. Assim, as associações cuja distância das imagens exceda um limiar pré-determinado são desconsideradas.

Outro método para remoção dos falsos positivos é o *Optimized Random Sampling Algorithm* (ORSA) (ESPUNY; MONASSE; MOISAN, 2013), cuja ideia principal é fazer uma amostragem no conjunto de associações e calcular a matriz fundamental, conforme indicado em (HARTLEY; A. ZISSERMAN, 2004). Na sequência, é calculado o erro global referente à projeção dos pontos da imagem original com a imagem alvo. Este processo se repete até que seja encontrado o melhor conjunto de associações, baseando-se na matriz fundamental que tenha o menor erro de projeção da imagem origem para a alvo.

Estes descritores são exemplos de atributos que são utilizados em aplicações de recuperação de imagens. Vale ressaltar que esta lista não é exaustiva, visto que, para cada tarefa um conjunto distinto de atributos pode ser definido. Contundo, a maior parte dos trabalhos encontrados na literatura utiliza descritores locais, porém é digno de nota que as abordagens mais atuais utilizam como extrator de atributos redes neurais. Exemplos de aplicações que utilizem atributos de baixo nível são apresentadas na sequência (seção 2.3).

Os descritores locais são amplamente utilizados na tarefa de recuperação de imagem, pois permitem a execução de buscas de maneira efetiva, isto é, em tempo aceitável e ainda podem sofrer um pós processamento para aumentar a acurácia das imagens retornadas. Como exemplo de aplicação tem-se o descrito em (PHILBIN; SIVIC; ZISSERMAN, 2008).

2.3 APLICAÇÕES UTILIZANDO ATRIBUTOS DE BAIXO NÍVEL

As técnicas apresentadas nas seções 2.1 e 2.2 servem como base para outras aplica- ções. Por exemplo: detecção de objetos e classificação. Estas técnicas fundamentam a área de detecção de objetos e recuperação de imagens, ambas as quais são empregadas neste trabalho.

2.3.1 DETECÇÃO DE OBJETOS

A competição PASCAL Visual Object Classes Challenge definida em (EVERINGHAM et al., 2010) fomentou o desenvolvimento recente da área de reconhecimento de objetos. Esta competição provê uma base de imagens homogênea permitindo a comparação de maneira uniforme entre diversas técnicas para detecção de objetos. O método Deformable Parts Model (DPM) apresentado em (FELZENSZWALB et al., 2010) obteve altas taxas de acurácia e consequentemente serviu de base para inúmeras aplicações desenvolvidas posteriormente.

O DPM consiste em dois estágios: treinamento e detecção. Assim, para cada objeto que se deseja detectar deve ser calculado um modelo. Consequentemente, como premissa é necessário ter um conjunto de imagens que corresponde ao objeto de interesse (exemplos positivos)

e um outro conjunto com imagens que não possuem o objeto (exemplos negativos). Então, na etapa de treinamento, para cada imagem existente na base é criada uma pirâmide contendo diferentes escalas em cada um dos níveis. É definido um tamanho de janela homogêneo entre todas as escalas. Posteriormente, para cada uma destas janelas é calculado o descritor histograma de orientação de gradientes (HOG). Cada um dos histogramas HOG gerado é classificado atra- vés de um Máquinas de Vetor de Suporte (SVM) binário. Em síntese, são extraídos descritores HOG em múltiplas escalas, então estes descritores são classificados, de maneira supervisionada, em duas classes: positivos e negativos.

A etapa de detecção segue os mesmos procedimentos para a extração de atributos. Isto é, divide-se a imagem em múltiplos espaços de escala e busca predizer quais destas regiões contém o objeto de interesse. O resultado desta classificação é um conjunto de retângulos limítrofes (*bounding boxes*) que são combinados para detectar a localização do objeto.

2.3.2 CLASSIFICAÇÃO ATRAVÉS DE BOW

O desenvolvimento dos descritores locais possibilitou a aplicação de algumas das técnicas utilizadas em processamento de textos no processamento de imagens. Uma das técnicas que seguiram esta direção foi o método *Bag of Words* (BOW), cujo intuito é indexar textos de acordo com a frequência das palavras contidas nele (YANG et al., 2007). Traçando um paralelo para o contexto de imagens tem-se uma adaptação, comumente referenciada como *Bag of Visual Words* (BOVW), onde um banco de imagens é indexado de acordo com os descritores presentes em cada imagem (GRAUMAN; LEIBE, 2011).

O BOVW é um método que atua em duas fases distintas: indexação e consulta. Na fase de indexação, para cada imagem presente no banco, descritores locais são extraídos. Em geral, utiliza-se *Scale Invariant Feature Transform* (SIFT) como técnica de descrição dos pontos de interesse devido à sua maior capacidade de realizar associações espaciais entre imagens distintas (J. PHILBIN et al., 2007). Consequentemente, para fins de clareza, este trabalho referenciará atributos no método BOVW como sendo descritores SIFT de **128** dimensões.

Após a extração de atributos de todas as imagens, é necessária à criação de um vocabulário agrupando estes atributos similares em diferentes palavras (*visual words*). Nesta etapa, aplica-se o algoritmo *K-means* com um determinado número de centróides. Em geral, o número é determinado através de experimentação ou de heurísticas. Contudo, quanto menor o número de centróides, menos discriminante será o método. Em contrapartida, ao se aumentar o número de centróides, é possível que atributos similares sejam atribuídos à diferentes grupos degradando o desempenho.

O resultado do agrupamento gera um conjunto k de atributos representando médias de atributos. A geração destes k grupos proporciona a quantização das imagens do conjunto original. Isto é, a redução da dimensão da imagem em função da quantidade de grupos.

A quantização da imagem ocorre da seguinte maneira: para cada atributo presente em uma imagem, é calculado qual o grupo que melhor representa este atributo através de uma medida de distância. Em outras palavras: dado um conjunto $k \times 128$, cada descritor SIFT é atribuído a um destes k grupos através do cálculo da distância de Hellinger. Por fim, as imagens são representadas de acordo com a frequência que cada k grupo aparece nesta imagem, transformando a dimensão original da imagem $n \times m$ para uma representação $1 \times k$, onde $k << (n \times m)$.

A aplicação do processo de quantização para todas imagens gera um conjunto de histogramas. Para aumentar a eficiência da busca, um índice invertido é utilizado (YANG et al., 2007). Este índice invertido consiste na criação de uma referência bidirecional que permite identificar quais k grupos correspondem à uma imagem, assim como quais imagens possuem atributos representados pelo grupo k_i.

O processo de consulta extrai atributos utilizando a mesma abordagem aplicada no treinamento e quantifica a imagem utilizando o conjunto de k centróides gerado na etapa de indexação. Então, através do índice invertido são verificadas quais imagens da base compartilham o maior número de atributos. Estas imagens são consideradas imagens candidatas a possuírem o mesmo objeto presente na imagem de consulta.

Um dos problemas do BOVW é a não utilização da localidade onde cada atributo é detectado na imagem, consequentemente, isto eleva o número de falsos positivos. Uma abordagem proposta para contornar este problema é realizar a verificação espacial (J. PHILBIN et al., 2007). A verificação espacial consiste na associação dos descritores extraídos das imagens originais, conforme descrito na Seção 2.2.3.

Esta técnica BOVW é aplicada especialmente em recuperação de imagens baseada em conteúdo (YANG et al., 2007) e classificação de cenas (L. FEI-FEI; PERONA, 2005).

2.3.3 ATRIBUTOS EXTRAÍDOS DE MANEIRA NÃO SUPERVISIONADA

Abordagens mais recentes tem investigado a extração de atributos de um conjunto de imagens de maneira não supervisionada. Estes métodos fazem uso de redes neurais profundas para a determinação da representação de atributos que melhor se aplica à um conjunto de dados de acordo com um contexto específico (KRIZHEVSKY; G. E. HINTON, 2011). Baseado no desempenho dos modelos *Convolutional Neural Networks* (CNN), que visam encontrar uma representação que melhor separe as diferentes classes previamente identificadas em um conjunto de dados (KRIZHEVSKY; SUTSKEVER; G. E. HINTON, 2012), o trabalho apresentado em (KRIZHEVSKY; G. E. HINTON, 2011) busca de maneira não supervisionada identificar o conjunto de atributos que melhor represente as imagens existentes.

A grande vantagem destas técnicas, em detrimento das apresentadas nas seções 2.1 e 2.2, é a capacidade de generalização. Isto é, devido a variações no contexto um conjunto de atributos que se mostra eficaz em determinadas tarefas pode se mostrar falho para outra tarefa.

Visto que descritores SIFT são modelados para obtenção de atributos de maneira genérica o que não garante que os atributos detectados sejam ótimos para o conjunto de imagens anali- sados, em contrapartida a obtenção de atributos de maneira não supervisionada consegue uma representação específica para cada conjunto de dados (KRIZHEVSKY; G. E. HINTON, 2011). Adicionalmente, o que impulsionou este desenvolvimento recente foi a criação de bancos de imagens contendo milhões de exemplares, como por exemplo o (RUSSAKOVSKY et al., 2015), aliado ao desenvolvimento de *hardware* através de placas *Graphics Processor Unit* (GPU) as quais proporcionam a computação deste grande conjunto de dados. Vale ressaltar que esta tarefa é supervisionada, pois, em geral, busca-se realizar a classificação de uma base de imagens com relação a um conjunto de classes, porém a extração de atributos é realizada através de filtros convolucionais cujo núcleo é atualizado visando minimizar o custo do algoritmo. Em resumo, a tarefa de classificação é supervisionada, contudo a extração de atributos é automática e específica para o domínio.

A classificação de objetos através de redes neurais convolucionais se tornou o padrão *de facto* em visão computacional após 2012, em grande parte aos resultados obtidos por (KRIZHEVSKY; SUTSKEVER; G. E. HINTON, 2012) na competição ImageNet. Posteriormente, essa tendência se consolidou, especialmente, com os trabalhos de (SZEGEDY et al., 2015; K. HE et al., 2016).

A abordagem mais simples de classificação busca mapear uma entrada x para uma saída y através da aproximação de uma função $f(x) = y$. As redes neurais possuem a particularidade de possuirem diversas camadas que podem ser combinadas conforme o problema em questão. Pode-se considerar a rede neural mais simples, conhecida como *Perceptron*, como um conjunto de pesos W sobre a entrada x, com um parâmetro adicional b. Em resumo, uma rede neural mínima, é expressa por:

$$f(x, W, b) = Wx + b \quad . \tag{7}$$

Através de técnicas de otimização, especialmente gradiente descendente, os parâmetros (W, b) são atualizados.

As redes neurais apresentam uma grande versatilidade, pois podem ser estendidas através da adição de diferentes camadas. Focando-se em aplicações de visão computacional, especialmente classificação de imagens, o principal característica é a inclusão das camadas convolucionais. Conforme descrito na seção 2.4.

Um dos principais pontos entre a classificação de imagens utilizando redes neurais é a seleção de atributos, isto é, os atributos para o classificador são *aprendidos*, visto que a extração de atributos ocorre de maneira implícita, especialmente nas camadas convolucionais. Este fato proporciona um modelo muito mais coerente com o domínio aplicado e por outro lado evita um custoso processamento como, por exemplo, a extração de descritores locais.

Visto que o foco deste trabalho é a recuperação de imagens, e consequentemente é necessário extração de atributos de imagens, optou-se pela utilização das redes neurais. E dado este fato, e por trabalhar com imagens, aplicou-se as redes neurais convolucionais.

2.4 REDES NEURAIS CONVOLUCIONAIS

As redes neurais tornaram-se a principal técnica para a aplicação em problemas de reconhecimento de padrões na área de visão computacional. Esta afirmação é corroborada principalmente por competições como a ImageNet (RUSSAKOVSKY et al., 2015), cujo resultado em 2012 obtido por Krizhevsky, Sutskever e G. E. Hinton (2012) representou uma ruptura na área em termos de acurácia.

Assim, na área de classificação de imagens pode-se definir um marco, dividindo as abordagens em dois grupos: as técnicas anteriores à publicação de Krizhevsky, Sutskever e G. E. Hinton (2012) e após. De tal modo que, na era pré Krizhevsky, Sutskever e G. E. Hinton (2012), talvez o principal método para reconhecimento de padrões em imagens era a utilização do *Bag-of-Words (BOW)*, com a extração de atributos através de descritores locais, como o SIFT (LOWE, 2004) ou descritores globais como o GIST (DOUZE et al., 2009). E posteriormente, a grande maioria de trabalhos na área de classificação de imagens faz uso de redes neurais convolucionais.

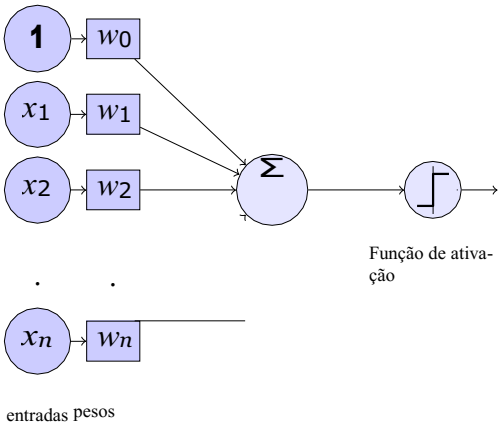
Considerando a rede neural mais simples existente conhecida como *Perceptron*, ilustrada na figura 2, tem-se a base para a construção de outras arquiteturas. Assim, conforme ilustrado na figura 2, tem-se as entradas $I = 1, \cdots, x_n$ para um conjunto n-dimensional. Cada entrada x é conectada à um neurônio através de um peso $W = w_0, w_1, \cdots, w_n$. Adicionalmente, este neurônio aplica uma função degrau que é geralmente uma sigmoide ou uma função tangente. Além dos pesos aplicados à cada entrada, ainda tem-se um parâmetro adicional que não sofre influência da entrada conhecido como *bias*. O viés serve para evidenciar características que podem não ser expressas considerando apenas a informação de entrada disponível, que em termos práticos pode ser ambígua e ruidosa. Por fim este exemplo de neurônio é expresso da seguinte maneira:

$$\theta(x) = \sum_{i=1}^{n} w_i x_i + b \quad . \tag{8}$$

Este modelo é praticamente o modelo de regressão logística. O processo de aprendizado do modelo apresentado na figura 2 é, em geral, realizado através da técnica de gradiente descendente. A técnica do gradiente descendente consiste em comparar o resultado obtido com o esperado através de uma função de custo. Como exemplo, pode-se considerar a seguinte função de custo:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (\theta(x_i) - y_i)^2 \quad , \tag{9}$$

Figura 2 – Modelo de rede neural conhecido como *Perceptron*.



Fonte: Autor.

sendo:

a) m a quantidade de exemplos;

b) $\theta(x_i)$ é a função de ativação para a entrada x_i;

c) y_i a saída esperada para a entrada x_i.

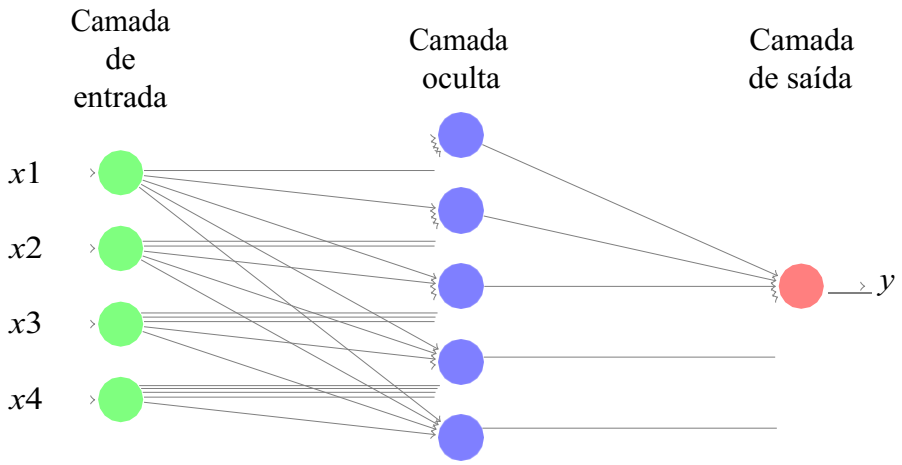
A função de custo possui uma superfície onde cada conjunto de parâmetros distintos estará em uma posição diferente. O objetivo do gradiente descendente é identificar o mínimo da função de custo, consequentemente aproximar a saída do valor esperado. Assim, através do cálculo do gradiente é possível saber a direção que o algoritmo deve tomar considerando-se a superfície da função de custo. Por fim, através de uma taxa de aprendizado λ, os pesos W e o *bias* b são atualizados proporcionalmente conforme a importância para o resultado previamente obtido. Este processo de calcular a derivada para cada peso de acordo com a função de custo e consequentemente atualização dos pesos através de uma taxa de aprendizado é denominado retropropagação.

Este processo é entendido como uma iteração, isto é, o cálculo da função $\theta(x)$, o cálculo do gradiente de acordo com a função de custo $J(\theta)$ e a atualização dos pesos e *bias* (W, b). Este procedimento repete-se até um número máximo de iterações definido ou o modelo atingir o mínimo global da função de custo.

Um fator importante ao se considerar o modelo apresentado na figura 2 é a existência de apenas um neurônio. Isto é, a ilustração na figura 2 contém apenas uma única função de ativação. Em aplicações práticas é comum incluir mais funções e conectá-las às distintas entradas. Este conjunto de funções, cuja entrada parte do mesmo conjunto X recebe a nomenclatura de camada. Na figura 3 três camadas distintas são apresentadas.

Em síntese, é possível combinar distintos neurônios, e consequentemente este conjunto de neurônios forma uma camada. Adicionalmente, e seguindo este mesmo paradigma existe

Figura 3 – Modelo de rede neural multicamadas.



Fonte: Autor.

a possibilidade de se combinar camadas. Em síntese, a saída de um modelo, aplicando uma ou mais funções como por exemplo o apresentado na figura 2, se transforma na entrada de um segundo modelo e assim sucessivamente, esta outra arquitetura é apresentada na figura 3. O modelo apresentado na figura 3 contém três camadas, cada uma contendo um número distinto de neurônios. A nomenclatura das camadas em um modelo como o apresentado na figura 3 é expressa pelo seguinte padrão: a camada entrada de dados, a camada de saída de dados e as demais são denominadas camadas ocultas.

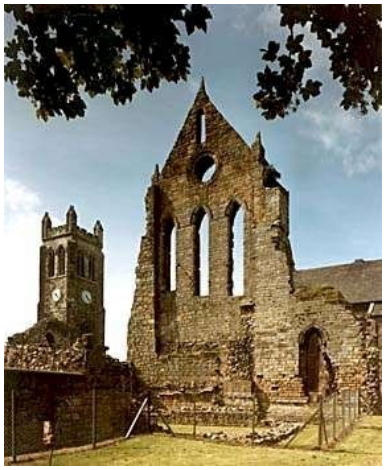
A inclusão de camadas aumenta o número de parâmetros que devem ser aprendidos e tende a aumentar a precisão, contudo ao se adicionar muitos parâmetros o modelo perde em generalização e consequentemente degrada a acurácia. Vale ressaltar que, conforme o modelo apresentado na figura 3, a camada oculta é do tipo densamente conectada, pois todas as dimensões da camada de entrada, isto é $X = x\mathbf{1}, x\mathbf{2}, \cdots, xn$ possuem ligação com todos os neurônios existentes na camada oculta. Estes modelos multicamadas são comumente referenciados na literatura como modelos de aprendizado profundo (*deep learning*).

Focalizando-se na classificação de imagens, um tipo de camada em especial é adotada, que é a camada convolucional. De maneira sucinta, a camada convolucional recebe uma imagem representada através de uma matriz de três dimensões e aplica um filtro convolucional através de um núcleo *k*. Este filtro aplica a operação de convolução entre uma imagem de entrada e este núcleo *k*. Este núcleo é em geral de um tamanho menor que a imagem, assim pode ser utilizado em janelas na imagem que possuam o mesmo tamanho. Isto é, aplicando a convolução em janelas na imagem. Adicionalmente, através do deslocamento distintas janelas da imagem são convolucionadas, extraindo assim informação local. Para cada convolução apenas um valor é considerado através de uma função de normalização, sendo a mais usada uma função que limita os valores negativos. Este procedimento é ilustrado na figura 4, a figura 4a representa uma imagem original, a figura 4b exibe a imagem da figura 4a convolucionada com um núcleo cujo o efeito é aplicar a suavização de maneira moderada, a figura 4c apresenta a mesma

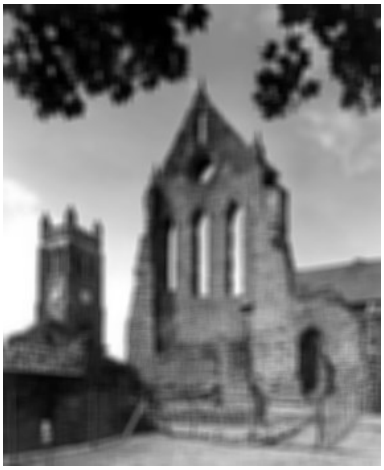
imagem da figura 4a convolucionada com um núcleo cuja função é amplificar a suavização, e por fim a figura 4d apresenta a imagem 4a convolucionada com um filtro cujo o objetivo é a detecção de linhas.

Figura 4 – Exemplo de convolução em imagens com diferentes núcleos.

(a) Imagem original.

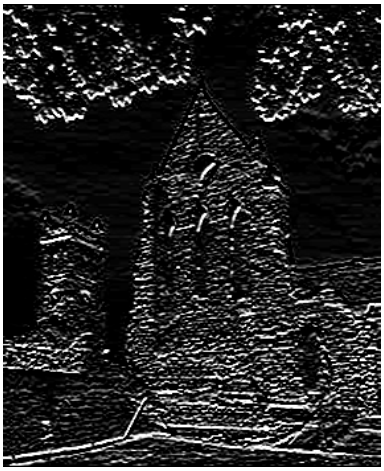


(b) Exemplo de convolução com suavização moderada.



(c) Exemplo de convolução com suavização intensa.

(d) Exemplo de convolução para detecção de linhas.



Fonte: Autor.

Em resumo, uma camada convolucional aplica a convolução em diversas janelas da imagem e um único valor (*máximo*) é considerado para cada bloco. Cada um destes valores é combinado formando uma outra matriz resultante, conhecida como mapa de ativação. Uma camada convolucional, possui distintos filtros, assim o resultado da aplicação de uma camada convolucional é um conjunto de mapas de ativação. Por fim, vale ressaltar que da mesma maneira que o *Perceptron*, estas camadas convolucionais também podem ser encadeadas. No processo de aprendizado, através da regra da cadeia, tem-se a derivada de cada camada e considerando a taxa de aprendizado, atualiza-se o valor dos filtros.

Traçando-se um paralelo com as abordagens clássicas que utilizam descritores, como por exemplo SIFT (LOWE, 2004), as camadas convolucionais criam mapas de ativação, que podem ser entendidos como atributos, de acordo com os distintos valores de intensidade dos *pixels* existentes na imagem. Contudo, a extração baseia-se no domínio aplicado, visto que é um processo iterativo que é refinado a cada iteração, em contraste com os descritores que são genéricos. Em síntese, fazendo uma abstração, o intuito da aplicação de uma rede neural convolucional é identificar o melhor conjunto de atributos para descrever um domínio e consequentemente aumentar a acurácia de um classificador neste domínio.

A arquitetura básica de uma rede neural convolucional consiste em algumas camadas convolucionais na entrada e a inclusão de algumas camada densas e por fim uma sigmoide que fornece a verossimilhança da imagem pertencer à alguma das classes alvo. No entanto, neste trabalho, para a segmentação de objetos, antes da camada de sigmoide é aplicada uma convolução transposta, que resulta em um mapa de ativação com tamanho da imagem de entrada original. A convolução transposta é basicamente a mesma operação realizada na convolução no retrocesso. E de maneira análoga, no retrocesso a operação realizada é a convolução. Essa propriedade permite escalar a matriz de verossimilhança para uma matriz de verossimilhança com tamanho da imagem original. Assim, a saída da rede neural será uma matriz com dimensões $M = (N, W, H)$ sendo:

a) N número de classes esperado;

b) W largura da imagem de entrada;

c) H altura da imagem de entrada.

Cada elemento da matriz M possui a probabilidade de ser de uma das classes do conjunto N . Ao considerar apenas o máximo da probabilidade, tem-se uma matriz de dimensões (W, H) com as mesmas dimensões da imagem original. Consequentemente com cada *pixel* fazendo referência a um objeto de interesse. Vale ressaltar que a probabilidade é obtida através da função *Softmax*, a qual é utilizada para representar uma distribuição de categorias, isto é, a distribuição de probabilidades com relação as diferentes possíveis saídas.

A técnica adotada inclui-se no conjunto de aprendizado de máquina. Adicionalmente, situando-se no grupo de aprendizado supervisionado. Esta abordagem divide-se em duas etapas complementares e hierárquicas, que são: etapa de treinamento e etapa de previsão.

A etapa de treinamento necessita de um conjunto de imagens anotadas onde cada *pixel* é identificado de acordo com a classe a que pertence. Assim, cada *pixel* recebe um único e exclusivo valor. Este valor deve ser um rótulo, como por exemplo o valor RGB: $(255, 255, 255)$, que é mapeado para uma classe. A figura 5 ilustra como deve ser a anotação para treinamento da rede neural. A figura 5a pertence à uma imagem em seu formato original, a imagem 5b corresponde à anotação realizada. É importante ressaltar que na figura 5b objetos similares, por exemplo pessoas, devem possuir a mesma classe, ou seja, terem a mesma cor. No paradigma

adotado neste trabalho, o fundo, assim como objetos que não sejam de interesse são rotulados como RGB=**(0,0,0)**, ou seja, o fundo da imagem é anotado com valor zero.

Figura 5 – Exemplo de anotação para treinamento da rede neural.

(a) Imagem original.

(b) lmagem anotada.

Fonte: Autor.

Este trabalho faz um largo uso de redes neurais conforme apresentado no capítulo 4. De acordo com a informação gerada pela rede neural é construída uma representação deste conhecimento.

2.5 VALIDAÇÃO DOS RESULTADOS DA RECUPERAÇÃO

Considerando os sistemas de recuperação de informação, a informação relevante para uma dada consulta deve ter como resultados relevantes apenas uma fração mínima de toda a base de informação. Contextualizando para este trabalho, dada uma consulta estabelecendo uma relação espacial entre dois objetos, apenas um subconjunto das imagens deve ser rele- vante. Assim, para avaliar o resultado da informação que o sistema retornou busca-se verificar o conteúdo retornado com o padrão ouro. Consequentemente, a posição em que a imagem é retornada tem uma maior importância, isto é, imagens nas primeiras posições tendem a ter um peso maior na avaliação do sistema.

A medida mais adotada para avaliar sistemas de recuperação de informação é o *Average Precision* (AP). O AP consiste em avaliar a qualidade da informação recuperada, ponderando-a pela posição em que aparece. Adicionalmente, ao se considerar um sistema de recuperação de imagens, pode-se desejar medir apenas os primeiros *r* resultados. Para se calcular o AP aplica-se a seguinte fórmula expressa, para cada consulta realizada:

$$AP = \frac{\sum^{r} P\,r}{R} \quad , \tag{10}$$

sendo, *r* a posição de cada elemento retornado, *R* o número total de documentos relevantes, e *P r* é a precisão dos primeiros *r* documentos (L. LIU; ZSU, 2009). A avaliação de apenas uma consulta é ineficiente para avaliar o desempenho de um sistema, portanto, adota-se a média do AP entre todas as consultas realizadas para verificar a qualidade da informação recuperada. Assim, em termos práticos, o índice que é utilizado para medir o desempenho de sistemas de

recuperação de informação é o *Mean Average Precision* (MAP), que corresponde à média de AP de todas as consultas executadas.

2.6 REPRESENTAÇÃO DE CONHECIMENTO

Esta seção elucida a abordagem adotada para construção da representação do conhecimento oriunda de uma análise quantitativa - baixo nível - em termos de uma ontologia espacial.

Existem diversas definições de ontologia encontradas na literatura (SANKAT; THAKUR; JALOREE, 2016). Neste trabalho, ontologia é referenciada de acordo com a definição proposta em (GRUBER, 1993), que define ontologia como uma representação formal de conhecimento, em geral, restrita a um domínio específico (LEHMANN; VÖELKER, 2014). Esta representação é expressa de tal maneira que possa ser compreendida por processos computacionais, em especial, na área de inteligência computacional (SANKAT; THAKUR; JALOREE, 2016). Por conseguinte, ontologia pode ser referenciada como um entendimento comum a respeito de um domínio de interesse, sendo expressa por: conceitos, relacionamento entre conceitos, funções e instâncias (BHOWMICK; ROY; BASU, 2010).

Ontologias são consideradas como uma fração da área de representação de conhecimento. Sistemas de representação de conhecimento, em geral, utilizam uma representação formal de ontologia e mecanismos de inferência aplicados sobre esta ontologia a fim de empregar algum tipo de raciocínio de alto nível. Adicionalmente, o foco de ontologias é o compartilhamento de conhecimento (USCHOLD; GRUNINGER, 2004). O conhecimento sobre um domínio pode ser representado em diferentes maneira, na sequência são apresentadas as técnicas mais comuns para organizar o conhecimento, de acordo com o descrito em (SIKOS,2017):

a) **Tesauros** trabalhos de referência que agrupam listas de palavras de acordo com o seu significado;

b) **Taxonomia** palavras categorizadas de maneira hierárquica;

c) **Bases de conhecimento** terminologia sobre um domínio de conhecimento, incluido relacionamento entre as partes que o compõem;

d) **Bancos de dados** coleção de declarações que são interpretadas por máquinas;

e) **Ontologias** conceitualização utilizando um formalismo sobre o conhecimento à respeito de um domínio, com relacionamentos complexos e apropriado para inferir novas declarações de maneira automática;

No entanto, o conhecimento não é limitado e consequentemente pode ser mais abrangente. Por exemplo, o *software* mais simples possível pode ser considerado uma linguagem informal, que expressa conhecimento e relações sobre um determinado domínio, o domínio neste sentido sendo a própria linguagem de programação em que o *software* foi escrito. Evoluindo este conceito um *software* utilizado por uma empresa pode ter um banco de dados com

um particular *schema*. Este *schema* pode ser considerado uma ontologia, apenas não sendo expressa em uma linguagem formal (Adam PEASE, 2011).

Um avanço com relação ao formalismo, em comparação com as linguagens informais, é encontrado nas linguagens de quadro (do inglês *frame languages*). O conjunto de quadros forma uma taxonomia, e cada quadro pode possuir valores específicos que representam um compartimento. Contudo estas linguagens de quadro carecem da possibilidade de expressar regras entre pares de instâncias.

A Lógica descritiva (LD) é construída com base nas linguagens de quadro e com o intuito de introduzir semântica para tais linguagens (Adam PEASE, 2011). LD é um fragmento de lógica de primeira ordem e provê meios para estabelecer relações entre as entidades que compõem o domínio de interesse. Assim, todas os tipos de LD são baseados em um vocabulário contendo nomes de indivíduos, nomes de conceitos e nomes de funções, cuja cardinalidade é binária. O fato das funções expressas em LD utilizarem a cardinalidade binária faz com que uma nova proposição deve ser adicionada para expressar novos relacionamentos. Consequentemente, o número de declarações pode se tornar elevado e inviabilizar a utilização, adicionalmente, tal fator impacta a expressividade da representação de conhecimento utilizando LD (Adam PEASE, 2011).

Com o intuito de evitar tais implicações, é possível utilizar lógica de primeira ordem. Ao aplicar lógica de primeira ordem, o poder de expressividade aumenta, pois a cardinalidade não é limitada às relações binárias. Desta maneira é possível dividir as ontologias de acordo com sua expressividade, conforme o apresentado em (USCHOLD; GRUNINGER, 2004). Esta divisão é apresentada na tabela 1. Na tabela 1 as ontologias são divididas em formais e informais, contudo, considerando-se especialmente as formais, tem-se uma evolução no poder de expressividade da ontologia, partindo da menos expressiva (taxonomia formal), para a mais expressiva (lógica geral) [1].

Além da expressividade, outra característica importante no desenvolvimento de uma ontologia é a especificidade, isto é, o quão específicos são os conceitos representados por uma ontologia. Existem ontologias que se focam em um domínio específico, que são as ontolo- gias de domínio. Em contrapartida, existem as ontologias que se baseiam em uma represen- tação mais abstrata dos conceitos que compõem o domínio, estas são as ontologias gerais ou ontologias superiores. Assim, a utilização de uma ontologia geral como base para estrutu- ração do conhecimento reduz o esforço, visto que uma grande parte dos conceitos necessá- rios pode já ter sido definido, ainda que se tenha uma curva de aprendizado para compreen- der efetivamente tal representação (Adam PEASE, 2011). Vale ressaltar que distintas ontolo- gias podem ser combinadas gerando uma unificação que engloba os conceitos expressos em ambas (V. MASCARDI; LOCORO; P. ROSSO, 2010). Na literatura são encontradas inú- meras ontologias gerais sendo as mais comuns: *Basic Formal Ontology* (BFO), disponível em: `http://ifomis.uni-saarland.de/bfo/` (SMITH, 1998), Cyc, disponível em:

[1]Lógica geral compreendo basicamente, lógica em primeira ordem e lógica de ordem superior.

Tabela 1 – Ontologias divididas pela formalidade .

Formalismo	Ontologia	Categoria
Informal	Termos Glossários Classificação de usuários *Web* dicionários Dicionários de dados	Glossários e dicionários de dados
	Thesauri Glossário estruturado Hierarquias informais	Taxonomias
	XML DTD *Schema* de banco de dados XML *schema* Modelo de dados (UML, STEP)	Metadados
Formal	Taxonomias formais Quadros Lógica descritiva Lógica geral	Ontologias formais e inferências

Fonte: Autor.

`http://www.cyc.com`, DOLCE disponível em: `http://www.loa.istc.cnr.it/old/DOLCE.html`. Uma lista exaustiva e análise das diversas ontologias existentes está além do escopo deste trabalho, contudo pode ser encontrada em (Viviana MASCARDI; CORDÌ; Paolo ROSSO, 2006).

O *Suggested Upper Merged Ontology* (SUMO) foi construído a partir de um conjunto de ontologias gerais, fazendo a junção das ontologias que não possuíam restrições de licença à época da criação (Ian NILES; Adam PEASE, 2001). O SUMO é uma iniciativa do IEEE Standard Upper Ontology Working Group que visa promover interoperacionalidade de dados, busca e recuperação de informação, inferência automática e processamento de linguagem natural (NOY, 2004).

A representação interna do SUMO é construída utilizando-se uma extensão da linguagem *Knowledge Interchange Format* (KIF) (GENESERETH et al., 1992). Assim, a linguagem interna do SUMO é a *Standard Upper Ontology Knowledge Interchange Format* (SUO-KIF). Vale ressaltar que o intuito em se utilizar uma ontologia é a capacidade de realizar inferências que façam uso do conhecimento incorporado e consigam generalizar para situações inicialmente não mapeadas. O SUMO não tem um mecanismo de inferência incorporado, contudo a ontologia do SUMO pode ser representada em *The Thousand of Problems for Theorem Provers* (TPTP), descrito em (SUTCLIFFE, 2010), e consequentemente traduzida no formato que programas, específicos para realizar tal inferência, consigam processar. É digno de nota, que o *software* utilizado neste trabalho é descrito em (SCHULZ, 2013) e denominado E. O E aceita como parâmetros uma base de conhecimento expressa em TPTP, e um conjunto de consultas também expresso em TPTP e faz a verificação se este conjunto de consultas é satisfeito de

acordo com a base de conhecimento. É importante mencionar que o SUO-KIF é um formato mais amigável do que o TPTP, visto que o SUO-KIF utiliza predicados que são construídas em língua inglesa, em contrapartida o TPTP usa símbolos específicos para representar a mesma informação. Porém, é importante enfatizar que a conversão do SUO-KIF para o TPTP tem a cardinalidade **1 : 1**.

O SUMO foi construído baseado em alguns conceitos, estes conceitos são apresentados na sequência, conforme descrito em (Ian NILES; Adam PEASE, 2001):

- *Physical*
  - *Object*
    - *SelfConnectedObject*
      - *ContinuousObject*
      - *CorpuscularObject*
    - *Collection*
  - *Process*
- *Abstract*
  - *SetClass*
    - *Relation*
  - *Proposition*
  - *Quantity*
    - *Number*
    - *PhysicalQuantity*
  - *Attribute*,

contudo tais conceitos básicos já foram evoluídos e consequentemente o SUMO em sua versão atual é muito mais abrangente, conforme o indicado em (Adam PEASE, 2011). Adicionalmente, as definições no SUMO são divididas em módulos, partindo de uma ontologia estrutural que contém as definições mais básicas da linguagem evoluindo para conceitos mais específicos, ainda que compondo uma ontologia superior.

Considerando a construção da ontologia, o SUO-KIF oferece alguns operadores padrão. Tais operadores são: quantificação, termos funcionais, relacionais (no sentido de indicação da existência uma relação), equações e inequações, além de operadores de verdadeiro e falso. Voltando-se para as relações espaciais, o SUMO inclui os seguintes atributos posicionais: *Above, adjacent, below, downhill, downstairs, downstream, downwind, horizontal, inside, left, near, on, outside, right, uphill, upstairs, upstream, upwind* e *vertical*. Por fim, para o estabelecimento de uma relação especial considerando algum destes atributos deve-se utilizar o

predicado *orientation*, além dos objetos que compartilham a relação. Por exemplo, para expressar a seguinte relação: ***John is near San Francisco***, tem-se a seguinte declaração, apresentada na fórmula 2.1, conforme apresentado em (Adam PEASE, 2011).

Fórmula 2.1 – Exemplo de construção de uma relação no SUMO usando SUO-KIF.

```
(orientation John SanFranciscoCalifornia Near)
```

adicionalmente, pode se definir John como instância de Person, conforme o apresentado à seguir:

Fórmula 2.2 – Instanciação de um indivíduo da classe Person.

```
(instance John Person)
```

Como o SUMO é uma ontologia superior, os axiomas existentes não são aplicados para todos os domínios naturalmente, consequentemente são necessárias extensões. As extensões aplicadas neste trabalho são descritas em maiores detalhes na seção 4.2.

O problema da interpretação da imagem é definido pela extração automática da semântica presente em uma imagem. Contudo, a semântica não é representada de forma implícita na imagem, portanto é necessário conhecimento sobre o domínio (I. BLOCH; C. HUDELOT; J. ATIF, 2007). O domínio neste trabalho refere-se as relações espaciais existentes em uma imagem, adicionalmente, a interpretação do que é extraído na análise quantitativa (seção 4.1) é realizada através de uma ontologia, sendo a ontologia adotada o SUMO (Adam PEASE; Ian NILES; J. LI, 2002). O SUMO possui uma ampla cobertura da teoria envolvendo as relações do *Region Connection Calculus* (RCC), além disto possui axiomas expressos em lógica de primeira ordem o que provê recursos para aplicações representando relações mais complexas. É importante ressaltar isto, visto que a maioria das abordagens que trabalham com RCC carecem de expressividade, geralmente tendo que incorporar uma outra linguagem para representar estes axiomas (I. BLOCH; C. HUDELOT; J. ATIF, 2007). Adicionalmente, o SUMO é baseado em uma ontologia cuja construção foi realizada manualmente mapeando toda a WordNet (A. PEASE; I. NILES, 2003), o que permite uma conexão entre o que é expresso em linguagem natural e a própria ontologia. Estes fatores combinados constituem uma base importante para este trabalho, consequentemente a aplicação do SUMO é essencial no método proposto.

2.7 CONCLUSÃO

O BOVW utilizando descritores locais é a técnica mais encontrada na literatura referente a tarefa de recuperação de imagens (GRAUMAN; LEIBE, 2011). Contudo, uma grande parte das abordagens atuais faz uso de redes neurais para extração de atributos que sejam mais pertinentes ao domínio, consequentemente aumentando a qualidade dos resultados obtidos (WAN et al., 2014).

Apesar das redes neurais conseguirem melhores resultados em tarefas de visão computacional, especialmente em classificação de imagens, ainda existe o problema do *semantic*

gap. Este problema persiste, pois a semântica expressa em imagens não pode ser capturada apenas por processos de baixo nível (I. BLOCH; C. HUDELOT; J. ATIF, 2007). Pois, dada a dimensionalidade das imagens é necessário criar uma representação compacta que viabilize o processamento, contudo essa representação não consegue capturar a semântica em toda a sua extensão.

O presente trabalho estabelece relações entre conceitos, baseado em uma ontologia superior (SUMO). Tais conceitos são representados pelos objetos existentes em imagens e estabelecer relações espaciais com o intuito de aumentar a percepção, em termos semânticos, sobre as imagens.

3 TRABALHOS RELACIONADOS

Este trabalho baseia-se nas áreas de: recuperação de imagens e interpretação de imagens. Neste trabalho, a interpretação de imagens é realizada através da construção de uma mereotopologia entre objetos em imagens digitais. Desta maneira, o foco é criar relações entre regiões detectadas em um conjunto de imagens e, em seguida, recuperar imagens através de consultas que expressam uma relação espacial.

3.1 RECUPERAÇÃO DE IMAGENS

A evolução tecnológica recente proporcionou a rápida criação de grandes bancos de imagens a um custo baixo, por exemplo, o banco descrito em (TORRALBA; FERGUS; FREEMAN, 2008) e também o descrito em (RUSSAKOVSKY et al., 2015). Assim, faz-se necessário o desenvolvimento de métodos que sejam capazes de recuperar imagens em bases com milhares de elementos, de maneira rápida e confiável. Dado o elevado custo computacional desta tarefa é necessário criar uma representação compacta de cada elemento existente no conjunto de imagens a ser buscado (XIA et al., 2017). Por um lado, ao se utilizar esta representação compacta, construída apenas com atributos de baixo nível, que baseiam-se em propriedades como luminosidade e textura, tem-se uma representação que não tem poder de expressividade e consequentemente prejudica a comparação entre imagens (XIA et al., 2017). Por outro lado, esta representação esparsa viabiliza a execução de buscas. Em resumo, as técnicas para recuperação de imagens realizam buscas em uma base de imagens procurando similaridades de acordo com uma imagem de consulta (WAN et al., 2014), contudo a compreensão do conteúdo de uma imagem é diferente da representação construída através de atributos de baixo nível (ISLAM et al., 2017). Este problema, de construir uma representação de uma imagem que compreenda a semântica, é conhecido como lacuna semântica (do inglês *semantic gap*) (PARK; JEONG; LEE, 2007).

Uma aplicação de Recuperação de imagens baseada em conteúdo (CBIR), em geral, inclui três componentes: extração de atributos, processamento destes atributos e indexação (H. WANG et al., 2015). Para a extração de atributos, um grande número de abordagens é encontrado na literatura, variando a utilização de atributos globais como o GIST (OLIVA; TORRALBA, 2001), para atributos locais como o SIFT (LOWE, 2004). Trabalhos recentes têm aplicado redes neurais convolucionais visto que através desta técnica é construída uma representação específica para cada tarefa (XIA et al., 2017). No entanto, ainda se encontra muitos trabalhos utilizando as abordagens convencionais. Abordagens convencionais definidas por um algoritmo genérico, geralmente focado em informação local, como SIFT. Pois uma das maneiras de reduzir o impacto do *semantic gap* é através da utilização de atributos locais (ISLAM et al., 2017). Neste contexto, os atributos locais podem ser entendidos como representação de características de partes de objetos. Ao se criar uma representação cujos atributos representam

partes de objetos é possível reduzir o *semantic gap* (HONG et al., 2017). Assim, conclui-se que para reduzir a lacuna semântica, a maneira como a imagem é representada torna-se crucial. Por fim, a tarefa de recuperação de imagens é muito dependente do contexto, o que faz com que não exista um única técnica vigente (WAN et al., 2014).

Uma das principais técnicas na área de recuperação de imagens é o BOW (L. FEI-FEI; PERONA, 2005). Esta técnica originalmente é adotada na busca de texto, em síntese, a aplicação desta abordagem para recuperação de imagens utiliza atributos da imagem de maneira análoga à utilização de palavras em textos. O BOW pode ser dividido nas seguintes etapas: extração de atributos, agrupamento destes atributos e indexação. Através do BOW é possível indexar bases com mais de um milhão de imagens, como por exemplo o trabalho realizado em (PHILBIN; SIVIC; ZISSERMAN, 2008) onde as imagens são atribuídas a um grafo de conexões e depois grupos de classes são criados através de técnicas de agrupamento. Na etapa de consulta apenas um subconjunto das imagens é verificado o que proporciona ganhos em tempo e acurácia. A maioria dos trabalhos existentes na literatura utiliza o SIFT como representação dos descritores de uma imagem (Gabriella CSURKA et al., 2004; PHILBIN; SIVIC; ZISSERMAN, 2008; J. PHILBIN et al., 2007).

A abordagem padrão no processo de agrupamento, considerando o BOW, é a utilização de um método de classificação não supervisionado, em geral o *K-Means*. Outras técnicas também são referenciadas na literatura, como o *Hierarchical K-Means* (HKMeans), ou *Aproximate K-Means* (AKMeans) (J. PHILBIN et al., 2007), *Latent Dirichlet Allocation* (LDA) (LARLUS; JURIE, 2009), *Gaussian Mixture Models* (GMM) (J. PHILBIN et al., 2008; Gabriela CSURKA; PERRONNIN, 2011). Apesar destas alternativas, o método que ainda é mais utilizado continua sendo o *K-Means* devido a sua simplicidade aliada aos bons resultados obtidos. Após a identificação dos grupos de atributos é necessário reduzir a dimensão original das imagens, este processo é obtido ao se calcular o histograma de frequências com relação aos centróides gerados pelo *K-Means*. Isto é, as imagens são projetadas no espaço do *K-Means* e passam a ser representadas através de um vetor com a frequência de cada centróide para a imagem em questão. Um problema da criação do histograma de frequências é que esta etapa possui um elevado custo computacional (TSAI, 2012). Consequentemente, conforme se aumenta o número de grupos, degrada-se o desempenho, assim pode ser utilizada uma aproximação para melhorar o desempenho ao se processar um número elevado de grupos (MUJA; LOWE, 2009). É importante ressaltar que é desejável que o histograma de frequência de atributos para cada imagem seja esparso. Contudo, ao se aumentar o número de grupos, isto é o número de centróides, pode incidir a introdução de ruído (ARANDJELOVIC, 2012).

Em termos práticos, para uma recuperação das imagens mais eficiente, aplica-se um índice invertido (GRAUMAN; LEIBE, 2011; YANG et al., 2007). Este índice consiste em manter uma referência cruzada entre as imagens e os atributos indexados. Esta etapa é essencial, pois faz a junção entre uma representação (atributos) para o objetivo (imagem).

Esta abordagem padrão BOVW sofre com o número de falsos positivos devido a ruídos no processo de vetorização das imagens. Para garantir a consistência do resultado, uma etapa de verificação espacial é realizada. A verificação espacial consiste em comparar cada descritor de uma imagem de consulta com todos os outros de uma imagem candidata. Uma associação é válida caso a distância entre os descritores esteja abaixo de um determinado limiar. Esta etapa é similar ao processo realizado em (PHILBIN; SIVIC; ZISSERMAN, 2008). Adicional- mente, também aplica-se o *Term frequency inverse document frequency* (TFIDF) para melhorar o resultado e reduzir o custo computacional (JÉGOU; DOUZE; SCHMID, 2010; J. SIVIC; A. ZISSERMAN, 2006). Isto é, pondera-se cada atributo de acordo com a relevância que ele possui no conjunto. Assim, atributos que aparecem menos, considerando-se o conjunto total de atributos, tendem a ser mais discriminantes.

Em síntese, ao se aplicar BOVW, as imagens originais são projetadas em um espaço de menor dimensão, permitindo comparações de menor custo computacional. Este espaço vetorial de menor dimensão também pode ser utilizado como base para classificação de acordo com categorias pré-definidas. Neste sentido, utiliza-se largamente SVM, ou *Naïve Bayes* (Gabriella CSURKA et al., 2004) para classificação, principalmente com foco no reconhecimento de cenas (L. FEI-FEI; PERONA, 2005).

Uma outra técnica para recuperação de imagens é a criação de um *hash* semântico, como o proposto em (SALAKHUTDINOV; G. HINTON, 2009), aplicado para imagens (XIA et al., 2017). A proposta destes métodos é criar uma representação que seja compacta, e cuja construção não considere apenas uma única propriedade da imagem. Esta abordagem é uma aproximação da busca dos k vizinhos mais próximos considerando um conjunto de imagens (R. XIA et al., 2014). Algoritmos de *hash* são originalmente uma alternativa para representações em árvore, e sua aplicação na busca visa realizar aproximações em termos de similaridade entre o elemento de consulta e as imagens da base. A consulta através de um *hash* possibilita diminuição dos tempos de busca, porém pode diminuir a precisão da busca (GRAUMAN; LEIBE, 2011). Um dos principais fatores para recuperação de imagens, ainda mais evidenciado ao se tratar de abordagens que utilizam um *hash* semântico, é a métrica de comparação. Em resumo, estes algoritmos criam uma tabela *hash* e as colisões são items similares, assim para uma consulta aplicando a mesma função H para gerar o *hash*, a colisão terá as imagens similares. Na prática, múltiplas tabelas *hash* são usadas, com funções independentes, e a imagem de consulta é comparada em todas as tabelas através da união dos pontos em comum entre a função de *hash* da entrada e das imagens recuperadas. As técnicas de recuperação baseadas em *hash* podem ser vistas como uma especialização da abordagem padrão utilizando TFIDF (SALAKHUTDINOV; G. HINTON, 2009).

3.2 SEGMENTAÇÃO SEMÂNTICA

Neste trabalho o intuito é identificar os objetos presentes em uma imagem, portanto é necessário aplicar algum algoritmo de segmentação. Esta seção revisa os trabalhos recentes na área de segmentação semântica. O intuito da segmentação semântica é atribuir a cada *pixel* da imagem uma classe, a qual é relacionada a um conceito de alto nível. Estes conceitos fa- zem referência, em geral, a objetos existentes no mundo real (RUSSAKOVSKY et al., 2015). Portanto, esta tarefa pode ser entendida como uma classificação supervisionada no domínio da imagem (Gabriela CSURKA; LARLUS; PERRONNIN, 2013; Gabriela CSURKA; PERRONNIN, 2011). Analisando a tarefa por outra perspectiva, este problema é uma etapa para a interpretação de cena. Em resumo, o objetivo da segmentação semântica é identificar cada uma das estruturas que compõem uma cena.

Dada a complexidade da tarefa, isto é, a classificação de cada *pixel* da imagem, muitos autores tem utilizado abordagens combinadas para a detecção dos diversos objetos existentes na imagem. De maneira geral, para a detecção e consequente classificação dos objetos, as distintas regiões compreendidas pelos objetos são identificadas nas imagens e aplica-se um conjunto de técnicas para representar estas regiões. A aplicação de diversas técnicas para representar uma região visa cobrir o maior espectro possível de variações entre os objetos de interesse.

Em aplicações reais, a utilização de apenas um conjunto específico de atributos mostra-se insuficiente para atingir o nível de generalização esperada (ARBELAEZ et al., 2012). Por exemplo, ao se combinar um descritor de textura com um de cor obtém-se melhores resultados do que a utilização destes descritores de maneira isolada. Na literatura são encontrados trabalhos usando vários conjuntos de atributos para obter uma representação da imagem que cubra o maior número possível de variações (ARBELAEZ et al., 2012; Gabriela CSURKA; LARLUS; PERRONNIN, 2013; SANDE et al., 2011; SOCHER et al., 2011). Assim, como na área de classificação de imagens, as abordagens mais recentes estão voltando-se para a aplicação de redes neurais, sendo (LONG; SHELHAMER; DARRELL, 2015) o trabalho pioneiro na área à utilizar redes neurais para segmentação semântica.

Em geral, para se realizar a segmentação semântica, inicialmente é aplicado algum algoritmo de segmentação não supervisionado para extração de regiões candidatas na imagem, usualmente denominadas como *superpixels*. Esta abordagem contudo não é suficiente para a correta detecção de objetos, visto que na maioria dos casos é gerado uma *oversegmentation* (GOULD; X. HE, 2014), pois os *superpixels* são, em geral, menores do que o objeto de interesse (GOULD; X. HE, 2014). Faz-se então necessário consolidar regiões, o que pode ser tratado por meio de modelos de Markov (TIGHE; LAZEBNIK, 2010). Porém, também é possível realizar a fusão destes segmentos de acordo com métricas de distância entre uma classe de objetos ou através de um processo de classificação não supervisionado realizado previamente (SOCHER et al., 2011). Este conjunto de técnicas aplicadas na segmentação semântica permite a decomposição de imagens em múltiplos objetos. Em Zheng et al. (2015) a fusão destas

regiões identificadas na imagem é realizada através de redes neurais recorrentes. Uma das vantagens ao se utilizar redes neurais é que este processo complexo de fazer a fusão de regiões é realizado através do aprendizado do modelo, o que possibilita uma abordagem mais focada para cada tarefa. Adicionalmente, um único modelo que consegue identificar os objetos facilita a experimentação e, consequentemente obtém melhores resultados. Assim, o estado-da-arte na área é a aplicação de redes neurais, como pode ser corroborado pelos resultados obtidos em Zhou et al. (2017), Zhou et al. (2016) e Zheng et al. (2015).

Inspirado em resultados oriundos da área de Processamento de linguagem natural (PLN), trabalhos recentes tem investigado a relação entre os diversos objetos que constituem uma imagem (SOCHER et al., 2011). Por exemplo, em (L.-J. LI; SOCHER; Li FEI-FEI, 2009) são extraídas regiões e classificadas conforme conceitos oriundos da *Word Net* (MILLER, 1995), buscando obter uma representação de alto nível de maneira automática sem interação humana relacionando conceitos com a hierarquia de estruturas visuais existentes na imagem. Avanços na área de interpretação de imagens também buscam estabelecer relações espaciais entre as estruturas encontradas em imagens, tais estruturas sendo usualmente objetos ou parte de objetos (C. HUDELOT; J. ATIF; I. BLOCH, 2008).

3.3 REPRESENTAÇÃO DO CONHECIMENTO

Uma das propostas em aplicações Recuperação de imagens baseada em conteúdo (CBIR) é relacionar atributos de baixo nível oriundo dos *pixels* da imagem com conteúdo semântico. Relações espaciais entre objetos que compõem uma imagem podem revelar informações sobre propriedades desta imagem (MORALES-GONZÁLEZ; GARCÍA-REYES, 2013) e melhorar a eficácia em aplicações de CBIR (C. HERNÁNDEZ-GRACIDAS; L. Enrique SUCAR, 2007). Adicionalmente, a noção do contexto é essencial em aplicações que busquem representar o conhecimento, pois os próprios objetos no mundo real existem em um determinado contexto (Y. LIU, 2010). Este trabalho tem foco em relações espaciais que são apenas um fragmento da semântica existente. Para a representação do contexto visual, focando-se em relações espaciais, de acordo com o apresentado em (C. HERNÁNDEZ-GRACIDAS; L. Enrique SUCAR, 2007) tem-se os seguintes tipos de relação:

a) **Relações topológicas** Se mantém em alterações de rotação, escala e translação. Por exemplo: *sobreposição*, *adjacência* e *contido por*.

b) **Relações de ordem** Preservadas em escala e translação mas se alteram em rotação. Como: *acima-abaixo* e *à direita* ou *à esquerda*.

c) **Relações de medida** Variam em escala mas se mantém em rotação e translação. Exemplo: *10 metros de distância*.

d) **Relações *Fuzzy*** São relações de teor mais vago, como *perto* ou *longe*.

Estas relações não são completas, e em (Y. LIU, 2010) são definidas relações de caráter mais extenso. A tabela 2 apresenta tais relações. Assim, as relações são divididas em cinco tipos: relações semânticas, relações topológicas, relações direcionais, relações de coocorrência e por fim relações que não se caracterizam em nenhuma das categorias anteriores.

Tabela 2 – Relações contextuais ontológicas, adaptado de (Y. LIU, 2010)

Tipo	Nome	Símbolo	Significado	Exemplo	
				c_1	c_2
Relações semânticas	similaridade	$sim(c_1, c_2)$	c_1 é similar à c_2	automóvel	carro
	parte	$part(c_1, c_2)$	c_1 é parte de c_2	roda	carro
	especificação	$spec(c_1, c_2)$	c_1 é especificado por c_2	urso	animal
	combinação	$comb(c_1, c_2)$	combinação de c_1 e c_2	Mar e areia formam praia	
Relações topológicas	adjacência	$adj(c_1, c_2)$	c_1 é adjacente a c_2	rocha	areia
	dentro	$ins(c_1, c_2)$	c_1 esta dentro de c_2	nuvem	céu
Relações direcionais	acima	$ab(c_1,c_2)$	c_1 esta acima de c_2	céu	mar
	abaixo	$bel(c_1, c_2)$	c_1 está abaixo de c_2	mar	sol
	esquerda	$left(c_1,c_2)$	c_1 está à esquerda de c_2		
	direita	$right(c_1,c_2)$	c_1 está à direita de c_2		
Coocorrência	coocorrência	$co(c_1,c_2)$	c_1 coexiste com c_2	areia	mar
Outras	localidade	$loc(c_1,c_2)$	c_1 é a localidade de c_2	campo	barraca
	propriedade	$pr(c_1,c_2)$	c_1 é a propriedade de c_2	maduro	maçã
	exemplo	$ex(c_1,c_2)$	c_1 é um exemplo de c_2		

Obs.:c_1 e c_2 são conceitos de alto nível pertencentes ao domínio de interesse.
Fonte: Autor.

Uma relação espacial é um conceito com propriedades exclusivas e, cuja principal característica é realizar a ligação entre outros conceitos de caráter distintos, os quais são representados por objetos ou parte de objetos (C. HUDELOT; J. ATIF; I. BLOCH, 2008). Em Logan e Sadler (1996) são definidos três tipos de relações espaciais:

a) **básica** expressam uma relação de um objeto com relação ao observador.

b) **deíctica** especificam uma relação entre um objeto (localizado) em termos de outro (referência).

c) **intrínseca** também especificam uma relação entre um objeto (localizado) em termos de outro (referência). Contudo, diferem das relações deícticas pois a relação é com referência a um quadro projetado de maneira intrínseca de acordo com o objeto de referência.

Neste trabalho apenas as relações intrínseca serão consideradas, visto que a base de conhecimento será construída através de imagens digitais. Trabalhos recentes têm investigado as relações entre os distintos objetos presentes em uma imagem, visando traduzir a informação existente em termos de linguagem natural (KARPATHY; JOULIN; Li FEI-FEI, 2014; MALINOWSKI; FRITZ, 2014).

3.4 RECUPERAÇÃO DE IMAGENS ATRAVÉS DA SEMÂNTICA

A área de recuperação de imagens possui uma literatura bastante rica, porém a maioria dos trabalhos utiliza busca por exemplo, isto é, procurando imagens semelhantes à partir de outra que serve de exemplo, ou busca a partir de textos, porém com o enfoque mais abrangente do que relações espaciais. Em resumo, o principal problema da abordagem de busca por exemplo é a diferença entre a representação construída e o conceito esperado. Em contraste, a busca por texto apresenta um alto índice de recuperação, contudo é dependente de custosas anotações na imagem que podem variar de acordo com a consulta de interesse acarretando um problema de escala (C. A. HERNÁNDEZ-GRACIDAS; Luis Enrique SUCAR; MONTES-Y-GÓMEZ, 2013). No entanto, detecção de objetos e estabelecimento de relações é uma maneira de reduzir a lacuna semântica, sem incidir em uma tarefa de pré-processamento custosa, ainda que a anotação dos objetos em uma imagem tenha um custo alto. Esta seção faz uma análise de trabalhos encontrados na literatura que aplicam métodos para reduzir a lacuna semântica fazendo uso do reconhecimento de objetos ou através de detecção relações espaciais.

A maioria das abordagens que utilizam semântica para a recuperação de imagens utiliza a WordNet para construir uma representação formal de maneira léxica. As relações são construídas através de hiperônimos e hipônimos. Contudo, existem outras relações, por exemplo entre partes de objetos (RIAD; ELMINIR; ABD-ELGHANY, 2012). Consequentemente, sistemas de recuperação de imagens baseados apenas na WordNet não são capazes de responder consultas utilizando preposições espaciais (RIAD; ELMINIR; ABD-ELGHANY, 2012).

Em P. L. Lin e Tan (2003) relações topológicas são aplicadas na tarefa de recuperação de imagens. Através da detecção dos objetos existentes na imagem, a representação destes objetos é construída através do *bounding box* do objeto. Contudo, a utilização do *bounding box* pode elevar o número de ruído, pois além do objeto ainda incorpora parte do fundo em alguns cenários (P. L. LIN; TAN, 2003). Adicionalmente, em (P. L. LIN; TAN, 2003) as seguintes relações topológicas são investigadas: *disjoint, meet, overlap, covered by, inside, equal, covers*, e *contains*.

Em Kesorn e Poslad (2009) é construída uma representação formal para recuperar imagens, que expressem atividades esportivas, através de consultas realizadas em linguagem natural. Em Kolla e Venu Gopal (2017) uma representação formal é aplicada com o intuito de refinar os resultados obtidos através da interação dos usuários do sistema.

Em C. A. Hernández-Gracidas, Luis Enrique Sucar e Montes-y-Gómez (2013) é apresentado um arcabouço que engloba segmentação e estabelecimento de relações para recuperação de imagens. Esta abordagem, desconsiderando as particularidades de cada método, é bem similar ao proposto neste trabalho na análise quantitativa, descrita na seção 4.1, porém diferente no trabalho proposto, o método descrito em (C. A. HERNÁNDEZ-GRACIDAS; Luis Enrique SUCAR; MONTES-Y-GÓMEZ, 2013) não faz uma análise qualitativa com o intuito de generalizar novas consultas.

Adicionalmente, na sequência são apresentadas as abordagens que este sistema utilizou como base de comparação. Ainda que utilizando distintas técnicas todas seguem o objetivo proposto em (LAN et al., 2012).

3.4.1 ABORDAGENS DE COMPARAÇÃO

Um fator importante em sistemas de recuperação de imagens envolve a própria consulta, isto é, muitos trabalhos visam reduzir a ambiguidade das consultas com o intuito de eliminar divergências, entre a consulta e a representação da imagem, que possam degradar sistema. Assim, em (LAN et al., 2012) um formato de consulta é proposta para expressar buscam com o intuito de recuperar imagens cujos objetos compartilhem uma relação espacial específica. Por ser um formato estruturado esta representação diminui a ambiguidade da consulta.

O método adotado em (LAN et al., 2012) situa-se na área de aprendizado de relevância, isto é, é proposto um conjunto de treinamento referente a cada consulta, assim através de um classificador estatístico identificar quais são as imagens mais relevantes para uma consulta específica. Porém, este método não faz nenhum tipo de reconhecimento de objetos. Uma abordagem posterior é apresentada em (MALINOWSKI; FRITZ, 2014) onde através de uma média ponderada uma relação espacial é estabelecida, apesar de a detecção de objetos ser essencial, em (MALINOWSKI; FRITZ, 2014) não é aplicado nenhum método de segmentação automático, nos testes realizados são utilizados objetos manualmente identificados. Por fim, o trabalho em (SIDDIQUIE et al., 2014) constrói uma representação dos objetos identificados e faz uma associação entre os distintos modelos que compreendem objetos e relações. Assim, à partir de um conjunto de modelos, construidos através de imagens sintéticas representando uma relação espacial, método descrito em (SIDDIQUIE et al., 2014) associa cada par de objeto identificado à um destes modelos e, consequentemente, estabelece uma relação espacial.

3.5 CONSIDERAÇÕES FINAIS

Nos últimos anos, a área de classificação de imagens e reconhecimento de objetos obteve grandes avanços. Contudo, estas abordagens não levam em consideração as relações existentes entre os objetos que compõem estas imagens. Tais relações fornecem meios para aumentar a compreensão de cena (Jamal ATIF; Céline HUDELOT; Isabelle BLOCH, 2014).

De maneira geral, a utilização de representação formal como representação do conhecimento visa descrever imagens utilizando atributos de nível intermediário oriundos de atributos de baixo nível. Por fim, a representação dos dados contidos em imagens por meio de uma representação formal possibilita a realização de inferências para obter informação semântica de alto nível (L.-J. LI; SOCHER; Li FEI-FEI, 2009).

Neste trabalho o objetivo é relacionar imagens utilizando uma representação de alto nível, consequentemente o emprego de relações modeladas utilizando um formalismo lógico possibilita a criação destas relações entre objetos.

4 MÉTODO

Este capítulo descreve o método aplicado neste trabalho em maiores detalhes. Este método visa fazer a integração de informação de baixo nível com informação de alto nível. A obtenção de informação de baixo nível é obtida através de métodos numéricos que são descritos na seção 4.1. A representação formal utilizada responsável por prover informação de alto nível referente ao domínio é apresentada na seção 4.2. A seção 4.3 apresenta um exemplo detalhado de todas as partes do método. Por fim a seção 4.4 consolida a informação apresentada de maneira sintetizada.

4.1 ANÁLISE QUANTITATIVA

Todos os processos computacionais efetuados que baseiam-se em informação de baixo nível são considerados como parte da análise quantitativa. A figura 6 ilustra os diversos procedimentos adotados neste trabalho. Conforme apresentado na figura 6, é possível dividir o método proposto em três etapas distintas (a, b, c): inicialmente é necessário fazer a segmentação dos objetos presentes na imagem (a), depois identificar as relações topológicas existentes entre os diversos objetos (b) e, por fim, estimar as preposições espaciais existentes entre pares de objetos detectados (c). Estas três etapas são detalhadas nas seguintes seções: a segmentação de objetos na seção 4.1.1, o estabelecimento de relações topológicas na seção 4.1.2 e na seção 4.1.3 será descrita o método adotado para estimação de preposições espaciais.

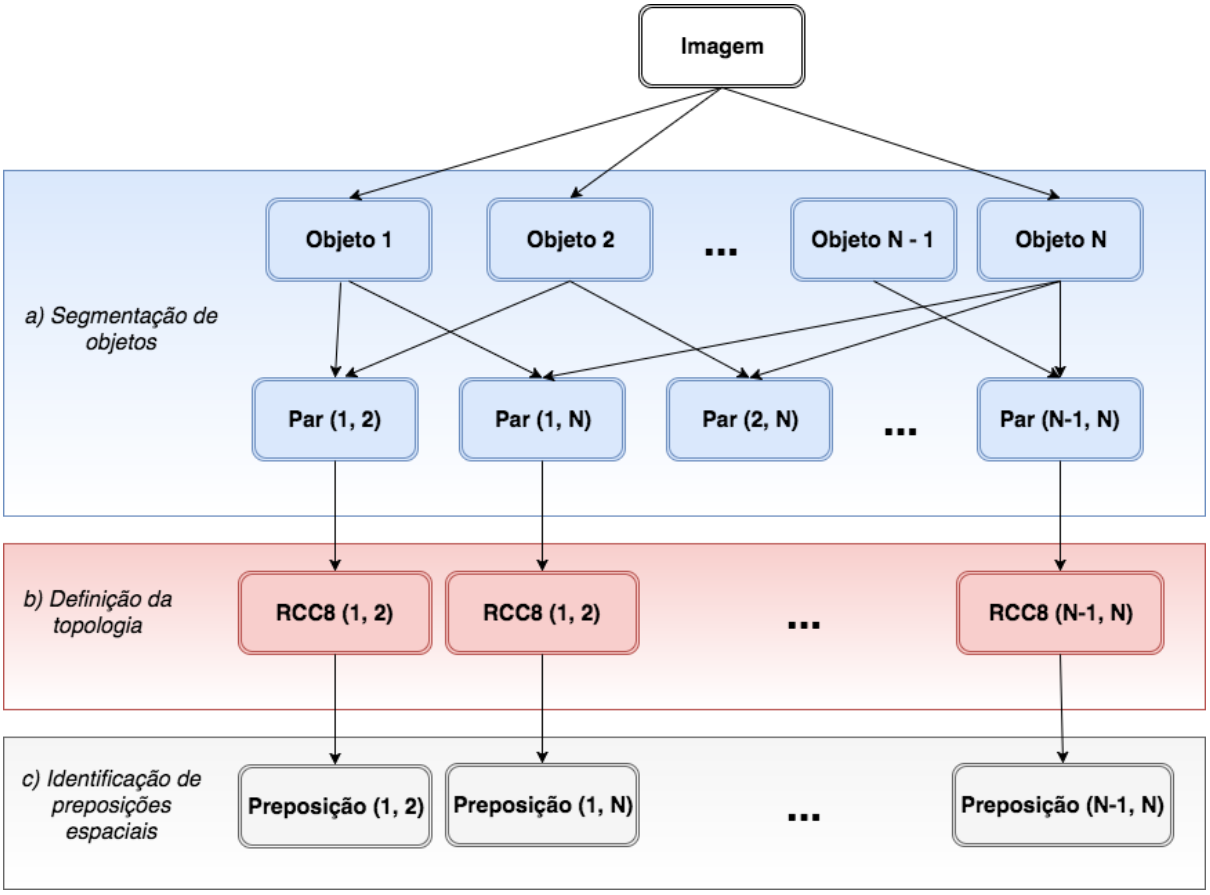
4.1.1 SEGMENTAÇÃO DE OBJETOS

Considerando-se apenas a análise quantitativa, a primeira etapa deste trabalho é identificar os objetos existentes na imagem de uma maneira automática. Este processo é ilustrado na figura 6, através da etapa (a). A técnica aplicada neste trabalho é referenciada como segmentação semântica. Esta abordagem visa classificar cada *pixel* de uma imagem de acordo com um conjunto preestabelecido de classes. Para execução desta tarefa adotou-se um método baseado em redes neurais convolucionais (seção 2.4).

Considerando a rede neural adotada neste trabalho, pode-se separar o processamento de acordo com as diversas camadas que são combinadas. As primeiras camadas, do tipo convolucional, funcionam como se fossem filtros que extraem atributos da imagem. Vale ressaltar que os atributos podem ser globais, ou seja, considerando a imagem como um todo, ou locais, isto é, atributos em locais específicos da imagem. As outras camadas, que são densamente conectadas, contribuem projetando os atributos em um hiperespaço. Por fim, há uma camada que aplica uma função sigmóide e consequentemente define a classe alvo.

O processo de treinamento é iterativo e o conjunto de treinamento é dividido em grupos denominados *batches*. Cada grupo é processado pela rede neural e a saída é comparada com

Figura 6 – Visão geral da etapa de análise quantitativa número do método proposto.



Fonte: Autor.

o esperado. Através de uma função de custo previamente definida é calculada a direção do gradiente para cada camada e os pesos são atualizados conforme uma taxa de aprendizado. Este processo é padrão para o aprendizado de redes neurais e é conhecido como retrocesso (do inglês *backpropagation*).

O processo de segmentação, ou previsão, é o processamento da imagem através da rede neural. A saída da rede será um vetor multidimensional, para cada *pixel* proveniente da imagem, de tamanho igual ao número de classes. Este vetor multidimensional corresponde à probabilidade referente à cada classe. Por fim, para obter a classe correspondente à cada *pixel*, é considerado apenas o valor máximo deste vetor multidimensional.

Para a construção do modelo de segmentação, as imagens foram todas normalizadas em escala de acordo com o tamanho original utilizado em (ZHOU et al., 2016). Adicionalmente, utilizaram-se os pesos disponibilizados por Zhou et al. (2016) para a inicialização do modelo adotado neste trabalho. Todas as imagens tiveram seu tamanho alterado para **(384, 384)** *pixels*, considerando-se a altura e largura, respectivamente. Além disto, utilizou-se uma paleta de cores discriminando cada classe com uma cor específica. A figura 7 ilustra um exemplo da preparação das imagens para o processamento. A figura 7a contém um exemplo de uma imagem original de acordo com o conjunto SUN09. A figura 7b é a versão da imagem da figura 7a respeitando

a paleta de cores pré-definida. O mesmo conjunto de cores é aplicado em todas as imagens, assim o mesmo objeto em imagens distintas terá a mesma cor. Um aspecto digno de nota é que todas as marcações de objetos também tiveram que ser redimensionadas para estarem no mesmo tamanho de $(384, 384)$. Consequentemente, o *pixel* localizado na posição (x, y) na imagem 7a corresponde ao *pixel* localizado na mesma posição (x,y) na imagem 7b. Por fim, todos os objetos que pertencem ao conjunto SUN09 original, mas que não são representados por consultas estruturadas, isto é, não estão listados na tabela 15, tiveram a sua cor definida como zero, representando uma outra classe indicando o fundo das imagens e os objetos que não são de interesse.

Neste trabalho, para realizar a segmentação de objetos utilizou-se uma rede neural com uma arquitetura idêntica à definida em (ZHOU et al., 2016), pois constitui o estado-da-arte em aplicações de segmentação semântica. A arquitetura da rede utilizada, separada por camadas, é apresentada na tabela 3. A rede possui um total de $134.443.370$ parâmetros, incluindo o *bias*. Para o treinamento deste modelo utilizou-se o arcabouço *Caffe* (JIA et al., 2014). O *Caffe* é uma biblioteca escrita na linguagem C++ com enfoque na construção de redes neurais convolucionais.

A tabela 3 contém a nomenclatura indicada para cada camada e o tipo da camada referente. Os tipos de camada fazem referência à função realizada pela camada. Por exemplo, a camada *Data* possui tipo de entrada com dimensões $(3, 384, 384)$, o que significa que esta camada recebe uma imagem em formato RGB de $(384, 384)$ dimensões e cada um dos canais em RGB é fornecido à rede totalizando os $(3, 384, 384)$. Para esta camada não há valores na coluna parâmetros pois ela não possui pesos que sejam atualizáveis, em outras palavras, esta camada apenas alimenta a rede neural sem ser responsável de maneira direta pelo *aprendi- zado*. Adicionalmente, tem-se as camadas convolucionais as quais tem o funcionamento idên- tico entre si, alterando apenas o tamanho do núcleo, isto é, o número de parâmetros que são atualizados e consequentemente *aprendidos*. Por fim, todos os parâmetros que são atualizados pela rede totalizam $134.443.370$. A principal diferença entre a rede adotada e a utilizada por (KRIZHEVSKY; SUTSKEVER; G. E. HINTON, 2012), são as três ultimas camadas apresentada na tabela 3. Visto que, o modelo proposto em (KRIZHEVSKY; SUTSKEVER; G. E. HINTON, 2012) estima uma categoria para a imagem inteira, no modelo utilizado neste trabalho o intuito é estimar uma classe para cada um dos *pixels* que compõem a imagem. Assim, a camada final, *score* de acordo com a tabela 3, possui a probabilidade de cada uma das N classes para cada um dos *pixels*.

A figura 7 consolida todos os passos realizados para segmentação de imagens. Considerando a figura 7a, rotulam-se todos os objetos de interesse de maneira similar ao feito na imagem 7b. O modelo, conforme descrito na tabela 3, atualiza os parâmetros de maneira iterativa, de modo a projetar a imagem original 7a em um hiperespaço que a aproxime mais do objetivo 7b. Esta projeção no hiperespaço atribui para cada *pixel* a probabilidade de pertencer à uma das classes de interesse. Ao se considerar a maior probabilidade para cada *pixel* tem-se

Tabela 3 – Arquitetura da rede utilizada.

Camada	Tipo	Tamanho	Parâmetros
Data	Entrada	**(3, 384, 384)**	
Conv 1.1	Convolucional	**(64, 3, 3, 3)**	1.792
Conv 1.2	Convolucional	**(64, 64, 3, 3)**	36.928
Pool 1	*Pooling* (Pool)	**(64, 192, 192)**	
Conv 2.1	Convolucional	**(128, 64, 3, 3)**	73.856
Conv 2.2	Convolucional	**(128, 128, 3, 3)**	147.584
Pool 2	Pool	**(128, 96, 96)**	
Conv 3.1	Convolucional	**(256, 128, 3, 3)**	295.168
Conv 3.2	Convolucional	**(256, 256, 3, 3)**	590.080
Conv 3.3	Convolucional	**(256, 256, 3, 3)**	590.080
Pool 3	Pool	**(256, 48, 48)**	
Conv 4.1	Convolucional	**(512, 256, 3, 3)**	1.180.160
Conv 4.2	Convolucional	**(512, 512, 3, 3)**	2.359.808
Conv 4.3	Convolucional	**(512, 512, 3, 3)**	2.359.808
Conv 5.1	Convolucional	**(512, 512, 3, 3)**	2.359.808
Conv 5.2	Convolucional	**(512, 512, 3, 3)**	2.359.808
Conv 5.3	Convolucional	**(512, 512, 3, 3)**	2.359.808
fc6	Densa	**(4096, 512, 7, 7)**	102.764.544
fc7	Densa	**(4096,4096,1,1)**	16.781.312
score	Convolucional	**(42, 4096, 4096)**	172.074
upsample	Convolução transposta	**(42, 1, 16, 16)**	10.752
score	Softmax	**(42, 384, 384)**	
Total			134.443.370

Fonte: Autor.

a imagem apresentada na figura 7c. Por fim, a figura 7d sobrepõe ambas as imagens (7a e 7d) para melhor visualização dos objetos com relação à segmentação realizada.

É importante ressaltar que a área de segmentação de imagens é vasta e diversas técnicas podem ser empregadas na solução deste problema. Assim, a escolha da utilização de uma rede neural para realizar a segmentação semântica baseia-se em dois conceitos principais (ZHU et al., 2016):

a)a seleção de atributos é realizada de maneira automática, o que elimina a complexa tarefa de construir extratores de atributos. Neste trabalho, esta característica mostra-se essencial, visto que objetivo principal deste trabalho não é contribuir com novos métodos de segmentação;

b)o bom desempenho das recentes abordagens que utilizam redes neurais para executar a tarefa de segmentação semântica em conjuntos de dados públicos (ZHOU et al., 2016; ZHENG et al., 2015).

Por fim, é digno de nota que todavia não existe uma técnica que possa ser utilizada em todos os problemas com desempenho superior as outras.

Figura 7 – Etapas do processo de segmentação..

(a) Imagem original.

(b) Imagem anotada.

(c) Imagem segmentada.

(d) Imagem segmentada sobre imagem original.

Fonte: Autor.

Este processo é eficiente para identificar os objetos conhecidos que existem em uma imagem. Isto é, os objetos previamente definidos que compõem a base de treinamento. Este trabalho utilizou os objetos apresentados na tabela 15, exibida no anexo. Após realizar a segmentação dos objetos, ainda é necessário relacionar os diferentes objetos detectados conforme a relação espacial apresentada entre eles para interpretar a semântica espacial existente na imagem. Em síntese, esta tarefa visa realizar a segmentação dos objetos à partir de uma imagem de entrada, o resultado da segmentação é uma imagem, com as mesmas dimensões da entrada, e para cada *pixel* é atribuída uma classe do conjunto de interesse.

4.1.2 DETECÇÃO DE RELAÇÕES TOPOLÓGICAS

A tarefa de segmentação ilustrada na seção 4.1.1 consiste na atribuição de uma classe para cada *pixel* existente na imagem. Estas classes representam os distintos objetos de interesse que compõem o domínio. Com o intuito de definir relações espaciais, este trabalho utiliza relações mereotopológicas. É importante mencionar que apenas em termos da mereologia não é possível definir uma representação espacial (CASATI; VARZI, 1996). Pois, não é possível defi-

nir, por exemplo, a relação entre um objeto e sua superfície. Contudo, a topologia complementa a mereologia, isto é, é possível definir a topologia *a priori* e através de primitivas existentes na topologia definir conceitos mereológicos (CASATI; VARZI, 1996). Assim, este procedimento de atribuição de relações com enfoque mereotopológico é definido como etapa (b) na figura 6.

A investigação de formalismos para representar relações de parte-todo existe desde que filósofos começaram a estudar o mundo físico (ARTALE et al., 1996). Para estudar este problema, muitos trabalhos encontrados na literatura investigam relações mereológicas em conjunto com um arcabouço que representa a topologia do espaço em estudo. Consequentemente, a Mereotopologia é a combinação do formalismo de parte e todo realizada pela área da Mereologia em combinação com o formalismo do espaço através da Topologia (ARTALE et al., 1996). Assim, relações mereotopológicas consistem em relações espaciais cujo foco é a análise de parte e do todo entre objetos distribuidos em um espaço topológico. As relações mereotopológicas são as mais estudadas na área de *Qualitative Spatial Reasoning* (QSR) (Anthony G. COHN; RENZ, 2008). Neste trabalho, para representar relações mereotopológicas, é utilizado um subconjunto do *Region Connection Calculus* (RCC). O RCC é um formalismo que define relações sobre pares de regiões espaciais, introduzido em (RANDELL; CUI; A. G COHN, 1992). As relações compreendidas pelo RCC são conjuntamente exaustivas e disjuntas em pares, do inglês *Jointly Exhaustive and Pairwise Disjoint* (JEPD). RCC modela um conjunto de relações mereotopológicas por meio da relação primitiva de conectividade entre pares de regiões. Todas as oito relações básicas definidas pelo RCC são exibidas na figura 8. Dadas regiões espaciais x e y, tem-se que:

a) x está desconectado de y, do inglês *Disconnected* (DC(x, y));

b) x está externamente conectado com y, do inglês *Externally Connected* (EC(x, y));

c) x está parcialmente sobreposto com y, do inglês *Partially Overlapping* (PO(x, y));

d) x é igual à y, do inglês *Equal* (EQ(x, y));

e) x é parte própria tangencial de y, do inglês *Tangential Proper Part* (TPP(x, y));

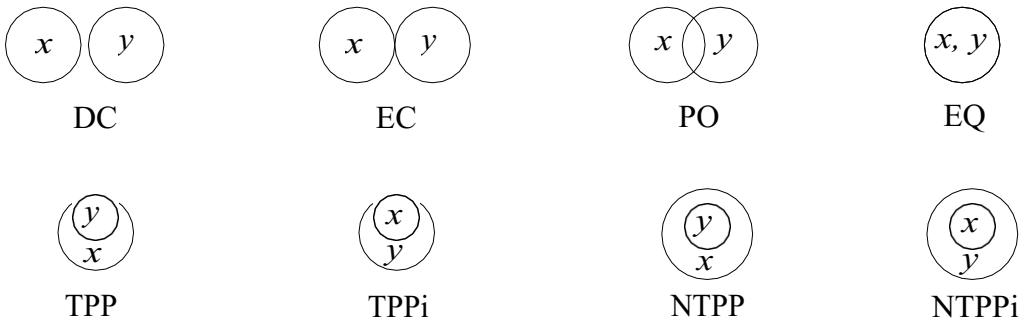
f) x é parte própria tangencial inversa de y, do inglês *Non-tangential Proper Part* (TPPi(x, y));

g) x é parte própria não tangencial de y, do inglês *Tangential Proper Part Inverse* (NTPP(x, y));

h) x é parte própria não tangencial inversa de y, do inglês *Non-tangential Proper Part Inverse* (NTPPi(x, y));

De acordo com a figura 8, considerando duas regiões (x, y), é necessário combinar pares de objetos distintos encontrados na tarefa de segmentação para que seja possível estabelecer as relações topológicas. Assim, para cada elemento I_i identificado em uma imagem I é necessário

Figura 8 – Relações topológicas existentes no formalismo RCC

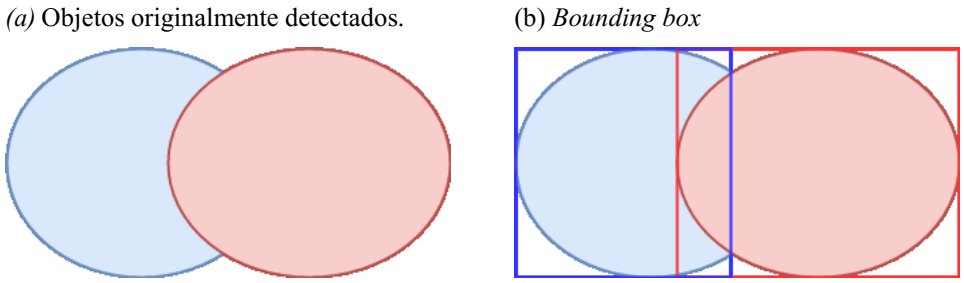


Fonte: Randell, Cui e A. G Cohn (1992).

definir a relação deste com cada um dos elementos encontrados na imagem $I = I_1, I_2, \cdots, I_n$, para as n regiões segmentadas na imagem.

Neste trabalho, a segmentação atribui uma classe referente à um objeto para cada pixel existente na imagem. Em contrapartida o RCC é baseado em conexões entre distintas regiões (ou entre partes de uma mesma região). Assim, as regiões necessárias para definição do RCC são obtidas, de maneira análoga, através das regiões que os objetos ocupam na imagem. Neste contexto, dois objetos se conectam caso estejam sobrepostos, mesmo que parcialmente. Consequentemente, se cada *pixel* é atribuído à uma classe não há sobreposição, portanto não é possível estabelecer a relação topológica.

Figura 9 – Objetos detectados e as respectivas *bounding box*.



Fonte: Autor.

Entretanto, neste trabalho optou-se por expandir a região originalmente detectada visando estimar uma possível sobreposição, isto é, utilizou-se a *bouding box* de cada objeto identificado, conforme ilustrado na figura 9b. Por este fator, ou seja, a utilização do *bounding box*, o próprio objeto (representado por uma região na imagem) nunca é utilizado.

Em resumo, regiões segmentadas são combinadas em pares e, para estimar a relação topológica, utiliza-se uma região limítrofe (*bounding box*), ao invés do próprio objeto detectado. Para cada par combinado, uma relação do RCC é atribuída.

Para estabelecer uma relação RCC adota-se o seguinte procedimento. Considerando duas regiões x e y, é necessário obter a intersecção z entre tais regiões. Assim, para estas regiões (x, y) duas imagens binárias são criadas (I_x, I_y), para x e y respectivamente. Adicionalmente

, uma imagem $I_z = I_x \wedge I_y$ é calculada. Esta imagem I_z contém a intersecção entre as duas regiões segmentadas. Por fim, com base na área (a_x, a_y, a_z) das imagens (I_x, I_y, I_z), isto é, a contagem de valores maiores que zero e, de acordo com um conjunto preestabelecido de regras, atribui-se uma relação do RCC. Estas regras são apresentadas a seguir no algoritmo 1.

Algoritmo 1 – Algoritmo proposto para a atribuição de uma relação RCC

```
1 Entrada: ax, ay,az
2 Saída: Relação RCC
3 se ax = ay = az então
4     retorna EQ(x, y)
5 fim
6  se ax ƒ= az ∧ ay   az ∧ az > θ então
7     retorna PO(x, y)
8 fim
9  se ax ≠ az    ay ≠ az ƒ az ∧  θ então
10    retorna≤EC(x, y)
11 fim
12  se az > 0 ∧ (ax = az ∨ ay = az)    tan(x, y) então
13    se x é igual ao objeto de referência então
14       retorna TPP(x, y)
15    fim
16    senão
17       retorna TPPi(x, y)
18    fim
19 fim
20  se az > 0 ∧ (ax = az ∨ ay = az) ∧ ¬ tan(x, y) então
21    se x é igual ao objeto de referência então
22       retorna NTPP(x, y)
23    fim
24    senão
25       retorna NTPPi(x, y)
26    fim
27 fim
28 senão
29    retorna DC(x, y)
30 fim
```

O algoritmo 1 recebe a área coberta pelos objetos de interesse com respeito às imagens: (a_x, a_y, a_z). Caso a área de $(a_x = a_y = a_z)$ atribui-se a relação EC(x, y). Caso a_x ƒ= a_z e a_z ƒ= a_z existe intersecção. Neste cenário, caso a intersecção seja maior que um limiar θ atribui-se PO(x, y) do contrário atribui-se EC(x, y). Por fim, há a detecção de parte própria, assim $a_z > 0$ e $a_z = a_x$ ou $a_z = a_y$ é detectado parte própria. Considerando x como referência, é feita uma verificação se a região $tan(x, y)$, então é atribui-se TPP(x, y) caso contrário NTPP(x, y). Caso a referência seja y, segue-se o mesmo procedimento, contudo atribui-se NTPP(x, y) ou NTPPi(x, y). Por fim, caso não seja parte própria, atribui-se DC(x, y).

Neste trabalho, é aplicado o procedimento descrito no algoritmo 1 para identificar as possíveis relações mereotopológicas como etapa preliminar à definição de relações espaciais expressas usando preposições.

4.1.3 IDENTIFICAÇÃO DE PREPOSIÇÕES ESPACIAIS

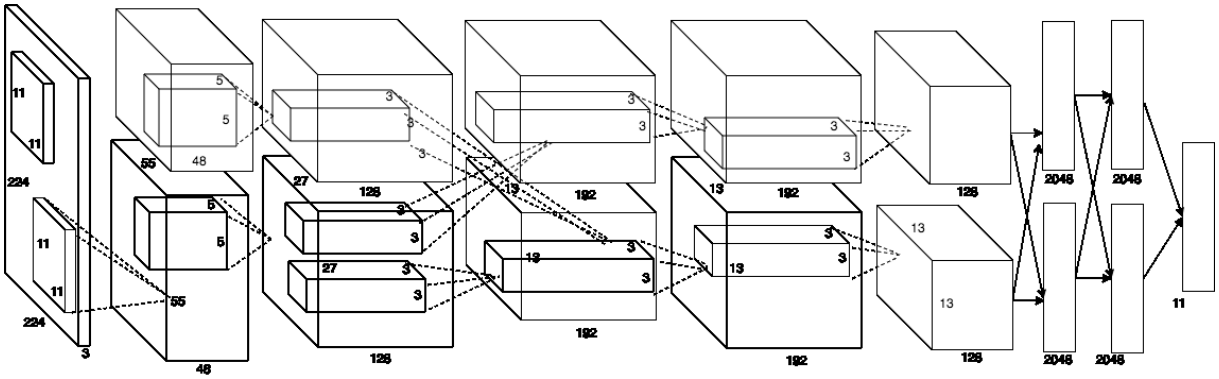
O RCC é fundamentado na conexão entre regiões, as quais são estabelecidas neste trabalho conforme a intersecção entre duas regiões segmentadas em uma imagem. Considerando duas regiões segmentadas, apenas à partir do posicionamento de ambas regiões não é possí- vel identificar se existe sobreposição, conforme é apresentado na figura 9. Contudo, ao incluir informação como, por exemplo, o tipo de objeto que compõe a cena, extraído através de um processo de segmentação, este problema é minimizado. É possível de maneira rudimentar estimar a forma de um objeto conhecido. Tomando novamente como exemplo a figura 9, ao incluir as regiões limites, figura 9b, é possível apenas com a identificação de quais são os ob- jetos que estão apresentados nesta imagem prever qual é o grau de conectividade deste objeto. Adicionalmente, se considerarmos a semântica envolvida na cena, apenas uma aproximação da forma do objeto pode ser suficiente para inferir a relação que ambos objetos possam ter e consequentemente estimar uma preposição. Este trabalho segue este paradigma, isto é, combinar informação a respeito do objeto com a relação topológica para estimar a preposição espacial correspondente ao arranjo espacial. Por fim, considerando-se todo o processo referenciado na seção 4.1 como análise numérica, esta etapa refere-se ao item (c) na figura 6. Em resumo, o objetivo desta etapa, neste trabalho, é estimar preposições espaciais considerando um par de regiões identificadas na imagem. Estas regiões podem ser identificadas de maneira automática, através de um processo de segmentação, ou manualmente, por indicação de especialistas. Adicionalmente, este trabalho investiga se é possível atribuir uma preposição espacial existente entre regiões sem a explícita identificação da topologia existente, isto é, se a topologia expressa pelo RCC é necessária para estimar uma preposição espacial entre dois objetos.

As regiões obtidas através das imagens são mapeadas para um conjunto de relações definido por preposições espaciais através de um método de classificação supervisionado. Um classificador visa mapear um conjunto de atributos com um conjunto de anotações previamente identificados, para posteriormente aplicar as mesmas medidas e prever a anotação em atributos não disponíveis inicialmente.

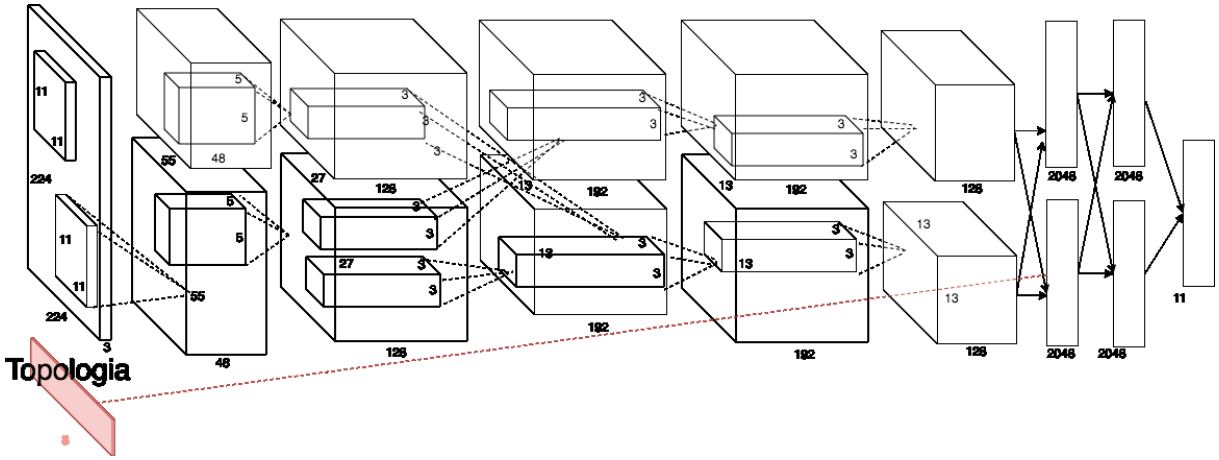
Em síntese, para realizar a estimativa da preposição espacial entre dois objetos, é necessário informação sobre os objetos em questão, além de sua relação topológica. Assim, neste trabalho utilizou-se a detecção realizada na etapa de segmentação como informação dos objetos. A informação topológica é fornecida através do RCC. A segmentação fornece um vetor $O_x = O_1, O_2, \cdots, O_{n_O}$, representando a verossimilhança estimada de cada *pixel* ser um objeto O para os no possíveis objetos, com relação à uma região x identificada na imagem. Este vetor O_x é extraído para cada objeto segmentado, consequentemente, o valor máximo de O_x indica

Figura 10 – Arquiteturas das redes utilizadas para estimar preposições

(a) Arquitetura implícita



(b) Arquitetura explícita



Fonte: Autor. Adaptado de Krizhevsky, Sutskever e G. E. Hinton (2012).

o objeto para considerando-se uma imagem I, assim: $O^I{}_x = max(O_x)$. De maneira similar, o

RCC define um conjunto de relações $R_{x,y} = R_1, R_2, \cdots, R_{n_R}$ para cada par de objetos (x, y). Consequentemente, dados dois objetos detectados (x, y), em uma imagem I é construída uma representação $F_{I,x,y} = (O^I_x, O^I_y, R_{x,y})$, onde O_x é a verossimilhança de cada objeto existente na região x, O_y representa a detecção da região y e $R_{x,y}$ é a relação identificada em termos do RCC. Por fim, o vetor $F_{x,y}$ é utilizado para realizar o treinamento da rede neural, considerando todo o conjunto de objetos detectados.

A figura 10 ilustra ambas as arquiteturas usadas no problema da estimação de preposições espaciais. Cada tipo de camada é colorida de uma maneira específica. A figura 10b possui duas entradas de dados, representadas pelos nós *data* e *topology*. Em contraste, a rede apresentada na figura 10a possui uma camada para entrada de dados representada pelo nó *data*. Em ambos os modelos a entrada através da camada *data* é uma imagem com mapa de cores no formato RGB.

Em resumo, este método aplica uma técnica de segmentação, posteriormente identifica a relação topológica existente entre pares de objetos. Por fim, aplica uma rede neural para estimar

preposições espaciais. Assim, é possível sair de uma representação de baixo nível, isto é, *pixels* para um conceito de mais alto nível, representado por preposições espaciais.

4.1.4 CRIAÇÃO DE ÍNDICE

A representação das imagens descrita na seção 4.1.3 permite a construção de um índice possibilitando a recuperação de imagens através de um conjunto de dados estruturado. Esta estrutura deve expressar um par de objetos de interesse além de uma preposição espacial entre ambos os objetos. Assim, considerando um par de regiões:

$$x, y \in I^a_{R1}, I^a_{R2}, \ldots, I^a_{RN}; a \in I_n,$$

sendo I^a uma imagem proveniente do conjunto de dados, Ri sendo a i-ésima região detectada na imagem I^a de um total de N^a regiões segmentadas; tem-se uma função $R(x, y)$, sendo uma preposição espacial, para R sendo estimada através do classificador descrito na seção 4.1.3. Assim, para uma imagem I^a, tém-se uma combinação tal que $\sum_{2}^{N^a}$, visto que as repetições são consideradas, pois a ordem dos objetos é importante. Adicionalmente, existem relações topológicas inversas, assim como preposições com sentido contrário, portanto ao inverter a order dos objetos na função R, tem-se a alteração dos objetos alvo e referência, portanto um resultado distinto é obtido. Seguindo este paradigma, aplica-se a função R para cada par de regiões detectado na imagem I^a, assim tem-se o conjunto de preposições R^a. Esta estrutura é armazenada de maneira que facilite a recuperação. Assim, é criado um índice para cada objeto que indica qual imagem da base possui aquele objeto, e o mesmo procedimento é aplicado para as preposições.

De maneira análoga ao procedimento apresentado para uma única imagem, este processo é aplicado em todas as imagens existentes na base, resultando em um conjunto composto pelas regiões identificadas na imagem e a preposição estimada entre pares de regiões. Consequentemente, para cada imagem tem-se um conjunto estruturado da seguinte maneira: <*objeto alvo, preposição espacial, objeto de referência*>. Este conjunto é referenciado através de um índice invertido, isto é, uma referência de cada conjunto para a imagem que contém tais regiões existentes no conjunto. Este índice permite a recuperação de imagens ao se consultar a existência dos objetos de interesse. Em síntese, para cada imagem extrai-se um conjunto de atributos que podem ser consultados para verificar quais imagens possuem uma determinada relação espacial expressa na forma de <*objeto, preposição espacial, objeto*>.

Por fim, a relevância de uma imagem para uma determinada consulta é relativa ao número de acertos que possui ao se verificar o índice criado, ponderado pela quantidade de objetos extraídos. Isto é, no momento da consulta, imagens que tiveram mais objetos extraídos são penalizadas pois são mais suscetíveis a ruídos. Um exemplo detalhado deste processo completo é apresentado na seção 4.3

Adicionalmente, as imagens recuperadas são comparadas com o esperado disponível no conjunto de dados adotado. Para realizar esta comparação optou-se por usar o mesmo sistema

de métricas utilizado em (MALINOWSKI; FRITZ, 2014), assim adotou-se o índice de *average precision*, conforme será visto na seção 4.3.

4.2 ANÁLISE QUALITATIVA

Neste trabalho estendeu-se o conjunto de anotações fornecido por (MALINOWSKI; FRITZ, 2014) com o intuito de inferir novas relações derivadas das relações originais, isto é, dado o conjunto de anotações fornecido por (MALINOWSKI; FRITZ, 2014) o intuito foi analisar as propriedades destas relações, como por exemplo a transitividade, através de uma representação formal. Em outras palavras, através do uso de uma representação formal o objetivo é expandir o conhecimento sobre relações espaciais e avaliar novas atribuições de preposições que eram desconhecidas na análise quantitativa.

Em princípio, este trabalho focou-se na transitividade e comutatividade entre relações espaciais. Para avaliar a abordagem proposta foram feitos diferentes tipos de consultas, preservando a semântica original. Para realizar esta tarefa aplicou-se a relação inversa, por exemplo, considerando as preposições *above* e *below*, assim espera-se inverter as relações e recuperar o mesmo conjunto de imagens. Em outras palavras, para dois objetos **(***x, y***)** e a consulta **(***x*, *above*, *y***)**, avaliou-se o retorno para a consulta **(***y*, *below*, *x***)** baseado no mesmo conjunto de anotações para a preposição *above* na consulta original, isto é, manualmente anotada.

No contexto deste trabalho, a construção de fórmulas foi necessária para representar as diferentes relações entre os objetos existentes no conjunto de dados analisados. Considerando, a imagem apresentada na figura 11, tem-se a imagem original 11a e a imagem que passou pela análise quantitativa 11b. Nesta imagem, figura 11b, são destacados três tipos de objetos: *building*, *sky* e *grass*. Vale ressaltar que a segmentação dos objetos deve ser mapeada para o idioma inglês visando facilitar a compatibilidade com a representação formal, cuja construção se baseia na língua inglesa. Portanto, dados estes três objetos, e suas relações topológicas, tem-se:

Tabela 4 – Exemplo de relações topológicas de acordo com uma imagem.

Objeto 1	Objeto 2	RCC
building	*sky*	*Partially overlapping*
building	*grass*	*Partially Overlapping*
sky	*grass*	*Disconnected*

Fonte: Autor.

Desta maneira, é necessário representar cada objeto detectado através da representação formal. Por exemplo, considerando a imagem e os objetos encontrados na figura 11, criou-se as definições apresentadas na fórmula 4.1. Para simplificar a notação, esta imagem é denominada como *A*, assim todos os objetos capturados por esta imagem recebem *A* como sufixo.

Fórmula 4.1 – Instâncias

Figura 11 – Exemplo de imagem segmentada, com três objetos destacados e o fundo negro.

(a) Imagem original.

(b) lmagem segmentada.

Fonte: Autor.

(instance *Building*$_A$ Building).
(instance *Grass*$_A$ Grass).
(instance *Sky*$_A$ Sky).

A fórmula 4.1 instancia predicados que representam que cada objeto detectado pertence a uma classe específica. Além disto, existe uma fórmula que representa a relação espacial extraída através do classificador de preposições, apresentado na seção 4.1.3. Estes objetos, então, são expressos não apenas pelos atributos inerentes a cada um destes, são também considerados de acordo com a relação espacial que compartilham.

O SUMO fornece ferramentas para definir relações espaciais, garantindo definições precisas de cada relação. As relações criadas neste trabalho se baseiam em preposições espaciais, as quais são expressas na linguagem SUO-KIF para utilização em conjunto com o SUMO. A linguagem SUO-KIF é uma linguagem para a construção de axiomas e fórmulas para alimentar o SUMO. Grande parte das relações espaciais existentes com um senso mais abrangente, já está presente no SUMO, e foi inserida por especialistas, como exemplo de tal relação tem-se a relação *Above* [1]. Consequentemente, na maior parte dos casos apenas é necessário indicar quais objetos que compõem a relação, visto que as propriedades da relação já foram definidas. Considerando a base padrão do SUMO, para a definição de uma relação especial é necessá- rio, em geral, criar uma instância da classe *orientation*. A classe *orientation* define relações binárias entre dois objetos (ou neste caso regiões segmentadas na imagem). Considerando a saída do classificador, apresentado na seção 4.1.3, e os objetos ilustrados na figura 11, pode-se estabelecer a relação *Above* entre os objetos *sky* e *building*. Esta relação é expressa na fórmula 4.2. É importante ressaltar que toda base de conhecimento é levada em consideração na hora de se fazer inferências, consequentemente é necessário individualizar as regiões, isto é, utilizar as instâncias específicas correspondentes para cada região segmentada de acordo com a imagem

[1]Existe uma versão pública do SUMO que contém tais relações. O endereço para acesso é: `http://sigma.ontologyportal.org:8080/sigma/Browse.jsp?kb=SUMO`. Último acesso em 05 de Junho de 2017

em questão. Em outras palavras, visando realizar inferências para verificar a existência de uma relação espacial para uma imagem específica, utiliza-se a instância $Building_A$ para estabelecer a relação com outra instância Sky_A, visto que ambas fazem referência à imagem de interesse *A*. Ainda que a região compreendida pela classe *sky*, e consequentemente todas suas instâncias, possa ter a mesma relação *Above* sobre todas as instâncias da classe *Building*, dada a natureza de ambos os objetos, o interesse é apenas no que é interpretado através da análise da imagem *A*.

Fórmula 4.2 – Fórmula referenciando um objeto acima de outro

(orientation Sky_A $Building_A$ Above)

Uma propriedade essencial ao se tratar de espaço é a transitividade de relações. Assim, ao se considerar a figura 11, e a fórmula 4.2, é de grande valia se o sistema fosse capaz de identificar que a mesma relação entre Sky_A e $Building_A$ também se aplica a Sky_A e $Grass_A$, desde que exista a relação expressa na fórmula 4.3.

Fórmula 4.3 – Fórmula referenciando um objeto acima de outro

(orientation $Building_A$ $Grass_A$ Above)

Consequentemente, dadas as fórmulas 4.2 e 4.3, é desejado que a assertiva apresentada na fórmula 4.4 seja verdade.

Fórmula 4.4 – Relação inferida

(orientation Sky_A $Grass_A$ Above)

Essa propriedade é referenciada como transitividade, assim a relação adotada deve ser transitiva. No SUMO a definição de transitividade deve ser incluída na própria definição do axioma *orientation*. A atribuição expressa na fórmula 4.5 garante que a transitividade se estenda para as relações desejadas. A definição deste axioma é apresentada na seção 2.6.

Fórmula 4.5 – Relação transitiva

(instance orientation TransitiveRelation)

É digno de nota que as relações podem ser inseridas explícitamente como o representado nas fórmulas 4.2 e 4.3, ou podem ser inferidas de acordo com o estado da base de conhecimento. Além disso, as relações incluídas no SUMO podem ser variáveis e não constantes no tempo. Apesar deste trabalho lidar com imagens digitais, em cenários reais algumas relações podem variar no tempo. Por exemplo, considerando um carro numa estrada, podemos estabelecer a relação expressa na fórmula 4.6.

Fórmula 4.6 – Relação espacial para um carro numa estrada.

(orientation Car_x $Road_x$ On) .

A fórmula 4.6 representa que há um carro numa estrada [2], contudo, o fato representado por esta fórmula não é atemporal, é preciso então definir que a relação apresentada na fór- mula 4.6 não seja constante no tempo. A fórmula 4.7 corrige este problema, sendo assim mais abrangente do que a fórmula 4.6, ainda que ambas possam estar corretas dependendo do contexto. Neste trabalho optou-se por utilizar a versão dependente do tempo, para tornar o método proposto flexível o bastante para utilização em situações dinâmicas.

Fórmula 4.7 – Relação espacial para um carro numa estrada.

(holdsDuring Now (orientation Car_x $Road_x$ On))

Adicionalmente, é importante mencionar que a semântica com respeito as relações espaciais pode ser expressa por mais de uma preposição, consequentemente, diferentes preposições podem ser similares ou complementares. O SUMO também possui conectivos de implicação **=>** ou dupla implicação **<=>**. Entre inúmeras possíveis operações, estes operadores podem ser usados para representar sinônimos ou antônimos, por exemplo, ao se considerar os antônimos *Under* and *Above*, é necessário estabelecer o seguinte axioma, apresentado na fórmula 4.8.

Fórmula 4.8 – Definição de preposições com sentido oposto.

```
(<=>
        (orientation ?OBJ1 ?OBJ2 Under)
        (orientation ?OBJ2 ?OBJ1 Above))
```

É importante ressaltar que a ordem dos objetos é considerada na fórmula 4.8.

A cardinalidade das relações no SUMO também não é fixa, visto que a construção é realizada em lógica de primeira ordem. Assim é possível ter uma relação ternária, além disso, é possível definir uma relação ternária à partir de um conjunto de relações binárias. Assim, tem-se o seguinte axioma (fórmula 4.9):

Fórmula 4.9 – Relação ternária.

```
(=>
        (and
                (orientation ?X1 ?X2 Behind)
                (orientation ?X2 ?X3 Behind))
        (between ?X1 ?X2 ?X3))
```

Ao se analisar a fórmula 4.9 é possível perceber que o ponto de vista do observador não é levado em consideração. Como este trabalho focou-se em relações intrínsecas (seção 4.1.2), é necessário definir o observador. Assim, pode-se separar o conjunto de preposições adotado neste trabalho em dois tipos: orientação absoluta e orientação relativa. As relações definidas utilizando o axioma *orientation*, padrão no SUMO, são do tipo absoluta. Contudo, a orientação

[2]Para exemplificar e seguir o padrão utilizado anteriormente considerou-se uma imagem hipotética *X*.

relativa não é definida no SUMO padrão. Portanto foi necessário desenvolver uma extensão do SUMO para contemplar estas relações. A fórmula 4.10 define essa relação.

Fórmula 4.10 – Orientação relativa definida em termos do SUMO.

(orientationRelative OBJ1 OBJ2 SpatialRelation Observer)

É digno de nota que neste trabalho o observador é sempre constante, visto que são consideradas apenas as imagens do conjunto da base, as quais são estáticas, apesar do evento que é representado pela imagem ser dinâmico. Em aplicações reais, multi-agente, cada agente deve ser identificado como um observador.

Este conjunto de fórmulas apresentado nesta seção foi aplicado na informação extraída das imagens que compõem a base estudada.

A definição de uma representação formal e a criação do conhecimento necessário para a construção formam o método proposto neste trabalho. Em resumo, através de métodos numéricos, baseados em uma hierarquia de redes neurais, obtém-se informação que é utilizada para construir uma representação formal específica para este domínio, e consequentemente pode- se realizar inferências sobre os axiomas gerados. Na sequência (seção 4.3) é apresentado um exemplo detalhado de todas estas etapas.

4.3 EXEMPLO TRABALHADO

O procedimento completo realizado na análise quantitativa pode ser resumido nas seguintes etapas:

a)Segmentação de objetos;

b)Estimador de RCC;

c)Classificador de preposições;

d)Criação de índice.

Estas etapas são realizadas em sequência. Nesta seção é apresentado um exemplo considerando uma imagem para facilitar o entendimento do método proposto. Assim, para demonstrar o método proposto, a imagem ilustrada na figura 12 é considerada no restante da seção, como sendo a imagem original.

Considerando a imagem original, exibida na figura 12 incialmente, é realizada a etapa (a), isto é, a segmentação semântica. Como resultado deste procedimento, tem-se uma nova imagem gerada, com as mesmas dimensões da imagem original, neste exemplo a imagem apresentada na figura 13, onde para cada *pixel* é atribuído um valor que faz referência à uma classe. O processo de segmentação é realizado por uma rede neural convolucional. Neste exemplo, é assumido que esta tarefa já foi realizada. O treinamento deste modelo é detalhado na seção 4.1.1. Em resumo, na tarefa de segmentação, dada uma imagem de entrada, tem-se uma imagem

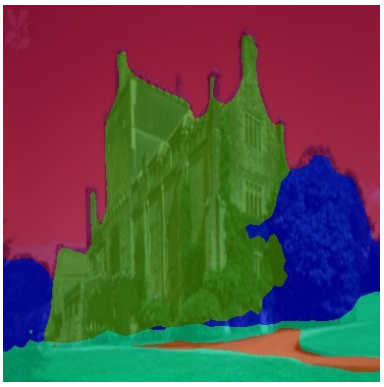
Figura 12 – Imagem de exemplo.

Fonte: Autor.

Figura 13 – Exemplo de imagem segmentada.

(a) Mapa de cores.

(b) Segmentação projetada na imagem original.



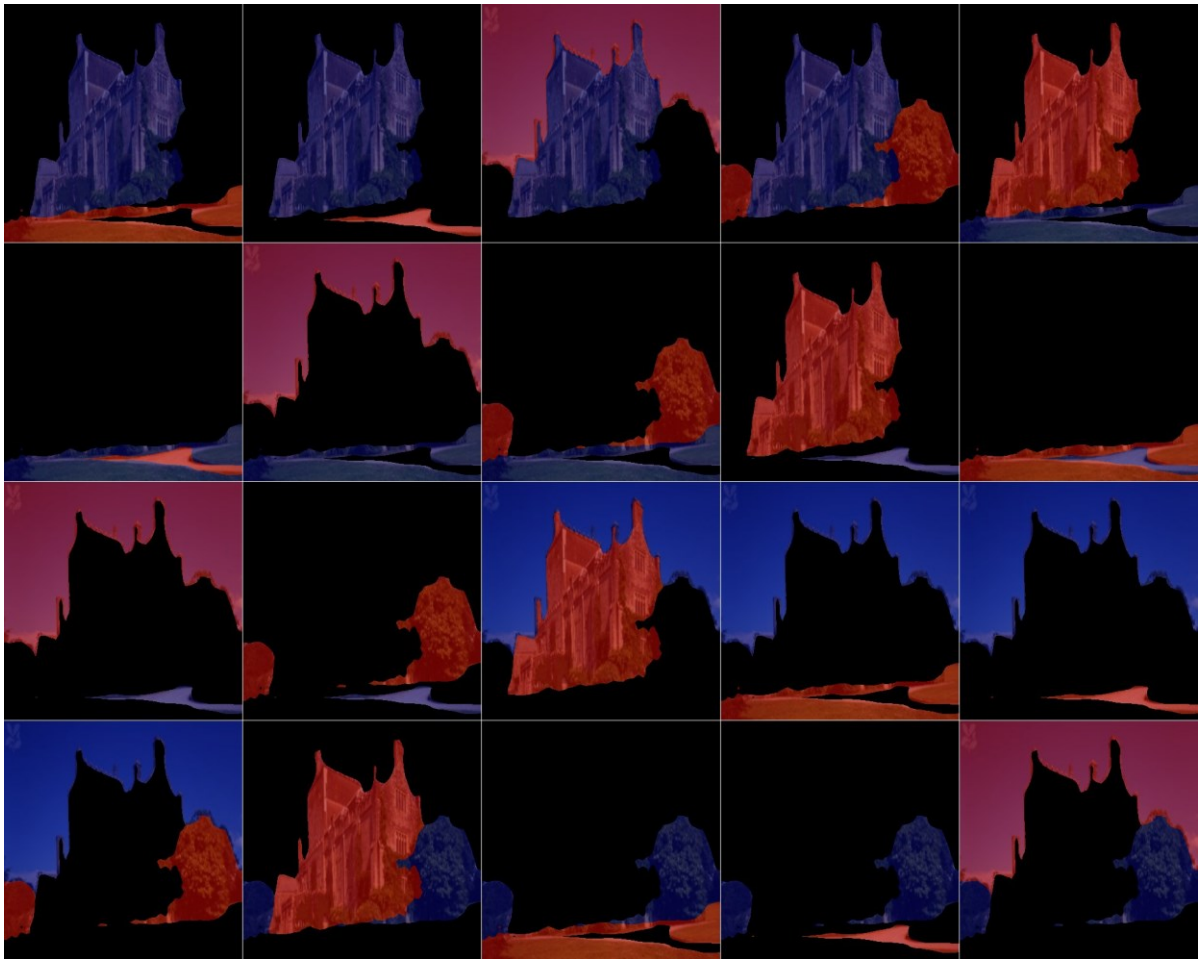
Fonte: Autor.

de saída cujo tamanho é idêntico, e cada *pixel* indica um dos objetos de interesse ou o fundo. Como exemplo desta imagem segmentada tem-se a figura 13. Na figura 13 tem-se a imagem gerada pela rede neural 13a e a imagem segmentada sobreposta na imagem original 13b.

Cada uma das regiões identificadas na imagem 13 por cores distintas tem como base um conjunto de objetos de interesse previamente definido. Apesar da rede neural utilizar informação espacial na construção do modelo de segmentação, não se tem informação de relações espaciais entre as distintas regiões segmentadas no resultado deste processo. Assim, estas regiões são agrupadas em pares através de um arranjo simples. A imagem da figura 13a contém cinco objetos distintos, identificados por cor. Vale ressaltar que objetos que aparecem múltiplas vezes são considerados como um único exemplar, novamente, como exemplo desta condição na figura 13, tem-se o objeto *árvore* [3], onde duas regiões são consideradas como um elemento único. O arranjo $A = \frac{n!}{(n-p)!}$ aplicado para a imagem obtida através da figura 13a, gera $A = \frac{5!}{(5-2)!} = 20$

[3]Identificado pela cor azul, melhor visualizado em cores.

Figura 14 – Objetos combinados referentes à imagem 13.



Fonte: Autor.

regiões combinadas. Consequentemente, são geradas **20** imagens contendo cada item do arranjo gerado. Estas **20** imagens são exibidas na figura 14.

Todas as imagens ilustradas na figura 14 são fornecidas ao classificador com o intuito de estimar as preposições. É importante observar que a ordem dos objetos é importante, de acordo com a sua ordem cada objeto recebe uma cor diferente: azul ou vermelho. Isto é, cada cor indica o objeto alvo (azul) e o objeto de referência (vermelho). Assim, busca-se estimar a relação entre as regiões identificadas na figura 14. Por exemplo, considerando-se um par de regiões x, y e a relação $R_1 = (x, below, y)$, ao inverter a ordem das regiões, tem-se y, x e a relação $R_2 = (y, above, x)$, para *below* sendo a relação inversa de *above*. Por fim, esta informação de $R = R_1, R_2, \cdots, R_N$, é indexada de tal maneira que indique a imagem original no caso a imagem ilustrada na figura 12.

A tarefa de recuperação de imagens consiste em listar as imagens que tenham as regiões (objetos) de interesse e as relações expressas por preposições. Cada imagem terá como atributos uma estrutura representando pares de objetos e, para cada par, uma preposição espacial. Para a recuperação, é criado um índice invertido, isto é, um índice que à partir de um objeto, ou uma

preposição, indica quais imagens possuem o objeto em questão. Consequentemente as imagens que mais aparecem, ponderadas pelo número de elementos que possuem, são as mais relevantes para a consulta.

A inclusão desta imagem na base de conhecimento do SUMO consiste na criação de fórmulas instanciando os objetos detectados conforme o ilustrado na imagem 13 e a representação das preposições encontradas para cada um dos pares ilustrados na figura 14. Adicionalmente, através de um *software* capaz de provar teoremas expressos em lógica de primeira ordem realizar inferências. O *software* adotado é descrito em (SCHULZ, 2013) e denominado *E*. O *E* é capaz de resolver teoremas expressos em *The Thousand of Problems for Theorem Provers* (TPTP) (SUTCLIFFE, 2010). É realizada uma tradução do SUO-KIF para a linguagem TPTP. Vale ressaltar que esta tradução é direta, isto é, as sentenças construídas em SUO-KIF para o SUMO tem uma correspondência direta para o TPTP, assim não há divergência entre ambas representações. Em resumo, dado uma relação específica é realizada a transformação do conhecimento expresso no SUO-KIF para TPTP e realizada a prova através do *E*. Neste contexto, as inferências que são realizadas consistem em verificar se uma imagem possui um determinado par de objetos e se estes objetos apresentam uma relação espacial específica.

4.4 CONCLUSÃO

Neste capítulo os distintos passos do método proposto foram ilustrados. O método divide-se em duas etapas, de acordo com a informação utilizada. A primeira etapa foca-se em análise numérica através de algoritmos. A segunda etapa baseia-se numa extensão da base de conhecimento padrão fornecida pelo SUMO. No capítulo 5 são apresentados os resultados obtidos de testes desta abordagem e a comparação com outras metodologias encontradas na literatura.

5 TESTES E RESULTADOS

Este capítulo detalha os resultados alcançados ao se avaliar o método proposto neste trabalho. Para a realização dos testes utilizou-se um conjunto de dados público conhecido como SUN09 (CHOI et al., 2010). Este conjunto de dados é composto por mais de **12.000** imagens contendo várias classes de objetos em cenas distintas. Para conduzir os experimentos, dois conjuntos de anotações referentes ao SUN09 foram usados. O primeiro destes conjuntos foi proposto no trabalho (LAN et al., 2012) (conjunto 1). Este conjunto de anotações contém marcações ternárias, que são denominadas em (LAN et al., 2012) como consultas estruturadas (do inglês *structured queries*). Estas consultas estruturadas são representadas da seguinte maneira (*substantivo, preposição, substantivo*). As preposições usadas neste conjunto de anotações são: *below* e *above*. Todas estas consultas estruturadas são representadas no idioma inglês. O segundo conjunto de anotações fornecido em (MALINOWSKI; FRITZ, 2014) (conjunto 2) expande o conjunto 1 incluindo um número maior de preposições, representando sete classes distintas: *above*, *across from*, *behind*, *below*, *in*, *in front of*, *inside*, *left*, *right*, *on* e *under*. Assim, o conjunto de dados utilizado na execução dos testes é um subconjunto do SUN09, contendo todas as consultas estruturadas. São utilizadas **4.366** imagens para treinamento e **4.316** imagens para testes.

A utilização de dois conjuntos de dados diferentes pode ser justificada da seguinte maneira: o primeiro conjunto é utilizado para verificar a viabilidade da proposta apresentada enquanto o segundo conjunto possui um número maior de preposições espaciais. Adicionalmente, o primeiro conjunto de dados pode ser considerado mais simples para um classificador cujo intuito seja detectar preposições, visto que a probabilidade de um acerto aleatório é de $\frac{1}{n} = \mathbf{0,5}$, com *n* sendo o número de classes, ou seja, *below* e *above*. Enquanto o segundo é considerado mais complexo, novamente com relação à tarefa de classificação pois a possibilidade de um acerto nestas condições cai de **0.5**, considerando o conjunto um, para $\frac{1}{n} = \mathbf{0,09}$, visto que tem-se 11 classes neste cenário.

5.1 PROCEDIMENTO DE TESTES

Para a realização dos testes foi feita uma avaliação dos módulos individuais e uma avaliação do método completo. Estes procedimentos foram aplicados em ambos os conjuntos de dados estudados. Em resumo, avaliou-se cada módulo utilizado individualmente para minimizar o impacto de ruídos ao módulo e no final avaliou-se o método proposto por completo, fazendo a medição do resultado final. O resultado final do método proposto foi avaliado através de uma tarefa de recuperação de imagens, enquanto que os módulos foram avaliados isoladamente através de métricas de precisão padrão utilizadas para classificadores.

Este trabalho utilizou dois conjuntos de dados. O primeiro disponibilizado em (LAN et al., 2012) e o segundo em (MALINOWSKI; FRITZ, 2014). O primeiro experimento, descrito

na seção 5.2, foca-se principalmente na validação do método proposto no que tange à análise quantitativa, visto que este conjunto de dados inicial é mais simples com relação as preposições consideradas. Adicionalmente, o primeiro conjunto de dados cobre menos possibilidades de análise do espaço, em contraste com o segundo conjunto de dados que contém mais preposições e consequentemente expressa mais características da disposição espacial dos objetos.

5.2 TESTES COM CONJUNTO DE DADOS 1

Conforme o método apresentado no capítulo 4, a primeira etapa deste trabalho é a análise quantitativa. O intuito é analisar se a interpretação de imagens é beneficiada pela identificação de objetos e estabelecimento de relações espaciais entre eles.

Este primeiro teste foca-se na recuperação de imagens que satisfaçam uma determinada consulta. Em Lan et al. (2012) são definidos cinco tipos distintos de consultas que variam da busca por objetos isolados a relações espaciais expressas por preposições. Este teste divide- se em três etapas, a verificação do algoritmo de segmentação (seção 5.2.1, o classificador de preposições (seção 5.2.2) e por fim uma avaliação do método completo na tarefa de recuperação de imagens (seção 5.2.3).

5.2.1 SEGMENTAÇÃO DE OBJETOS REFERENTE AO CONJUNTO DE DADOS 1

O primeiro conjunto de dados (conjunto 1) estudado contém os objetos apresentados na tabela 15, apresentada no anexo. A tabela 15 ilustra a ocorrência de cada objeto no conjunto de imagens. O nome original na língua inglesa é apresentado na coluna denominada Objeto. As próximas duas colunas indicam respectivamente: a quantidade de instâncias (freq.) para cada classe considerando todas as **4.367** imagens do conjunto de treinamento, e a área média percentual (área **%**) que cada objeto ocupa por imagem para o conjunto de treinamento (Treino). As próximas duas colunas contém a mesma informação (freq. e área **%**) porém para o conjunto de teste. As duas colunas seguintes contém a relação entre os valores do conjunto de treinamento com relação ao conjunto de teste (relação Treino / Teste). Por fim, a coluna mais à direita exibe a Intersecção sobre união (IoU) referente ao conjunto de testes obtido ao segmentar as imagens. A IoU é uma medida amplamente utilizada para avaliar a segmentação (RAHMAN; Y. WANG, 2016), cujo intuito é basicamente calcular o percentual em que a segmentação obtida (real) difere de uma segmentação manual (ideal). Para a realização deste cálculo combinam-se as duas regiões (real e ideal), a intersecção é dividida pelo união entre ambas regiões. O número de objetos distintos total é de **111** elementos. No entanto, na tabela 15, no apêndice 6.1, são exibidos apenas **105** objetos, pois algumas classes foram agrupadas a fim de se evitar ambiguidades, por exemplo as classes *cars* e *car* foram consideradas como *car*. Na concepção deste conjunto de dados levou-se em consideração o número de objetos visíveis em cada imagem, portanto em alguns casos existem duas classes para o mesmo objeto uma fazendo referência à objetos que

aparecem no singular e outra com o nome no plural. Por exemplo, considerando o objeto **carro** existem duas classes que o referenciam: *car* e *cars*. Como na hipótese verificada não há, primordialmente, distinção entre a quantidade de elementos para cada classe, optou-se por agrupar tais classes em uma representação comum para simplificar a tarefa do modelo de segmentação adotado. Todas as classes agrupadas estão listadas na tabela 15 com a frequência que aparecem no conjunto de dados (LAN et al., 2012), já considerando este agrupamento realizado.

Vale ressaltar que, neste teste, foram considerados apenas os objetos que tinham consultas estruturadas existentes no conjunto disponibilizado por (LAN et al., 2012). Optou-se por este critério de seleção de objetos, pois o conjunto total de objetos existentes originalmente no conjunto de dados SUN09 possui **6481** elementos. O processamento de todos os **6481** implicaria em um grande dispêndio de recursos computacionais, e, por outro lado, não seria possível a avaliação completa do processo proposto, visto que os objetos remanescentes não possuem anotação.

É importante observar na tabela 15 que a definição de objetos é genérica, no sentido que contém objetos concretos como *building* ao lado de entidades com definições mais vagas como *sky*. Além disto, dada a heterogeneidade na frequência de cada objeto, a tarefa de segmentação se torna particularmente difícil, especialmente em classes que contém poucos exemplos como *headstone*.

De posse dos objetos, o intuito é detectá-los de maneira automática. Este processo foi realizado através da etapa de segmentação, conforme descrito na seção 4.1.1. Com o intuito de verificar a qualidade da segmentação realizada utilizou-se a IoU. Esta medida indica não apenas se o objeto foi identificado, mas também indica se o objeto foi identificado na posição correta e com a região de ocupação do objeto desejada. Objetos que não são identificados são penalizados, pois a intersecção é zero e consequentemente a medida IoU é zero. Em contrapartida, menores inconsistências na forma estimada sofrem menor penalização. A tabela 15 apresenta os valores de IoU para cada objeto do conjunto de interesse. Em síntese, a tarefa de segmentação visa mapear o espaço original na figura 7a para uma representação como a exibida na figura 7b, a diferença entre o real obtido (figura 7c) e o ideal (7b) é computada em termos da IoU. A coluna da direita na tabela 15 exibe a IoU para as distintas classes, visto que quanto melhor a segmentação maior será o valor calculado.

Após a detecção automática dos objetos faz-se necessário estabelecer relações para aumentar o conhecimento sobre o domínio.

5.2.2 ATRIBUIÇÃO DE PREPOSIÇÕES ESPACIAIS

A principal dificuldade ao se atribuir uma preposição de acordo com um par de objetos em uma imagem é a sobreposição natural entre distintas preposições com significados semelhante. Isto é, existem diversas preposições que podem ser atribuídas a cada instante de tempo de maneira coerente.

Tabela 5 – Medidas de precisão e revocação para o modelo estimador de preposição.

Preposição	Precisão (%)	Revocação (%)	Medida F (%)	Quantidade
Above	88	88	88	14364
Below	93	94	94	26829
Consolidado	92	92	92	41193

Fonte: Autor.

Antes de atribuir uma preposição, as relações topológicas são detectadas, conforme descrito na seção 4.1.3. Nos experimentos realizados, cada imagem é representada como um conjunto, cuja estrutura segue o seguinte padrão: (objeto de projeção, objeto de referência, relação *Region Connection Calculus* (RCC)). Como definição dos objetos são utilizados as regiões segmentadas pela rede neural, de acordo com a seção 4.1.1. Adicionalmente, foi utilizado um limiar de **10** *pixels*, definido empiricamente, para definir as relações de sobreposição entre *Externally Connected* (EC**(***x, y***)**) ou *Partially Overlapping* (PO**(***x, y***)**). Ou seja, caso a sobreposição entre dois objetos seja menor ou igual que este limiar, a região é considerada EC**(***x, y***)** nos outros casos atribui-se PO**(***x, y***)**. Esta etapa é importante especialmente ao se considerar o conjunto de dados 1, pois as preposições: *above* e *below* não cobrem todo o espectro de relações espaciais possíveis, portanto é necessário informação topológica para tratar das ambiguidades. Contudo, ao se considerar um número maior de preposições, esta etapa pode eventualmente ser suprimida.

Para estimar as preposições espaciais utilizou-se uma outra rede neural. Como parâmetro para avaliação, adotou-se a precisão e a revocação para avaliar o modelo gerado. Os valores obtidos são apresentadas na tabela 5.

A tabela 5 exibe a estimativa da preposição para todos os possíveis pares existentes na base. Os valores na tabela 5 foram atualizados conforme o conjunto de treinamento, isto é, a segmentação foi realizada através da anotação manual ao invés de se utilizar o modelo cri- ado, de acordo com o descrito na seção 5.2. Adotou-se este procedimento de teste, conforme apresentado na seção 5.1, para a avaliação isolada de cada componente do método. Isto é necessário pois apenas a medição do classificador de preposições é de interesse, neste momento. É digno de nota que esta avaliação considerou como regiões segmentadas as regiões manualmente anotadas, isto é, o padrão ouro. Este procedimento foi aplicado para que erros na detecção de objetos, realizada de maneira automática, não fossem considerados na avaliação do classificador. Contudo, uma avaliação do método completo é apresentada na seção 5.2.3.

Assim, para um total de **41193** pares de objetos combinados existentes no conjunto de teste o classificador alcançou uma precisão de **88%** para a preposição *above* considerando as **14364** amostras da classe *above*. Isto significa que de um total de **14364** itens o classificador previu corretamente cerca de **12571** ocorrências. De maneira similar, considerando-se a classe *below*, a precisão foi de **93%**, o que confere uma taxa de acertos em valores absolutos de **25182** para **26829** possíveis. A revocação foi calculada de maneira similar, isto é de **14364** pares

de objetos cuja relação era *above* o classificador cobriu **12640** casos, totalizando **88%**. Em contrapartida, para a classe *below* a revocação foi de **94%** representando um total de **25219** acertos. Por fim, a coluna medida F (%) pondera ambos os valores, assim pode ser visto como um índice médio entre a precisão e a revocação. Os valores referentes à medida F (%) foram de **88%** e **92%** para as preposições *above* e *below* respectivamente, o valor consolidado da medida F foi de **92%**.

Nesta seção buscou-se avaliar o classificador de preposições considerando o cenário fornecido pelo conjunto de dados 1.

5.2.3 AVALIAÇÃO DA RECUPERAÇÃO DE IMAGENS

Para avaliar a recuperação de imagens, apresentada no seção 4.1.4, as anotações criadas por (LAN et al., 2012) foram usadas.

A base de consultas foi criada seguindo o protocolo proposto neste trabalho. Isto é, para cada imagem os respectivos objetos foram selecionados e combinados em pares. Cada par de objetos foi usado de duas maneiras: na definição da topologia em termos do RCC e como entrada para estimar uma preposição espacial. Toda essa informação, de maneira consolidada, compõe a base alvo, a qual será utilizada para se realizar as buscas de imagem.

Neste trabalho testou-se o método proposto com relação à todas as consultas estruturadas propostas por (LAN et al., 2012). As consultas estruturadas contém um substantivo, por exemplo, *pedestrians*, ou uma relação expressa na forma de uma tupla da seguinte maneira: (substantivo, preposição, substantivo). Por exemplo, *car on road*. Os tipos de consultas estruturadas disponíveis são representados da seguinte maneira:

a) **Structure a (Sa)** contém apenas uma relação. Exemplo: *car on road*;

b) **Structure b (Sb)** contém uma relação e um substantivo. Exemplo: *car on road, pedestrians*;

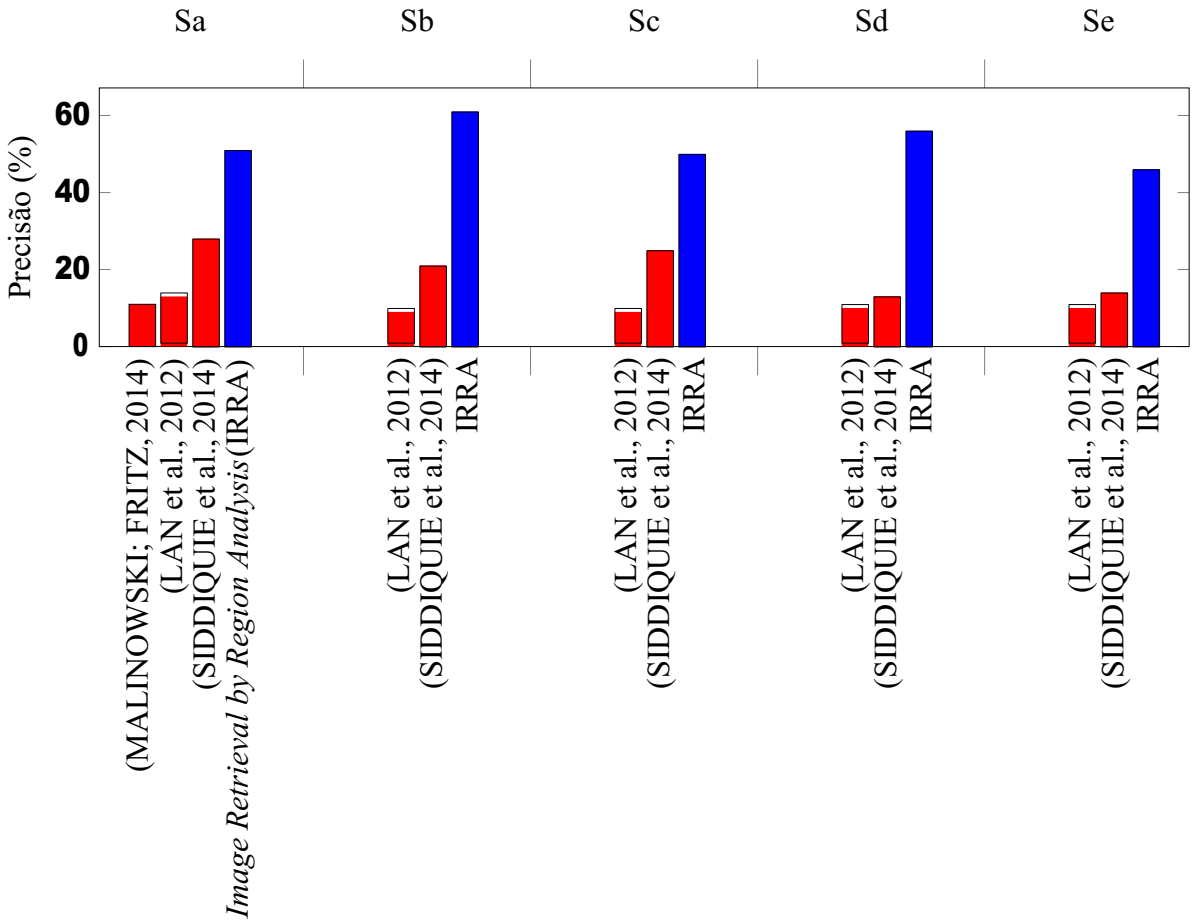
c) **Structure c (Sc)** contém dois conjuntos de relações. Exemplo: *car on road, sky above building*;

d) **Structure d (Sd)** contém dois conjuntos de relações e um substantivo. Exemplo: *car on road, sky above building, pedetrians*;

e) **Structure e (Se)** contém três conjuntos de relações. Exemplo: *car on road, sky above building, books inside of bookcase*.

A figura 15 ilustra as comparações para cada um dos tipos de consultas definidos por (LAN et al., 2012). Esta informação é consolidada na tabela 7. É possível verificar que o método pioneiro de (LAN et al., 2012) obteve **13%** para a precisão, considerando apenas a consulta do tipo **a**. Os métodos desenvolvidos posteriormente: (MALINOWSKI; FRITZ, 2014) e (SIDDIQUIE et al., 2014) alcançaram respectivamente **11%** e **28%**. Enquanto o método IRRA

Figura 15 – Comparação com outras abordagens.



Fonte: Autor.

(proposto nesta tese) alcançou **51%**. Com respeito à consulta do tipo **b**, o método original de (LAN et al., 2012) obteve **9%** e a técnica empregada por (SIDDIQUIE et al., 2014) atingiu **21%**. O método IRRA chegou a **61%**. Vale ressaltar que nos tipos de consulta **b** e **d** foi necessário fazer uso da topologia, além do estimador de preposições, visto que o substantivo foi buscado de acordo com a relação *Disconnected* (DC**(***x, y***)**) com respeito ao RCC. De maneira análoga a consulta **c** para (LAN et al., 2012) obteve **9%**, contra **25%** de (SIDDIQUIE et al., 2014) e **50%** do IRRA. A consulta **d** para (LAN et al., 2012) obteve **10%**, com (SIDDIQUIE et al., 2014) foram **13%** e **56%** para IRRA. Por fim, com respeito à consulta **e**, (LAN et al., 2012) atingiu **10%**, (SIDDIQUIE et al., 2014) obteve **14%** e o método proposto neste trabalho **46%**.

Adicionalmente, na tabela 6 são apresentados os valores obtidos, considerando-se os índices de positivos e negativos.

A tabela 6 exibe os valores de positivos e negativos obtidos pelo método. A coluna consulta ilustra o tipo de consulta que se refere a linha, a coluna VP apresenta os verdadeiros positivos, a coluna FN os falsos negativos, enquanto a coluna FP os falsos positivos e a coluna VN os verdadeiros negativos. Por fim, as colunas P e N apresentam a quantide de positivos e negativos respectivamente.

Tabela 6 – Positivos e negativos na tarefa de recuperação de imagens

Consulta	VP	FN	FP	VN	P	N	Medida F
a	520	41.900	0	1.680.066	520	1.721.966	
c	5.507	188.935	15.934	1.463.293	21.441	1.652.228	
b	803	363.109	0	1.168.649	803	1.531.758	
e	3.673	97.744	18.855	528.867	22.528	626.611	
d	9.536	337.427	5.038	900.683	14.574	1.238.110	

Fonte: Autor.

Tabela 7 – Comparação entre os distintos métodos

Consulta	Método	Precisão
a	Malinowski e Fritz (2014)	**11%**
	Siddiquie et al. (2014)	**28%**
	Lan et al. (2012)	**13%**
	IRRA	**51%**
b	Siddiquie et al. (2014)	**21%**
	Lan et al. (2012)	**9%**
	IRRA	**61%**
c	Siddiquie et al. (2014)	**20%**
	Lan et al. (2012)	**9%**
	IRRA	**50%**
d	Siddiquie et al. (2014)	**13%**
	Lan et al. (2012)	**10%**
	IRRA	**56%**
e	Siddiquie et al. (2014)	**10%**
	Lan et al. (2012)	**14%**
	IRRA	**46%**

Fonte: Autor.

De acordo com a figura 15 é possível observar que o método proposto na presente tese (IRRA) consegue identificar relações entre objetos melhor que os demais utilizados neste estudo. Este teste focou-se na recuperação de imagens utilizando consultas estruturadas, em síntese, consultas buscando imagens através de relações espaciais. Contudo, o conjunto de preposições utilizado neste experimento é limitado, pois conta apenas com um duas preposição. Em cenários mais próximos à ambientes reais espera-se que o número de preposições encontrados seja maior. Assim, uma análise adicional em um conjunto de dados que expressa mais situações entre pares de objetos foi realizada, tal qual apresentado na próxima seção.

5.3 TESTES COM CONJUNTO DE DADOS 2

Nesta seção serão apresentados novos testes utilizando um conjunto de anotações distintos. Serão testadas: a segmentação, a classificação de preposição, a recuperação de imagens, e por fim, a recuperação de imagens utilizando também uma representação formal.

5.3.1 IDENTIFICAÇÃO DE OBJETOS

O conjunto 2 possui **41** objetos distintos que estão compreendidos no SUN09. Como técnica para avaliação da qualidade da segmentação utilizou-se a métrica IoU. A tabela 8 apresenta os valores de IoU obtidos para as classes de interesse.

A tabela 8 apresenta os objetos de interesse com relação ao segundo conjunto de dados. A coluna da esquerda (Objeto) contém o nome dos objetos, as duas próximas colunas (freq. e área % para o conjunto de Treino) ilustram a frequência que cada objeto aparece. Na sequência, tem-se o valor médio da área ocupada pelo objeto com relação ao conjunto de teste (freq. e área % para o conjunto de Teste). As duas seguintes (freq. e área para a relação Treino / Teste), indicam a frequência e a área, contudo o valor apresentado é a relação entre o conjunto de treinamento e o de teste. Por fim a última coluna (IoU%), isto é, a mais à direita indica o valor de IoU obtido para o conjunto de testes.

Estes testes foram estendidos e buscou-se realizar a integração do contexto em qual o objeto está inserido para a melhora da qualidade da segmentação. Estes testes são apresentados na seção 5.3.2

5.3.2 IDENTIFICAÇÃO DE CENAS

Com o intuito de melhorar a qualidade da segmentação de objetos, criou-se dois modelos com arquitetura idêntica ao apresentado na seção 5.3.1. Contudo, dividiu-se o conjunto de objetos em duas classes com relação ao ambiente em que tais objetos aparecem: ambientes internos ou ambientes externos. Consequentemente, cada modelo foi especializado em um tipo destes objetos. A tabela 9 apresenta os resultados obtidos ao aplicar este paradigma.

A tabela 9 apresenta em sua coluna a cena correspondente, ambiente externo ou interno. Cada cena representa um conjunto de objetos que são detalhados na coluna objeto. Por fim, são apresentados dois valores de IoU, representando o modelo da seção 5.3.1 e os modelos que adotam a cena como etapa prévia de préprocessamento. É importante ressaltar que a aplicação desta técnica foi capaz de elevar em cerca de **4%** os valores obtidos de IoU. Adicionalmente, para a detecção da cena que a imagem pertence é necessário um outro processo de classificação. A classificação de cena é apresentada na tabela 10.

Na tabela 10 cada linha corresponde à uma das cenas aplicadas: ambiente interno ou ambiente externo. Na sequência, a coluna mais à esquerda indica a cena correspondente as

Tabela 8 – IoU para as classes de interesse

Objeto	Treino		Teste		Relação Treino / Teste		IoU %
	freq.	área (%)	freq.	área (%)	freq.	área (%)	
Airplane	31	0.05	28	0.03	1.11	1.45	**20,24**
Armchair	79	0.04	77	0.04	1.03	1.13	**23,64**
Balcony	57	0.06	56	0.07	1.02	0.97	**0,00**
Bar	19	0.13	13	0.09	1.46	1.45	**0,00**
Basket	40	0.01	51	0.01	0.78	0.78	**8,12**
Bench	56	0.03	79	0.03	0.71	1.03	**10,83**
Book	37	0.00	48	0.01	0.77	0.65	**0,70**
Bookcase	29	0.19	26	0.22	1.12	0.87	**9,65**
Bottle	46	0.00	44	0.00	1.05	1.27	**4,60**
Box	60	0.01	80	0.02	0.75	0.71	**4,64**
Building	930	0.21	941	0.23	0.99	0.89	**17,86**
Cabinet	68	0.06	77	0.07	0.88	0.84	**0,00**
Car	359	0.01	289	0.01	1.24	0.82	**11,73**
Chair	321	0.02	348	0.03	0.92	0.87	**10,76**
Closet	18	0.15	22	0.18	0.82	0.83	**26,60**
Curtain	139	0.08	165	0.09	0.84	0.86	**33,11**
Desk	67	0.08	89	0.08	0.75	0.94	**13,66**
Door	446	0.03	472	0.03	0.94	0.80	**13,76**
Fence	174	0.04	148	0.04	1.18	1.04	**20,33**
Floor	679	0.24	797	0.24	0.85	0.97	**49,30**
Grass	342	0.11	318	0.12	1.08	0.96	**21,67**
Ground	229	0.15	233	0.15	0.98	1.02	**19,86**
Mirror	96	0.06	107	0.07	0.90	0.93	**35,25**
Path	98	0.06	94	0.09	1.04	0.65	**16,78**
Plant	368	0.04	405	0.04	0.91	1.07	**10,67**
Poster	44	0.02	55	0.02	0.80	1.21	**9,51**
Refrigerator	22	0.08	25	0.09	0.88	0.87	**52,53**
Road	520	0.17	475	0.17	1.09	1.02	**37,45**
Rock	14	0.06	15	0.04	0.93	1.50	**3,28**
Sign	187	0.01	174	0.01	1.07	1.61	**11,94**
Sky	1247	0.27	1132	0.25	1.10	1.08	**71,51**
Streetlight	92	0.00	86	0.00	1.07	1.25	**6,56**
Table	307	0.05	318	0.05	0.97	1.05	**20,74**
Tree	872	0.06	780	0.07	1.12	0.88	**21,11**
Van	68	0.01	69	0.02	0.99	0.60	**14,75**
Vase	82	0.00	70	0.01	1.17	0.74	**9,87**
Wall	877	0.21	959	0.21	0.91	0.99	**22,42**
Water	85	0.21	73	0.22	1.16	0.94	**29,92**
Window	775	0.03	829	0.03	0.93	1.03	**0,00**
Média	256	0.12	258	0.12	0.98	1.00	**16,84**

Fonte: Autor.

Tabela 9 – IoU com a inclusão do contexto

Cena	Objeto	IoU %	
		Contexto	Sem contexto
Ambiente externo	*Airplane*	23,62	20,24
	Armchair	13,70	23,64
	Balcony	1,45	0,00
	Bench	9,67	10,83
	Building	25,02	17,86
	Car	11,39	11,73
	Door	10,95	13,76
	Fence	21,51	20,33
	Gate	0,71	0,00
	Grass	41,60	21,67
	Path	32,05	16,78
	Road	57,26	37,45
	Rock	15,60	3,28
	Sign	10,66	11,94
	Sky	78,73	71,51
	Streetlight	5,72	6.56
	Tree	21,23	21,11
	Van	17,96	14,75
	Water	59,17	29,92
Average		**24,11**	**18,60**
Ambiente interno	*Basket*	5,28	8,12
	Book	1,48	0,70
	Bookcase	27,68	9,65
	Bottle	0,56	4,60
	Cabinet	16,35	0,00
	Chair	5,41	10,76
	Closet	24,35	26,60
	Curtain	31,36	33,11
	Desk	8,48	13,66
	Floor	60,94	49,30
	Ground	17,94	19,86
	Mirror	33,44	35,25
	Plant	15,48	10,67
	Poster	4,48	9,51
	Refrigerator	57,23	52,53
	Table	16,13	20,74
	Vase	1,46	9,87
	Wall	25,18	22,42
	Window	27,82	0,00
Média		**16,57**	**14,07**
Consolidado		**22,66**	**18,67**

Fonte: Autor.

colunas seguintes indicam respectivamente os valores de precisão e revocação correspondentes a cada uma das cenas. A última coluna, mais à direita, corresponde aos exemplos disponíveis para cada cena. A última linha indica o valor consolidado considerando ambas cenas. Assim, o modelo aplicado para classificação das cenas obteve um precisão de **0,87** para uma revocação de **0,87** considerando-se as **4268** imagens existentes. É importante ressaltar que todas estas

Tabela 10 – Precisão e revocação para classificação de cenas.

Cena	Precisão	Revocação	*n*
Ambiente interno	0,86	0,84	1829
Ambiente externo	0,88	0,90	2439
Total	0,87	0,87	4268

imagens foram previamente rotuladas de acordo com a cena correspondente para viabilizar a classificação supervisionada.

A etapa seguinte consiste na atribuição de uma preposição espacial, considerando pares de objetos. Em resumo, cada uma das imagens foi processada através do método de segmentação, e a qualidade da segmentação é apresentada na coluna IoU % na tabela 8. As regiões segmentadas foram combinadas em pares para que fosse possível estimar as preposições espaciais entre pares de regiões, conforme o apresentado na seção 4.1.3.

5.3.3 IDENTIFICAÇÃO DE PREPOSIÇÕES

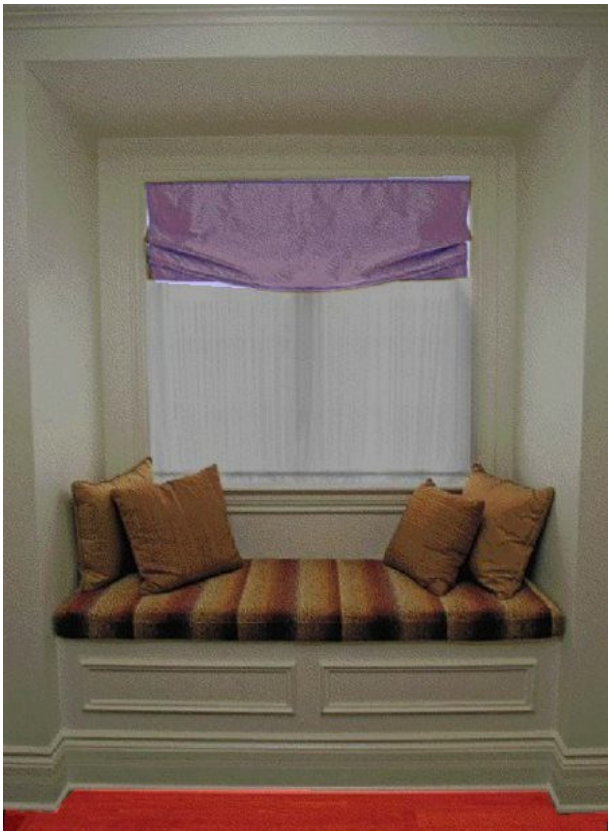
Para a identificação de preposições espaciais existentes entre um par de regiões segmentadas em imagens, duas abordagens distintas foram analisadas para a definição da topologia: explícita e implícita. A abordagem explícita inclui da topologia como parte do conjunto de atributos para a rede neural classificar a preposição. Em contraste, a implícita fornece apenas a própria imagem, como conjunto de atributos, para o classificador.

O classificador empregado utiliza como conjunto de dados uma imagem com pares de objetos destacados. Como exemplo, tem-se a imagem ilustrada na figura 16. A figura 16 contém duas imagens, a imagem original 16a e a imagem com os pares de objetos destacados 16b. Cada objeto teve uma camada de uma cor específica adicionada para indicar o objeto de referência e o objeto alvo. De tal maneira que, se as cores forem invertidas, espera-se que a preposição seja alterada. Visto que essa alteração seria interpretada como a visualização de outro ponto de vista, possivelmente por outro agente. A figura 16b exibe ambos os objetos, de acordo com o apresentado na figura 16b, o objeto azul é o objeto alvo enquanto o objeto vermelho é a referência. Adicionalmente tem-se duas variáveis considerando-se a informação espacial. Da imagem exibida na figura 16 é possível identificar a topologia e pelo menos uma preposição espacial. Consequentemente, a imagem 16b pode ser representada pela preposição *above*, considerando os objetos *curtain* e *floor*, e além disto pode ser referenciada pela relação topológica do RCC: $\mathrm{DC}(x, y)$.

Vale ressaltar que, de acordo com a imagem apresentada na figura 16, pode-se expressar a preposição *above* considerando o objeto de referência *curtain* e o objeto alvo *floor*. Ao se considerar a topologia, claramente ambos os objetos não compartilham nenhuma parte, assim a relação é $\mathrm{DC}(x, y)(curtain, floor)$. As preposições espaciais possuem uma grande sobreposição,

Figura 16 – Objetos de referência e alvo identificados. O objeto azul é a referência enquanto o vermelho é o alvo. Objetos destacados conforme anotação manual existente.

(a) Imagem original



(b) Imagem com pares de objetos destacados . A cortina (azul) é o objeto alvo, o chão (vermelho) o objeto de referência.

Fonte: Autor.

Tabela 11 – Valores de precisão e revocação da rede neural para detecção de preposições. Ambas as abordagens, implícita e explícita, são ilustradas.

Preposição	Implícito			Explícito			*n*
	precisão	revocação	*F-measure*	precisão	revocação	*F-measure*	
Above	0,76	0,58	0,66	0,74	0,46	0,57	166
Across from	1,00	0,03	0,06	0,43	0,03	0,06	387
Behind	0,65	0,62	0,63	0,43	0,65	0,52	329
Below	0,84	0,79	0,81	0,79	0,75	0,77	361
In	0,59	0,84	0,69	0,36	0,78	0,49	475
In front of	0,55	0,69	0,61	0,56	0,30	0,39	317
Inside of	0,00	0,00	0,00	0,00	0,00	0,00	65
Left of	0,01	0,01	0,01	0,09	0,05	0,07	187
On	0,60	0,77	0,67	0,55	0,65	0,60	208
Right of	0,00	0,00	0,00	0,15	0,08	0,11	59
Under	0,87	0,98	0,93	0,92	0,91	0,91	2399
Média	0,75	0,75	0,71	0,69	0,68	0,66	4953

Fonte: Autor.

visto que diferentes conceitos podem ser usados para expressar a disposição espacial alterando o sentido, porém mantendo a semântica coerente.

Voltando o foco para os modelos criados de classificação, utilizou-se duas redes neurais distintas. A arquitetura de ambas as redes pode ser vista na figura 10. As primeiras camadas das redes exibidas na figura 10 são idênticas. Contudo, tem-se a inclusão na informação topológica antes de se realizar a classificação (esta arquitetura é ilustrada na figura 10b). A topologia é expressa em termos de um vetor de oito posições onde todos os valores são zero, exceto a posição correspondente à relação detectada.

As duas técnicas investigadas, isto é, a detecção da topologia de maneira implícita e de maneira explícita, tiveram um comportamento bastante similar. Visto que o modelo adotado em ambos os cenários possui diversas características em comum. A comparação entre estas duas técnicas é apresentada na tabela 11.

Na tabela 11 a coluna mais à esquerda contém todas as preposições utilizadas neste estudo. As três colunas seguintes contém os valores de precisão, revocação e *F-measure* para a rede neural utilizando a representação implícita da topologia, isto é, o modelo ilustrado na figura 10a (arquitetura implícita). As próximas três colunas apresentam os valores para o outro modelo, estando distribuídos de maneira idêntica, ou seja, tem-se a precisão, revocação e *F-measure*. A última coluna mostra a quantidade de exemplos para cada caso de teste. Então, tomando como exemplo a preposição *Above*, havia 166 casos de exemplos. Estes 166 exemplos foram testados utilizando ambas as abordagens de redes neurais, previamente treinadas. Assim, o modelo cuja detecção da topologia era implícito obteve uma precisão de **0,76**, enquanto o valor no modelo explícito ficou em **0,74**. Este valor significa a taxa de acertos considerando os casos estimados como *Above*. O próximo valor de revocação significa o quão abrangente o

Tabela 12 – Matriz de confusão entre as preposições estimadas para o modelo implícito.

		Preposição estimada											
		Above	*Across from*	*Behind*	*Below*	*In*	*In front of*	*Inside of*	*Left of*	*On*	*Right of*	*Under*	Total
Padrão ouro	*Above*	96	0	13	0	4	0	0	0	35	0	18	166
	Across from	0	11	10	4	37	1	0	101	4	0	219	387
	Behind	11	0	204	2	68	2	0	1	24	0	17	329
	Below	0	0	2	286	7	39	0	7	2	0	18	361
	In	2	0	31	4	401	12	0	1	3	0	21	475
	In front of	0	0	6	26	29	218	0	5	2	0	31	317
	Inside of	0	0	5	3	50	3	0	2	0	0	2	65
	Left of	8	0	9	3	36	86	0	1	35	0	9	187
	On	9	0	25	5	8	0	0	0	160	0	1	208
	Right of	0	0	7	2	24	22	0	0	2	0	2	59
	Under	0	0	4	6	16	16	0	0	0	0	2357	2399

Fonte: Autor.

modelo é, por exemplo, com relação aos 166 casos de teste e à preposição *Above*, o primeiro modelo obteve uma revocação de **0,58**, enquanto que o segundo ficou em **0,46**. Estes valores estão definidos em um intervalo entre 0 e 1. Por fim, a medida F pondera os dois valores, precisão e revocação, assim o modelo implícito alcançou **0,66** contra **0,57** do modelo explícito. A média, considerando todas as preposições, é ilustrada na última linha da tabela 11, e o valor **4953** na coluna *n* representa o total de exemplos disponíveis.

Considerando apenas o modelo cujo desempenho foi melhor, isto é, o modelo cuja detecção da topologia é realizada de maneira implícita. Traçou-se uma matriz de confusão. Esta matriz é exibida na tabela 12. É importante ressaltar que apenas este modelo foi utilizado nos testes de recuperação de imagens, descritos na seção 5.3.4.

A tabela 12 ilustra a matriz de confusão representando os resultados da classificação de preposições para as imagens consideradas. Em cada linha da matriz tem-se os valores conhecidos, por exemplo, a primeira linha faz referência à preposição *Above*, somando-se todos os valores desta linha, tem-se **166** elementos, estes elementos são os mesmos ao se considerar a linha correspondente na tabela 11. No entanto, a tabela 12 foca-se no que o classificador previu de maneira errada, indicando em cada linha qual o valor esperado. Assim, considerando o exemplo mencionado para a preposição *Above*, é possível visualizar na tabela 12 que em **35** casos dos **166** o classificador estimou de maneira errada a preposição *On*. Um classificador ideal deveria ter valores apenas na diagonal principal. Seguindo esta abordagem, pode-se observar que as preposições *Inside of*, *Left of* e *Right of* não tiveram o máximo na posição correspondente à diagonal, o que é consistente com o apresentado na tabela 11. Pois, estas preposições tiveram os piores valores considerando a precisão, de acordo com a tabela 11.

O procedimento proposto neste trabalho foi comparado com o método apresentado por Malinowski e Fritz (2014) na tarefa de recuperação de imagens, esta comparação é descrita na seção 5.3.4.

5.3.4 RECUPERAÇÃO DE IMAGENS

De acordo com o procedimento de teste, descrito na seção 5.1, todos os módulos do método proposto foram avaliados em conjunto na tarefa de recuperação de imagens. Assim, aplicou-se o método proposto, IRRA descrito no capítulo 4, neste trabalho, adotando todas as fases, isto é: segmentação de objetos e estimativa de preposição entre pares, na tarefa de recuperação de imagens. Todas as imagens existentes no conjunto de teste foram processadas conforme o método proposto e representadas de maneira que fosse possível a recuperação. Esta representação foi construída utilizando-se uma estrutura de dados em mapa, cujo os valores dos objetos e das preposições indicam uma imagem. Os procedimentos estão descritos em maiores detalhes no capítulo 4. A tarefa de recuperação é simplesmente indicar os objetos e preposições de interesse e fazer a contagem das imagens que mais aparecem, conforme o apresentado na seção 4.1.4. A verificação dos falsos positivos foi realizada comparando com a anotação proposta em Malinowski e Fritz (2014).

A tabela 13 apresenta os valores obtidos na tarefa de recuperação de imagens, considerando as **53** consultas disponibilizadas por (MALINOWSKI; FRITZ, 2014).

Tabela 13 – Métricas da recuperação de imagens

Consulta	Verdadeiros positivos	Falsos positivos	Imagens Recuperadas	Padrão ouro	Precisão média IRRA
airplane in front of building	1	0	1	6	0,17
airplane under sky	15	6	21	24	0,66
armchair in front of wall	12	48	60	44	0,07
armchair right of plant	2	6	8	6	0,12
balcony above balcony	0	63	63	18	0,00
balcony in building	0	18	18	23	0,00
bench in front of wall	2	29	31	34	0,01
bench on floor	5	27	32	30	0,05
bench right of wall	2	58	60	15	0,00
books inside of bookcase	7	8	15	21	0,27
box inside of closet	0	6	6	3	0,33
building across from building	350	1398	1748	364	0,20
building behind wall	20	220	249	84	0,05
building on ground	24	180	204	55	0,10
building under sky	627	771	1398	697	0,42
car across from airplane	0	0	0	5	0,00
car across from cars	0	0	271	8	0,02
car across from van	0	17	17	10	0,00
car under sky	60	170	230	135	0,13
chair left of wall	46	232	278	117	0,07
chair right of tree	1	36	37	5	0,01
cupboard left of refrigerator	9	50	59	10	0,10
curtain above floor	50	152	202	148	0,13
desk on floor	22	28	50	65	0,13

door behind wall	0	23	23	42	0,00
fence in front of tree	10	63	73	37	0,06
fence left of road	15	182	197	26	0,11
floor under sky	37	995	1032	46	0,03
flowers in vase	0	0	0	37	0,00
flowers inside of vase	0	3	3	37	0,00
gate in wall	0	3	3	32	0,01
gate inside of building	0	1	1	11	0,00
grass below window	0	0	0	99	0,00
ground behind road	0	53	53	22	0,00
ground under sky	176	886	1062	216	0,18
mirror in wall	60	158	218	100	0,27
path under sky	38	114	152	77	0,15
plant behind rocks	1	8	9	6	0,03
road below building	91	150	241	302	0,20
road under sky	348	704	1052	379	0,30
rocks under sky	21	369	390	104	0,26
seats on floor	27	42	69	67	0,26
sign left of road	10	116	126	39	0,02
sign right of wall	14	159	173	33	0,09
streetlight under sky	12	15	27	78	0,12
table in front of wall	69	169	238	234	0,19
tree under sky	358	994	1352	434	0,26
wall behind chair	36	93	129	175	0,06
water under sky	52	471	523	256	0,38
window in wall	213	615	828	413	0,27
window under window	84	1508	1592	219	0,03
Mean Average Precision (MAP)					0,13

A tabela 13 apresenta os valores obtidos na tarefa de recuperação de imagens. Estes valores foram contrastados com o esperado. Assim, a coluna da esquerda (consulta) indica a consulta realizada, com os termos originais do inglês. Este conjunto de dados possui **52** consultas. A coluna seguinte (imagens recuperadas) indica o número de imagens recuperadas pelo método IRRA proposto. Na sequência (coluna padrão ouro) é apresentado o número de imagens esperadas, isto é, relevantes conforme a anotação manual. A última coluna (precisão média) indica a precisão considerado todas as imagens retornadas com o padrão ouro.

Considerando este segundo conjunto de dados o IRRA atingiu um valor de *mean average precision* de **13%**, em contraste com **11%** recentemente relatados por (MALINOWSKI; FRITZ, 2014), utilizando o mesmo conjunto de dados. Adicionalmente, o método proposto detecta múltiplas relações por imagem, pois cada relação é proveniente de um arranjo entre as regiões segmentadas, em contraste com o método proposto por (MALINOWSKI; FRITZ, 2014) que estabelece uma relação apenas por imagem. Esta característica aumenta a revocação sem degradar a precisão, como pode ser visto na tabela 13, pois gera mais termos para cada imagem na base aumentando a informação conhecida para cada elemento da base.

Tabela 14 – Consultas por diferentes relações espaciais.

Consulta	Original	Verdadeiros positivos	Recuperadas	Esperadas	*Average precision*
Airplane in front of Building	*Building behind airplane*	4	16	6	0,32
Armchair in front of wall	*Wall behind Armchair*	14	74	44	0,07
Bench in front of wall	*Wall behind bench*	3	27	34	0,02
Books inside of bookcase	*Books in bookcase*	2	4	21	0,05
Building behind wall	*Wall in front of building*	23	155	84	0,05
Curtain above floor	*Floor below curtain*	44	153	148	0,08
Fence in front of tree	*Tree behind fence*	9	96	37	0,02
Floor under sky	*Sky above floor*	25	496	46	0,03
Grass below window	*Window above grass*	14	25	99	0,06
Ground under sky	*Sky above ground*	55	187	216	0,09
Road below building	*Building above road*	11	54	302	0,01
Road under sky	*Sky above road*	186	370	379	0,23
Table in front of wall	*Wall behind table*	98	281	234	0,14
Wall behind chair	*Chair in front of wall*	30	106	175	0,05
Water under sky	*Sky above water*	38	61	256	0,09
Mean Average Precision (MAP)					0,09

Fonte: Autor.

É importante ressaltar que este conjunto de dados apresenta algumas inconsistências, com relação as imagens esperadas para cada consulta. Estes valores podem ser devidos à própria ambiguidade na definição de uma imagem possui uma relação espacial alvo. Contudo, tais inconsistências não inviabilizam a utilização, em termos práticos, penalizam em demasia o sistema sob análise.

5.4 EXPANSÃO DE CONSULTAS POR UMA REPRESENTAÇÃO FORMAL

As recuperação de imagens, realizada na seção 5.3.4, é fixa com respeito à disposição espacial dos distintos objetos que compõem a coleção de dados. Isto é, não há nenhum tipo de processamento para aumentar a revocação, por exemplo tratamento da transitividade entre as preposições. Contudo, investigou-se neste trabalho a aplicação de uma representação formal para conseguir abstrair as relações através de uma base de conhecimento, este procedimento é detalhado na seção 4.2. Portanto, para analisar este problema, neste trabalho, investigou-se a aplicação de uma representação formal expressa em lógica de primeira ordem.

A criação da representação formal considerando o conjunto de imagens SUN09 gerou mais de **13** mil termos referentes às imagens e objetos além de mais de **18** mil fórmulas referenciando as relações criadas. Cada termo é referente à uma instância de uma região com respeito à uma classe de objetos e, de maneira análoga, as fórmulas constituem relações entre estes objetos. A tarefa de recuperação de imagens baseada neste novo paradigma proposto foi realizada avaliando cada imagem através de um *software* com capacidade de provar teoremas em lógica de primeira ordem. O *software* adotado é descrito em (SCHULZ, 2013).

Cada imagem foi verificada através do novo conjunto de dados, o método proposto atingiu um valor MAP de **9%**. Em contraste com o conjunto original, manualmente anotado, que obteve um valor MAP de **13%**. Vale ressaltar que o conjunto de consultas é diferente em ambos os cenários.

A tabela 14 ilustra a recuperação de imagens através da representação formal. Utilizando a mesma base e as fórmulas definidas neste trabalho foi possível através de um provador de teoremas (SCHULZ, 2013) construir inferências de acordo com as consultas originais. Assim, novas consultas foram geradas alterando, em princípio, as preposições originais. Este procedimento construiu **15** novas consultas. A tabela 14 apresenta ambas as consultas: a consulta modificada na coluna à esquerda e a consulta original na segunda coluna. As colunas subsequentes são constituídas conforme as consultas modificadas, sendo na sequência da esquerda para a direita: a quantidade de imagens que foi possível realizar a inferência, quantas imagens eram esperadas no padrão ouro de acordo com a consulta original e por fim a métrica de precisão média. Vale ressaltar que utilizou-se o padrão ouro (MALINOWSKI; FRITZ, 2014), pois espera-se que, apesar de haver alterações na estrutura e na preposição adotada, a semântica se mantenha a mesma.

5.5 CONCLUSÃO

Neste capítulo foram apresentados os resultados obtidos à partir da realização de dois experimentos com base no conjunto de dados SUN09. Para isto dois conjuntos de anotação foram utilizados, o primeiro oriundo do trabalho (LAN et al., 2012) e o segundo proveniente de (MALINOWSKI; FRITZ, 2014).

Nestes experimentos aplicou-se o algoritmo IRRA proposto neste trabalho, ou seja, através de um modelo de segmentação, detectou-se objetos existentes na imagem, os quais foram combinados em pares. De posse destes pares de objetos buscou-se estimar preposições espaciais com o enfoque imediato na recuperação de imagens. Adicionalmente para o segundo conjunto de dados, disponibilizado em (MALINOWSKI; FRITZ, 2014), o conhecimento obtido foi estruturado em uma representação formal onde é possível fazer inferências de acordo com uma determinada consulta para verificar se uma determinada imagem possui os objetos de interesse e, além disso, se estes objetos compartilham uma relação espacial específica, dada por meio de uma preposição espacial.

Por fim, conclui-se que o arcabouço proposto neste trabalho tem um desempenho melhor que as abordagens atuais na recuperação de imagens cujo o foco seja relações espaciais entre os objetos exibidos nas imagens recuperadas

6 DISCUSSÃO E CONSIDERAÇÕES FINAIS

Este trabalho tratou um problema conhecido na área de visão computacional que é a lacuna semântica. Este problema refere-se ao fato de uma representação construída para processamento de imagens não expressar a semântica inerente a uma imagem. A semântica é abstrata e abrangente, então o foco neste trabalho foi apenas em uma área do conhecimento que trata de relações espaciais. Estas relações são, basicamente, relações binárias entre objetos que são exibidos numa imagem. A construção de relações espaciais baseia-se na característica de invariabilidade, isto é, as relações espaciais tendem a variar menos do que os próprios objetos ao se considerar imagens. Por fim, para expressar estas relações optou-se por preposições espaciais. Por um lado a utilização de preposições espaciais visando expressar relações espaciais parece natural, visto que são utilizadas em linguagem natural, por outro lado, inferir preposições considerando-se imagens é uma tarefa difícil pois estas preposições são, em certo grau, abstratas e o sentido varia de acordo com o contexto analisando. Em resumo, este trabalho investigou o problema da lacuna semântica expressando relações espaciais através de preposições espaciais. É digno de nota que as preposições utilizadas encontram-se na língua inglesa para viabilizar a comparação com outros trabalhos com a mesma perspectiva.

Neste trabalho o problema estudado foi decomposto em duas grandes partes: análise quantitativa e análise qualitativa. A análise quantitativa refere-se à aplicação de métodos computacionais para extração de informação, enquanto a análise qualitativa modela esta informação, extraída na etapa quantitativa, de acordo com uma representação formal previamente definida. A análise quantitativa, por sua vez, divide-se nas seguintes áreas: segmentação de objetos e estimativa de relações.

A primeira etapa da análise quantitativa é a segmentação de objetos. A descrição do método é apresentada na seção 4.1.1. A segmentação utilizada, considerando o conjunto de dados 2, obteve uma medida para Intersecção sobre união (IoU) de **16,84%**. O grande problema desta abordagem é a incidência de falsos positivos, os quais degradam esta métrica adotada. A segmentação de objetos é o ponto principal deste trabalho, pois uma segmentação precária incide em erros que são propagados no restante das etapas. Para realizar a segmentação, adotou-se uma abordagem publicada anteriormente que estabelece uma linha base na área (ZHOU et al., 2017), contudo este campo de pesquisa deve apresentar avanços, com relação a qualidade dos objetos detectados e a quantidade de classes de interesse, visto que tem recebido muita atenção recente. Especialmente pelo fato de grandes bases de dados públicos como os descritos em (T.-Y. LIN et al., 2014) e (EVERINGHAM et al., 2010) que fomentam o desenvolvimento da área. Além disto, aplicações de reconhecimento de objetos compõem o portfólio de empre- sas que disponibilizam serviços em nuvem, consequentemente a melhoria da qualidade destes sistemas tende a se traduzir em mais usuários e aumento de receitas.

O primeiro teste realizado foi a recuperação de imagens no conjunto de dados 1 que possui duas preposições espaciais: *above* e *below*, este teste é descrito em maiores detalhes na

seção 5.2.1. Neste teste foi comparado o método proposto com outras abordagens encontradas na literatura. De acordo com os resultados destes testes é possível verificar que ainda há espaço para melhorias. Estas melhorias se referindo a qualidade das imagens que são recuperadas em termos das relações espaciais de interesse. Visto que, um dos principais problemas nesta abordagem é o volume de falsos positivos, uma etapa de pré processamento alinhada com outra de pós processamento pode diminuir a incidência deste tipo de inconsistência. Contudo a melhoria do desempenho desta etapa em específico deverá ser realizada em trabalhos futuros, estando fora do escopo do presente trabalho. Outras abordagens podem aumentar a acurácia destes sistemas de segmentação, por exemplo, a classificação de cenas pode reduzir o número de objetos para cada cena analisada e, consequentemente, melhorar de maneira indireta a segmentação. Visto que, a quantidade de classes de objetos de interesse tende a diminuir. Ainda que a classificação da cena possa propagar erros. Contudo, a tarefa de segmentação é consideravelmente mais complexa. Enquanto para uma imagem de entrada $(n \times m)$ a classificação de k cenas vai resultar em um vetor de $(k \times 1)$ dimensões contendo as probabilidades de cada cena. Em contrapartida, a segmentação de objetos para uma imagem $(n \times m)$ resulta em um matriz de saída $(o \times n \times m)$, para o classes de objetos, sendo $k << o \times n \times m$. Consequentemente, a matriz possui uma entropia muito maior do que o vetor. Além disto, a tarefa de anotação é muito mais custosa, visto que é necessário um rótulo para cada *pixel*, ao invés da classificação que cada imagem possui um rótulo. E por fim, ao se separar a segmentação por cenas diminui- se também o número de classes para cada cena individualmente. Apesar do avanços recentes na área de segmentação poucos trabalhos visam realizar associações entre objetos detectados com o intuito de aplicar raciocínio de alto nível. De tal maneira que a estimativa da relação espacial entre dois objetos foi uma das principais contribuições deste trabalho.

A definição de um conjunto mais restrito de objetos, isto é objetos definidos de acordo com um contexto específico, deve melhorar o sistema em termos quantitativos com relação à medida IoU. Visto que os objetos, neste trabalho, fazem a fusão entre duas áreas de conhecimento: visão computacional e representação de conhecimento. Ambas as áreas abordam a representação de objetos de maneiras distintas. Isto é, em visão computacional um objeto pode ser definido como uma região em uma imagem digital, em contrapartida, para pesquisadores na área de representação de conhecimento um objeto está associado a um conceito. Isto posto, é possível encontrar divergências, visto que ao se construir uma representação de objetos baseada apenas em conceitos, objetos que ocorrem, por exemplo, em ambientes externos compartilham poucas características em comum com objetos em ambientes externos. Isto se amplifica ao considerarmos o espaço, pois uma distância absoluta de 10 metros pode ser perto considerando um ambiente externo, porém longe ao se referenciar um ambiente interno. Em resumo, a segmentação é independente com relação ao objeto de interesse, contudo essa característica não é propagada ao se fazer uma análise qualitativa. Avanços deste trabalho devem investigar esta relação e incluir este tipo de informação na base de conhecimento. Neste trabalho adotou-se a representação fundamentada na visão computacional, porém, com o intuito de reduzir a lacuna

semântica pode se definir os objetos de interesse com relação as características conceituais que tais objetos compartilham, por exemplo se são objetos físicos concretos como carros ou abstrações como o céu.

De acordo com os trabalhos encontrados na literatura para se definir uma relações espacial é necessário saber a topologia com relação aos objetos. Contudo, não há um método específico para definição da topologia. Neste trabalho, investigou-se a utilização do *Region Connection Calculus* (RCC). O intuito desta etapa era estimar uma relação binária entre dois objetos e expressar tal relação em termos de uma preposição espacial. Assim, este problema foi representado como um problema de classificação, de tal maneira que dado uma imagem com dois objetos o classificador expressaria uma preposição. As imagens foram representadas de maneira que se aproximassem da topologia. Baseado nesta representação dois modelos foram construídos, um que utilizava o RCC e outro cuja entrada era apenas tal imagem representando os dois objetos. O modelo utilizando o RCC, denominado como explícito e descrito na seção 4.1.2, não apresentou ganhos em relação ao modelo que utilizou apenas os objetos, denominado implícito. O modelo explícito obteve uma medida *F-measure* de **0,66** em contraste com o modelo implícito que obteve **0,71**. Isto deve-se ao fato da rede neural do modelo implícito conseguir generalizar a topologia de acordo com a ativação das camadas internas. Outro fator que deve ser mencionado é a dimensionalidade, a representação da imagem com os objetos em destaque possui dimensionaldade igual à da imagem original, em contraste, o RCC possui dimensionalidade oito. Este fator diminui a influência do RCC no resultado final. Em resumo, o RCC não apresentou vantagens ao ser utilizado para a estimativa das preposições espaciais. Por fim, a rede neural utilizada nos testes tem a possibilidade de ser melhorada caso se aplique algumas técnicas padrão como o aumento do conjunto de treinamento, um pré processamento das imagens e alterações no próprio modelo. Estas melhorias devem ser investigadas em trabalhos futuros.

Focando-se na recuperação de imagens, apenas alguns trabalhos investigam a recuperação de imagens utilizando preposições espaciais. Assim, seguiu-se o paradigma do trabalho apresentado em (LAN et al., 2012), e aplicou-se um texto estruturado para representar as consultas. É importante ressaltar que esta abordagem apresenta algumas diferenças em relação às consultas em linguagem natural, pois reduz o espaço para interpretação e consequentemente diminui a ambiguidade ao se considerar a consulta. Contudo, mesmo ao se utilizar um texto estruturado, caso o sistema não utilize uma representação formal não é possível abstrair novas relações. Pois, a criação do índice considera apenas a relação estimada pela rede neural, sem realizar abstrações guiadas pelas propriedades da relação em si. Este fato tende a diminuir o *recall*, especialmente ao se considerar sistemas de recuperação de imagens, pois o escopo de consultas é estático. Em contraste, pela representação formal é possível especificar as propriedades de uma determinada relação e, consequentemente, inferir novas relações à partir de um estado específico. Considerando o problema de recuperação de imagens, este fator provê fer-

ramentas para se adaptar às consultas não vistas, pois ao invés de buscas em um índice são realizadas inferências em uma base de dados.

A representação formal utilizada é uma ontologia superior, denominada como *Suggested Upper Merged Ontology* (SUMO). Uma das vantagens ao se utilizar uma ontologia superior, especialmente o SUMO é o fato de que os conceitos mais abstratos já estão previamente definidos, sendo necessário apenas fazer adequações pontuais para diferentes tarefas. No entanto, a construção de uma representação formal ou mesmo uma extensão como a realizada neste trabalho é um processo trabalhoso. Ao se considerar o SUMO em específico é necessário aprender a linguagem *Knowledge Interchange Format* (KIF) para representar os axiomas. Além disso o SUMO não conta com um módulo próprio para realizar inferências, sendo necessário utilizar *software* de terceiros para executar tal procedimento. Duas representações podem ser modeladas de tal maneira que expressem os mesmos conceitos, contudo de maneiras diferente, o que dificulta a integração de ambas, adicionalmente, a modelagem de uma representação formal segue decisões que podem ser específicas para um contexto, agravando este problema.

A proposta deste trabalho foi fazer a integração de informação de baixo nível, extraída de imagens, com informação de alto nível expressa por uma representação formal. Assim, após todo o processamento realizado, buscou-se utilizar ambas fontes de informação, para isto adotou-se o SUMO. De acordo com o cenário em que este trabalho está inserido validou-se a extensão do SUMO, apresentada neste trabalho, para realizar a recuperação de imagens. A hipótese investigada era se através da representação formal o sistema seria capaz de interpretar o conhecimento extraído da imagem e inferir novas relações baseando-se em propriedades expressas nesta representação formal. Assim, os resultados obtidos nos testes realizados foram de acordo com o esperado, visto que o sistema foi capaz de abstrair relações que não foram diretamente extraídas de maneira automática pois constituíam relações derivadas. Na execução dos testes as relações consideradas envolviam a propriedade da transitividade. Além disto, a abordagem proposta para extração de informação de baixo nível é limitada à relações binárias. No entanto, a topologia em um ambiente real é muito mais complexa e dinâmica sofrendo com variações inclusive devido ao ponto de vista dos observadores. O problema ao se utilizar uma representação formal é que a verificação da correção é computacionalmente complexa, assim, ao se aumentar gradativamente o tamanho da base de conhecimento é possível a inclusão de erros, e, adicionalmente, a detecção de tais erros se torna cada vez mais difícil de ser realizada, pois este é um problema NP (Adam PEASE; Ian NILES; J. LI, 2002). Dado a complexidade de validação dos axiomas, verificar se uma imagem possui uma determinada relação espacial, como o realizado neste trabalho, é uma operação custosa e consequentemente demanda um tempo que pode dificultar a aplicação onde este seja um requisito importante, fazendo com o que as possíveis consultas tenham que ser processadas previamente o que impacta no desempenho do sistema, acarretando um problema na inicialização, visto que as possíveis consultas podem ser desconhecidas *a priori*.

Como trabalhos futuros propõe-se a aplicação deste método em um sistema contendo multiagentes, onde cada agente possa fazer inferências a respeito de sua localização, e consequentemente do que é percebido à partir deste ponto, combinado em um arcabouço único, descentralizado, onde contenha informação de outros agentes localizados em pontos distintos. Assim, saindo de um domínio visual, para uma base de conhecimentos compartilhada entre os distintos agentes.

6.1 CONCLUSÃO

Este trabalho teve como foco a construção de relações espaciais através de imagens, de maneira *bottom-up*. Isto é, partindo dos dados de baixo nível, referentes à intensidade dos *pixels*, os distintos objetos existentes em uma imagem são identificados e combinados em pares para que se possa definir uma relação espacial. Estas relações espaciais são utilizadas para estender uma base de conhecimento. Assim, o sistema é capaz de abstrair novas relações não vistas anteriormente e adicionalmente expressar relações cuja cardinalidade seja diferente das relações binárias originais. Por fim, para evidenciar a proposta foram realizados testes cujo foco foi a recuperação de imagens.

Na área de recuperação de imagens existe um problema comum conhecido como lacuna semântica (do inglês *semantic gap* que é a diferença entre um conceito e a informação existente na imagem. Este problema ocorre pois a semântica pode ter um sentido mais amplo que a informação disponível, expressa através dos *pixels*, e consequentemente há uma distorção entre um determinado objeto expresso em uma imagem digital e o conceito que este objeto pode representar. De tal maneira que a tarefa de recuperação de imagens possa ser vista como a maximização da similaridade entre um conceito, expresso através de uma consulta, e uma determinada imagem. Quanto mais abrangente for o conceito de interesse mais ambígua é a informação consequentemente mais ruidosa será a recuperação. Este trabalho segue o paradigma adotado em algumas aplicações da área de recuperação de imagens e os passos para o mapeamento de conceitos para as imagens são executados na seguinte ordem, vale ressaltar de maneira *bottom-up*:

a) identificação dos objetos de interesse;

b) projeção em um hiperespaço;

c) definição de métricas de similaridade.

A identificação dos objetos de interesse é abordada através de uma técnica chamada segmentação semântica. Esta técnica visa identificar os distintos objetos existentes em uma imagem, e a posição onde se encontram. Por um lado, a identificação da localização do objeto na imagem tende a ser natural ao se buscar objetos em imagens. Contudo, a exata localização do objeto tem pouca importância ao se considerar as consultas realizadas, neste trabalho. Pois,

tais consultas aplicadas não especificam a posição do objeto na imagem. O impacto disto é que a métrica adotada para avaliação IoU pode penalizar em demasia o modelo de segmentação, visto que, pequenas variações na forma do objeto detectado não constituem um problema no resultado final, considerando o foco adotado neste trabalho, ainda que o valor de IoU seja impactado. No entanto, ao aumentar a precisão da detecção, de maneira análoga aumenta-se a precisão de sistemas como o proposto. Visto que, fazem uso da informação extraída por métodos de segmentação semântica.

A etapa seguinte consiste em identificar preposições espaciais, e consequentemente para a identificação da preposição é necessa´rio definir a topologia. Para a definição da topologia neste trabalho, utilizou-se o RCC. Em síntese, para a aplicação do RCC deve-se definir o objeto alvo e o objeto de referência considerando-se um par de objetos. Após a identificação da topologia, aplicou-se um classificador com o intuito de estimar uma preposição espacial. Assim, utilizou-se uma rede neural e um conjunto predefinido de preposições.

Para a definição dos objetos alvo e de referência utilizou-se uma máscara de cores. Esta máscara foi aplicada com o intuito de reduzir a ambiguidade nesta tarefa. Neste cenário foi possível testar algumas hipóteses, a principal hipótese verificada nesta etapa é se a informação da topologia acarretaria em uma melhora da detecção das preposições espaciais. Vale ressaltar que, de acordo com os experimentos realizados, a acurácia foi superior nos testes que foram realizados provendo apenas ambos os objetos, isto é, removendo-se o fundo da imagem. Assim, realizou-se um teste com uma imagem contendo um par de objetos processados através de uma máscara de cores e com o fundo removido, com o intuito de verificar o ganho ao se incluir a informação do RCC. Em resumo, o fornecimento da topologia, expressa em termos do RCC, não resultou em uma melhora na detecção, devido ao fato desta informação ser implicitamente extraída pela rede neural. Isto é, a rede neural é capaz de encontrar um hiperspaço e projetar a entrada e abstrair uma preposição do conjunto de preposições de interesse.

De maneira análoga, ao se manter o fundo da imagem, há a introdução de ruído. Contudo, se atribuir zero ao fundo este comportamento é reduzido, pois a entropia do núcleo da convolução será maior nos *pixels* que correspondem aos objetos de interesse e na normalização através do máximo apenas a informação referente aos objetos é mantido.

Assim, criou-se uma representação que contém pares de objetos e relações entre estes objetos para cada imagem do domínio. De posse desta informação foi possível verificar o desempenho deste método na recuperação de imagens. A tarefa de recuperação é particularmente complexa pois a relevância de cada imagem dado uma consulta se altera bastante. Isto é, diferentes consultas podem retornar o mesmo conjunto de imagens, pois a semântica existente em uma consulta é mais abrangente do que a informação extraída. Por um lado, o método apresentado neste trabalho aborda este problema aumentando a eficácia em comparação com outros métodos recentes encontrados na literatura, por outro lado, ainda situa-se distante do completo entendimento da consulta com relação às imagens. O método proposto focou-se apenas na extração de informação das imagens, pois acredita-se que esse entendimento possa ser utilizado

em tarefas além da recuperação de imagens. Adicionalmente, pode ser eficaz ao se aumentar o conhecimento sobre uma base de imagens. Outros métodos como por exemplo Lan et al. (2012), adotam um paradigma distinto e buscam modelar as consultas com relação às imagens relevantes, uma abordagem híbrida entre o método proposto e o paradigma investigado em Lan et al. (2012) pode ser mais eficaz caso a aplicação seja a recuperação de imagens.

O âmbito deste trabalho era fazer a fusão de informação de baíxo nível, extraída de imagens, com informação de alto nível fornecida por especialistas. De tal maneira que fosse possível verificar se esta informação de alto nível pode ser aproveitada por processos computacionais, consequentemente, melhorando a percepção destes agentes acerca do contexto em que estão inseridos. Assim, após todo o processamento realizado buscou-se fundir ambas fontes de informação, como forma de representação do conhecimento adotou-se a SUMO. De acordo com o cenário em que este trabalho está inserido validou-se esta extensão do SUMO para realizar a recuperação de imagens. O sistema foi capaz de abstrair relações que não foram diretamente extraídas de maneira automática pois constituiam relações derivadas, e a abordagem proposta para extração de informação de baíxo nível é limitada à relações binárias. No entanto, a topologia em um ambiente real é muito mais complexa e dinâmica sofrendo com variações inclusive devido ao ponto de vista dos observadores.

Como trabalhos futuros propõe-se a aplicação deste método em um sistema contendo multiagentes, onde cada agente possa fazer inferências a respeito de sua localização, e consequentemente do que é percebido à partir deste ponto, combinado em um arcabouço único, descentralizado, onde contenha informação de outros agentes localizados em pontos distintos.

REFERÊNCIAS

ALAHI, A.; ORTIZ, R.; VANDERGHEYNST, P. FREAK: Fast Retina Keypoint. In: PROCEEDINGS of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Washington, DC, USA: IEEE Computer Society, 2012. (CVPR '12), p. 510–517.

ARANDJELOVIC, Relja. Three Things Everyone Should Know to Improve Object Retrieval. In: PROCEEDINGS of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Washington, DC, USA: IEEE Computer Society, 2012. (CVPR '12), p. 2911–2918.

ARBELAEZ, Pablo et al. Semantic Segmentation using Regions and Parts. In: PROCEEDINGS of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Washington, DC, USA: IEEE Computer Society, 2012. (CVPR '12), p. 3378–3385.

ARTALE, Alessandro et al. Part-whole relations in object-centered systems: An overview. **Data and Knowledge Engineering**, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 20, n. 3, p. 347–383, nov. 1996.

ATIF, Jamal; HUDELOT, Céline; BLOCH, Isabelle. Explanatory Reasoning for Image Understanding Using Formal Concept Analysis and Description Logic. **IEEE Transactions on systems, Man, and Cybernetics: Systems**, v. 44, n. 5, p. 552–570, maio 2014.

BAY, Herbert; TUYTELAARS, Tinne; GOOL, Luc Van. Speeded-Up Robust Features (SURF). **Computer Vision and Image Understanding**, Elsevier Science Inc., New York, v. 110, n. 3, p. 346–359, jun. 2008.

BHOWMICK, P. K.; ROY, D.; BASU, A. A Framework For Manual Ontology Engineering For Management Of Learning Material Repository. **Journal of Computer Science and Applications**, v. 7, n. 2, p. 30–51, 2010.

BLOCH, I.; HUDELOT, C.; ATIF, J. On the Interest of Spatial Relations and Fuzzy Representations for Ontology-Based Image Interpretation. In: PROCEEDINGS of the 7th International Conference on Advances in Pattern Recognition. Kolkata, India: ICAPR, jan. 2007. (ICAPR'07), p. 15–25.

CALONDER, Michael et al. BRIEF: Binary Robust Independent Elementary Features. In: **Computer Vision –ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV**. Edição: Kostas Daniilidis, Petros Maragos e Nikos Paragios. [S.l.]: BRIEF: Binary Robust Independent Elementary Features, set. 2010. p. 778–792.

CASATI, Roberto; VARZI, Achille C. The Structure of Spatial Localization. **Philosophical Studies**, Springer, v. 82, n. 2, p. 205–239, 1996.

CHOI, M. J. et al. Exploiting hierarchical context on a large database of object categories. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], jun. 2010. p. 129–136.

COHN, Anthony G.; RENZ, Jochen. Chapter 13 Qualitative Spatial Representation and Reasoning. **Foundations of Artificial Intelligence**, Elsevier, v. 3, p. 551–596, 2008. Handbook of Knowledge Representation.

CSURKA, Gabriela; LARLUS, Diane; PERRONNIN, Florent. What is a good evaluation measure for semantic segmentation? In:_____. **Proceedings of the British Machine Vision Conference**. [S.l.]: BMVA Press, 2013.

CSURKA, Gabriela; PERRONNIN, Florent. An Efficient Approach to Semantic Segmentation. **International Journal of Computer Vision**, v. 95, n. 2, p. 198–212, 2011.

CSURKA, Gabriella et al. Visual categorization with bags of keypoints. In: WORKSHOP on Statistical Learning in Computer Vision, ECCV. [S.l.: s.n.], 2004. p. 1–22.

DALAL, N.; TRIGGS, B. Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). [S.l.: s.n.], jun. 2005. v. 1, 886–893 vol. 1.

DOLLAR, P. et al. Pedestrian Detection: An Evaluation of the State of the Art. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 34, n. 4, p. 743–761, 2012.

DOUZE, Matthijs et al. Evaluation of GIST Descriptors for Web-scale Image Search. In: PROCEEDINGS of the ACM International Conference on Image and Video Retrieval. Santorini, Fira, Greece: ACM, 2009. (CIVR '09), 19:1–19:8.

ESPUNY, Ferran; MONASSE, Pascal; MOISAN, Lionel. A New A Contrario Approach for the Robust Determination of the Fundamental Matrix. In: **Image and Video Technology –PSIVT 2013 Workshops: GCCV 2013, GPID 2013, PAESNPR 2013, and QACIVA 2013, Guanajuato, Mexico, October 28-29, 2013, Revised Selected Papers**. Edição: Fay Huang e Akihiro Sugimoto. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. p. 181–192.

EVERINGHAM, Mark et al. The Pascal Visual Object Classes (VOC) Challenge. **International Journal of Computer Vision**, v. 88, n. 2, p. 303–338, 2010.

FEI-FEI, L.; PERONA, P. A Bayesian Hierarchical Model for Learning Natural Scene Categories. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern

Recognition (CVPR'05). Washington, DC, USA: IEEE Computer Society, 2005. v. 2. (CVPR '05), p. 524–531.

FELZENSZWALB, P. F. et al. Object Detection with Discriminatively Trained Part-Based Models. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 32, n. 9, p. 1627–1645, 2010.

GENESERETH, Michael et al. **Knowledge Interchange Format Version 3.0 Reference Manual**. [S.l.: s.n.], 1992.

GOULD, Stephen; HE, Xuming. Scene Understanding by Labeling Pixels. **Commun. ACM**, New York, v. 57, n. 11, p. 68–77, out. 2014.

GRAUMAN, Kristen; LEIBE, Bastian. **Visual Object Recognition**. [S.l.]: Morgan e Claypool Publishers, 2011.

GRUBER, Thomas R. A translation approach to portable ontology specifications. **Knowledge Acquisition**, v. 5, p. 199–220, 1993.

HARRIS, C.; STEPHENS, M. A combined corner and edge detector. In:_____. **Proceedings of the Alvey Vision Conference**. [S.l.]: Alvety Vision Club, 1998. p. 23.1–23.6.

HARTLEY, Richard; ZISSERMAN, Andrew. **Multiple view geometry in computer vision**. [S.l.]: Cambridge University Press, ISBN: 0521540518, 2004.

HAYES, Steven C.; BARNES-HOLMES, Dermot; ROCHE, Bryan. **Relational frame theory: A post-Skinnerian account of human language and cognition**. [S.l.]: Kluwer Academic, 2001.

HE, Kaiming et al. Deep Residual Learning for Image Recognition. In: THE IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.: s.n.], 2016.

HERNÁNDEZ-GRACIDAS, Carlos Arturo; SUCAR, Luis Enrique; MONTES-Y-GÓMEZ, Manuel. Improving image retrieval by using spatial relations. **Multimedia Tools and Applications**, v. 62, n. 2, p. 479–505, 2013. Disponível em: <https://doi.org/10.1007/s11042-011-0911-1>.

HERNÁNDEZ-GRACIDAS, Carlos; SUCAR, L. Enrique. Markov Random Fields and Spatial Information to Improve Automatic Image Annotation. In: **Advances in Image and Video Technology: Second Pacific Rim Symposium, PSIVT 2007 Santiago, Chile, December 17-19, 2007 Proceedings**. Edição: Domingo Mery e Luis Rueda. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. p. 879–892.

HONG, Richang et al. Coherent Semantic-visual Indexing for Large-scale Image Retrieval in the Cloud. **IEEE Transactions on Image Processing**, v. 26, n. 9, p. 4128–4138, 2017.

HUDELOT, C.; ATIF, J.; BLOCH, I. Fuzzy Spatial Relation Ontology for Image Interpretation. **Fuzzy Sets Syst.**, Elsevier North-Holland, Inc., Amsterdam, The Netherlands, The Netherlands, v. 159, n. 15, p. 1929–1951, ago. 2008.

ISLAM, Sk Mazharul et al. Content-based image retrieval based on multiple extended fuzzy-rough framework. **Applied Soft Computing**, v. 57, p. 102–117, 2017.

JAIN, Ramesh; KASTURI, Rangachar; SCHUNCK, Brian G. **Machine Vision**. New York: McGraw-Hill, Inc., 1995.

JÉGOU, Hervé; DOUZE, Matthijs; SCHMID, Cordelia. Improving bag-of-features for large scale image search. **International Journal of Computer Vision**, v. 87, n. 3, p. 316–336, 2010.

JIA, Yangqing et al. Caffe: Convolutional Architecture for Fast Feature Embedding. In: PROCEEDINGS of the 22Nd ACM International Conference on Multimedia. New York, NY, USA: ACM, 2014. (MM '14), p. 675–678.

KARPATHY, Andrej; JOULIN, Armand; FEI-FEI, Li. Deep Fragment Embeddings for Bidirectional Image Sentence Mapping. In: PROCEEDINGS of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press, 2014. p. 1889–1897.

KE, Yan; SUKTHANKAR, Rahul. PCA-SIFT: a more distinctive representation for local image descriptors. In: PROCEEDINGS of the 2004 IEEE computer society conference on Computer vision and pattern recognition. Washington, DC, USA: IEEE Computer Society, 2004. v. 2. (CIVR '04), p. 506–513.

KESORN, Kraisak; POSLAD, Stefan. Semantic Restructuring of Natural Language Image Captions to Enhance Image Retrieval. **Journal of Multimedia**, v. 4, p. 284–297, 2009.

KOLLA, Morarjee; VENU GOPAL, T. Region Based Semantic Image Retrieval Using Ontology. In: edição: Suresh Chandra Satapathy. Singapore: Springer Singapore,2017. p. 421–428.

KRIZHEVSKY, Alex; HINTON, Geoffrey E. Using very deep autoencoders for content-based image retrieval. In: PROCEEDINGS of The European Symposium on Artificial Neural Networks (ESANN). [S.l.: s.n.], 2011.

KRIZHEVSKY, Alex; SUTSKEVER, Ilya; HINTON, Geoffrey E. ImageNet Classification with Deep Convolutional Neural Networks. In: PROCEEDINGS of the 25th International

Conference on Neural Information Processing Systems. Lake Tahoe, Nevada: Curran Associates, Inc., 2012. (NIPS'12), p. 1097–1105.

LAN, Tian et al. Image retrieval with structured object queries using latent ranking svm. In: PROCEEDINGS of the 12th European Conference on Computer Vision - Volume Part VI. Berlin, Heidelberg: Springer-Verlag, 2012. (ECCV'12), p. 129–142.

LARLUS, Diane; JURIE, Frédéric. Latent mixture vocabularies for object categorization and segmentation. **Image and Vision Computing**, v. 27, n. 5, p. 523–534, 2009.

LEHMANN, Jens; VÖELKER, Johanna. An Introduction to Ontology Learning. In: _____. **Perspectives on Ontology Learning**. Edição: Jens Lehmann e Johanna Vöelker. [S.l.]: AKA Heidelberg, 2014. p. ix–xvi.

LEUTENEGGER, Stefan; CHLI, Margarita; SIEGWART, Roland Y. BRISK: Binary Robust invariant scalable keypoints. In: PROCEEDINGS of the 2011 International Conference on Computer Vision. Washington, DC, USA: IEEE Computer Society, 2011. (ICCV '11), p. 2548–2555.

LI, Li-Jia; SOCHER, Richard; FEI-FEI, Li. Towards Total Scene Understanding:Classification, Annotation and Segmentation in an Automatic Framework. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], jun. 2009. p. 2036–2043.

LIN, P. L.; TAN, W. H. An Efficient Method for the Retrieval of Objects by Topological Relations in Spatial Database Systems. **Inf. Process. Manage.**, v. 39, n. 4, p. 543–559, jul. 2003.

LIN, Tsung-Yi et al. Microsoft COCO: Common Objects in Context. In: EUROPEAN Conference on Computer Vision (ECCV). [S.l.: s.n.], 2014. p. 740–755.

LINDENBERG, Tony. Feature Detection with Automatic Scale Selection. **International Journal of Computer Vision**, v. 30, n. 2, p. 79–116, 1998.

LIU, Ce et al. SIFT Flow: Dense Correspondence across Different Scenes. In: **Computer Vision –ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part III**. Edição: David Forsyth, Philip Torr e Andrew Zisserman. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. p. 28–42.

LIU, Ling; ZSU, M. Tamer. **Encyclopedia of Database Systems**. 1st. [S.l.]: Springer Publishing Company, 2009.

LIU, Yuee. **Ontology-Based Image Annotation**. 2010. Tese (Doutorado) – Queensland University of Technology.

LOGAN, Gordon D.; SADLER, Daniel D. **A computational analysis of the apprehension of spatial relations**. [S.l.]: MIT Press, 1996.

LONG, Jonathan; SHELHAMER, Evan; DARRELL, Trevor. Fully convolutional networks for semantic segmentation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.: s.n.], 2015.

LOWE, David G. Distinctive Image Features from Scale-Invariant Keypoints. **International Journal of Computer Vision**, v. 60, n. 2, p. 91–110, nov. 2004.

MALINOWSKI, Mateusz; FRITZ, Mario. **A Pooling Approach to Modelling Spatial Relations for Image Retrieval and Annotation**. [S.l.], 2014.

MARIUS MUJA, David G. Lowe. Fast Matching of Binary Features. In: PROCEEDINGS of the 2012 Ninth Conference on Computer and Robot Vision. Washington, DC, USA: IEEE Computer Society, 2012. (CRV '12), p. 404–410.

MASCARDI, V.; LOCORO, A.; ROSSO, P. Automatic Ontology Matching via Upper Ontologies: A Systematic Evaluation. **IEEE Transactions on Knowledge and Data Engineering**, v. 22, n. 5, p. 609–623, 2010.

MASCARDI, Viviana; CORDÌ, Valentina; ROSSO, Paolo. **A comparison of upper ontologies**. [S.l.], 2006.

MATAS, J. et al. Robust Wide Baseline Stereo from Maximally Stable Extremal Regions. In: _____. **Proceedings of the British Machine Vision Conference**. [S.l.]: BMVA Press, 2002. p. 36.1–36.10.

MIKOLAJCZYK, Krystian; SCHMID, Cordelia. A performance evaluation of local descriptors. **IEEE Trans. Pattern Anal. Mach. Intell.**, v. 27, n. 10, p. 1615–1630, 2005.

MILLER, George A. WordNet: A Lexical Database for English. **Communications of the ACM**, v. 38, n. 11, p. 39–41, 1995.

MORALES-GONZÁLEZ, Annette; GARCÍA-REYES, Edel B. Simple object recognition based on spatial relations and visual features represented using irregular pyramids. **Multimedia Tools and Applications**, v. 63, n. 3, p. 875–897, 2013.

MOREL, Jean Michel; YU, Guoshen. ASIFT: A New framework for fully affine invariant image comparison. **SIAM Journal on Imaging Sciences**, v. 2, n. 2, p. 438–469, 2009.

MUJA, Marius; LOWE, David G. Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration. In: IN VISAPP International Conference on Computer Vision Theory and Applications. [S.l.: s.n.], 2009. p. 331–340.

NILES, Ian; PEASE, Adam. Towards a Standard Upper Ontology. In: PROCEEDINGS of the International Conference on Formal Ontology in Information Systems - Volume 2001. New York, NY, USA: ACM, 2001. (FOIS '01), p. 2–9.

NOY, Natalya F. Semantic Integration: A Survey of Ontology-based Approaches. **SIGMOD Rec.**, v. 33, n. 4, p. 65–70, 2004.

OLIVA, Aude; TORRALBA, Antonio. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. **International Journal of Computer Vision**, v. 42, n. 3, p. 145–175, 2001.

PARK, Kyung-Wook; JEONG, Jin-Woo; LEE, Dong-Ho. OLYBIA: Ontology-Based Automatic Image Annotation System Using Semantic Inference Rules. In: **Advances in Databases: Concepts, Systems and Applications: 12th International Conference on Database Systems for Advanced Applications, DASFAA 2007, Bangkok, Thailand, April 9-12, 2007. Proceedings**. Edição: Ramamohanarao Kotagiri. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. p. 485–496. Disponível em: <https://doi.org/10.1007/978-3-540-71703-4_42>.

PEASE, A.; NILES, I. Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In: PROCEEDINGS of the 2003 International Conference on Information and Knowledge Engineering. [S.l.: s.n.], 2003.

PEASE, Adam. **Ontology: A Practical Guide**. [S.l.]: Articulate Software Press, 2011.

PEASE, Adam; NILES, Ian; LI, John. The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications. In: IN Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web. [S.l.: s.n.], 2002.

PHILBIN, James et al. Lost in Quantization: Improving Particular Object Retrieval in Large Scale Image Databases. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2008. p. 1–8.

PHILBIN, James et al. Object retrieval with large vocabularies and fast spatial matching. In: IEEE Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2007.

PHILBIN; SIVIC; ZISSERMAN. Object Mining using a Matching Graph on Very Large Image Collections. In: PROCEEDINGS of the 2008 Sixth Indian Conference on Computer

Vision, Graphics & Image Processing. Washington, DC, USA: IEEE Computer Society, 2008. (ICVGIP '08), p. 738–745.

POPESCU, Adrian; MILLET, Cristophe; MOELLIC, Pierre-Alain. Ontology Driven Content Based Image Retrieval. In: PROCEEDINGS of the 6th ACM international conference on Image and video retrieval. [S.l.: s.n.], 2007. p. 387–394.

RAHMAN, Md Atiqur; WANG, Yang. Optimizing Intersection-Over-Union in Deep Neural Networks for Image Segmentation. In: edição: George Bebis. Cham: Springer International Publishing, 2016. p. 234–244.

RANDELL, D. A.; CUI, Z.; COHN, A. G. A spatial logic based on regions and connection. In: 165-176. PROCEEDINGS of the 3rd International Conference on Knowledge Representation and Reasoning. [S.l.: s.n.], 1992.

RIAD, Alaa. M.; ELMINIR, Hamdy. K.; ABD-ELGHANY, Sameh. A Literature Review of Image Retrieval based On Semantic Concept. **International Journal of Computer Applications**, v. 40, n. 11, p. 12–19, 2012.

RUBLEE, Ethan et al. ORB: An efficient alternative to SIFT or SURF. In: PROCEEDINGS of the 2011 International Conference on Computer Vision. [S.l.: s.n.], 2011. p. 2564–2571.

RUSSAKOVSKY, Olga et al. ImageNet Large Scale Visual Recognition Challenge. **International Journal of Computer Vision**, v. 115, n. 3, p. 211–252, dez. 2015.

SALAKHUTDINOV, Ruslan; HINTON, Geoffrey. Semantic Hashing. **International Journal of Approximate Reasoning**, v. 50, n. 7, p. 969–978, 2009.

SANDE, Koen E. A. van de et al. Segmentation as Selective Search for Object Recognition. In: INTERNATIONAL Conference on Computer Vision. [S.l.]: IEEE, 2011. p. 1879–1886.

SANKAT, Monica; THAKUR, R. S.; JALOREE, Shailesh. Semi-Automatic Ontology Design for Educational Purposes. In: **Pattern and Data Analysis in Healthcare Settings**. Edição: Vivek Tiwari. Hershey, PA, USA: IGI Global, 2016. p. 124–142.

SCHULZ, Stephan. System Description: E 1.8. In: **Proceedings of the 19th LPAR**. Edição: Ken McMillan, Aart Middeldorp e Andrei Voronkov. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. p. 735–743.

SHI, Jianbo; TOMASI, Carlo. Good Features to Track. In: 1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 1994. p. 596–600.

SIDDIQUIE, Behjat et al. Multi-modal image retrieval for complex queries using small codes. In: PROCEEDINGS of International Conference on Multimedia Retrieval. New York, NY, USA: ACM, 2014. (ICMR '14), 321:321–321:328.

SIKOS, Leslie F. Knowledge Representation with Semantic Web Standards. In: [s.l.]: Springer International Publishing, 2017. p. 11–49.

SIVIC, Josef; ZISSERMAN, Andrew. Video Google: Efficient Visual Search of Videos. In: **Toward Category-Level Object Recognition**. Edição: J. Ponce. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006. p. 127–144.

SMITH, Barry. Basic Concepts of Formal Ontology. In: GUARINO, Nicola (Ed.). **Formal Ontology in Information Systems**. [S.l.]: IOS Press, 1998. p. 19–28.

SOCHER, Richard et al. Parsing Natural Scenes and Natural Language with Recursive Neural Networks. In: PROCEEDINGS of the 28th International Conference on International Conference on Machine Learning. Bellevue, Washington, USA: Omnipress, 2011. p. 129–136.

SUDOWE, Patrick; LEIBE, Bastian. Efficient Use of Geometric Constraints for Sliding-Window Object Detection in Video. In: PROCEEDINGS of the 8th International Conference on Computer Vision Systems. Berlin, Heidelberg: Springer-Verlag, 2011. (ICVS'11), p. 11–20.

SUTCLIFFE, Geoff. The TPTP World –Infrastructure for Automated Reasoning. In: edição: Edmund M. Clarke e Andrei Voronkov. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. p. 1–12.

SZEGEDY, Christian et al. Going Deeper with Convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.: s.n.], 2015. p. 1–9.

SZELISKI, Richard. **Computer Vision: Algorithms and Applications**. 1st. [S.l.]: Springer-Verlag New York, 2010.

TIGHE, Joseph; LAZEBNIK, Svetlana. SuperParsing: Scalable Nonparametric Image Parsing with Superpixels. In: **Computer Vision – ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part V**. Edição: Kostas Daniilidis, Petros Maragos e Nikos Paragios. Berlin, Heidelberg: Springer Berlin Heidelberg", 2010. p. 352–365.

TOLA, Engin; LEPETIT, Vincent; FUA, Pascal. DAISY: An Efficient Dense Descriptor Applied to Wide-Baseline Stereo. **IEEE Trans. Pattern Anal. Mach. Intell.**, v. 32, n. 5, p. 815–830, 2010.

TORRALBA, Antonio; FERGUS, Rob; FREEMAN, William T. 80 million tiny images: a large dataset for non-parametric object and scene recognition. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 30, n. 11, p. 1958–1970, 2008.

TORRALBA, Antonio; FERGUS, Rob; WEISS, Yair. Small Codes and Large Image Databases for Recognition. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], jun. 2008. p. 1–8.

TSAI, Chih-Fong. Bag-of-Words Representation in Image Annotation: A Review. **International Scholarly Research Network**, v. 2012, n. 19, p. 1–19, 2012.

TURK, Matthew A.; PENTLAND, Alex P. Eigenfaces for Recognition. **J. Cognitive Neuroscience**, v. 3, n. 1, p. 71–86, jan. 1991.

USCHOLD, Michael; GRUNINGER, Michael. Ontologies and Semantics for Seamless Connectivity. **SIGMOD Rec.**, v. 33, n. 4, p. 58–64, 2004.

WAN, Ji et al. Deep Learning for Content-Based Image Retrieval: A Comprehensive Study. In: PROCEEDINGS of the 22Nd ACM International Conference on Multimedia. Orlando, Florida, USA: ACM, 2014. (MM '14), p. 157–166.

WANG, Huafeng et al. Deep Learning for Image Retrieval: What Works and What Doesn't. In: PROCEEDINGS of the 2015 IEEE International Conference on Data Mining Workshop (ICDMW). Washington, DC, USA: IEEE Computer Society, 2015. (ICDMW '15), p. 1576–1583.

XIA, Rongkai et al. Supervised Hashing for Image Retrieval via Image Representation Learning. In: PROCEEDINGS of the Twenty-Eighth AAAI Conference on Artificial Intelligence. [S.l.]: AAAI Press, 2014. (AAAI'14), p. 2156–2162.

XIA et al. Deep convolutional hashing using pairwise multi-label supervision for large-scale visual search. **Signal Processing: Image Communication**, 2017.

YANG, Jun et al. Evaluating Bag-of-Visual-Words Representations in Scene Classification. In: PROCEEDINGS of the international workshop on Workshop on multimedia information retrieval. New York, NY, USA: ACM, 2007. (MIR '07), p. 197–206.

ZHENG, Shuai et al. Conditional Random Fields As Recurrent Neural Networks. In: 9. PROCEEDINGS of the 2015 IEEE International Conference on Computer Vision (ICCV). Washington, DC, USA: IEEE Computer Society, 2015. (ICCV 15, 9), p. 1529–1537.

ZHOU, Bolei et al. Scene Parsing through ADE20K Dataset. In: PROCEEDINGS of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2017.

ZHOU, Bolei et al. Semantic understanding of scenes through the ade20k dataset. In: ARXIV preprint arXiv:1608.05442. [S.l.: s.n.], 2016.

ZHU, Hongyuan et al. Beyond pixels: A comprehensive survey from bottom-up to semantic image segmentation and cosegmentation. **Journal of Visual Communication and Image Representation**, v. 34, p. 12–27, 2016.

APÊNDICE A – DISTRIBUIÇÃO DOS OBJETOS POR IMAGEM

Esta tabela apresenta a distribuição dos objetos por imagem com relação ao conjunto de dados proposto por (LAN et al., 2012).

Tabela 15 – Distribuição dos objetos no conjunto de treino e teste. Proporção entre os conjuntos de dados, e a medida IoU para cada classe

Objeto	Treino		Teste		Relação Treino / Teste		IoU
	freq.	área (%)	freq.	área (%)	freq.	área (%)	
Airplane	36	0,06	34	0,04	1,06	1,42	7,78
Armchair	144	0,05	126	0,05	1,14	1,03	1,27
Awning	90	0,03	81	0,03	1,11	1,00	2,76
Bag	61	0,01	82	0,01	0,74	0,88	0,00
Balcony	84	0,07	67	0,06	1,25	1,20	2,04
Ball	29	0,03	16	0,01	1,81	2,31	0,38
Bars	36	0,06	29	0,10	1,24	0,56	0,37
Basket	87	0,01	93	0,02	0,94	0,74	0,00
Bed	216	0,21	222	0,22	0,97	0,91	29,20
Bench	87	0,03	95	0,03	0,92	1,02	0,65
Boat	74	0,09	73	0,07	1,01	1,21	1,37
Book	199	0,01	207	0,01	0,96	0,97	0,00
Bookcase	70	0,19	66	0,21	1,06	0,91	6,94
Bottle	142	0,01	138	0,01	1,03	0,63	0,00
Bowl	53	0,01	47	0,01	1,13	0,53	0,00
Box	243	0,03	254	0,03	0,96	0,94	0,16
Bread	29	0,02	30	0,03	0,97	0,62	0,00
Building	1319	0,24	1321	0,23	1,00	1,03	17,91
Bus	38	0,02	30	0,02	1,27	0,81	0,00
Cabinet	136	0,07	135	0,07	1,01	0,98	1,08
Candle	48	0,00	35	0,01	1,37	0,63	0,00
Car	611	0,01	560	0,01	1,09	0,76	3,22
Chair	546	0,03	498	0,03	1,10	0,85	2,04
Chandelier	62	0,02	48	0,02	1,29	0,94	0,23
Clock	84	0,00	75	0,00	1,12	1,02	0,00
Closet	40	0,30	45	0,27	0,89	1,09	0,49
Clothes	32	0,05	26	0,05	1,23	1,10	3,74
Counter	75	0,12	67	0,16	1,12	0,78	0,45
Countertop	67	0,07	65	0,08	1,03	0,83	2,68
Cupboard	115	0,11	119	0,10	0,97	1,09	4,11
Curtain	255	0,09	225	0,09	1,13	0,95	11,88
Cushion	83	0,02	87	0,01	0,95	1,17	0,00
Desk	137	0,09	129	0,09	1,06	1,01	4,54
Dish	12	0,03	9	0,05	1,33	0,56	0,,00
Dishwasher	43	0,03	30	0,04	1,43	0,94	5,79
Dome	20	0,06	18	0,02	1,11	3,63	0,38
Door	760	0,04	725	0,04	1,05	0,90	3,20
Drawer	25	0,02	24	0,01	1,04	1,69	0,00
Easel	18	0,05	26	0,07	0,69	0,76	0,00

Fence	263	0,04	250	0,05	1,05	0,83	3,44
Field	190	0,27	211	0,27	0,90	0,99	11,04
Fireplace	34	0,06	38	0,08	0,89	0,80	3,57
Floor	1615	0,24	1591	0,24	1,02	1,00	43,50
Flower	181	0,02	159	0,03	1,14	0,73	1,03
Gate	56	0,02	64	0,01	0,88	1,24	0,81
Glass	53	0,01	69	0,00	0,77	1,94	0,00
Grass	491	0,15	454	0,14	1,08	1,04	15,26
Ground	326	0,18	337	0,17	0,97	1,07	9,22
Handrail	86	0,03	67	0,04	1,28	0,72	0,97
Headstone	5	0,02	3	0,03	1,67	0,77	0,00
Machine	18	0,05	13	0,21	1,38	0,22	0,00
Microwave	48	0,02	46	0,02	1,04	0,96	0,98
Mirror	153	0,07	149	0,07	1,03	0,94	6,00
Monitor	38	0,02	43	0,02	0,88	0,83	0,65
Mountain	616	0,19	647	0,19	0,95	1,04	20,24
Oven	31	0,04	32	0,08	0,97	0,51	3,53
Path	127	0,07	128	0,09	0,99	0,80	5,80
Person	860	0,03	888	0,04	0,97	0,92	3,61
Picture	132	0,01	135	0,01	0,98	1,46	0,05
Pillow	111	0,02	123	0,02	0,90	1,01	0,53
Plant	694	0,06	670	0,05	1,04	1,01	5,03
Plate	55	0,01	55	0,01	1,00	0,97	0,00
Platform	28	0,15	24	0,09	1,17	1,62	3,27
Poster	85	0,02	88	0,02	0,97	1,00	0,13
Pot	179	0,01	178	0,01	1,01	0,83	0,26
Railing	97	0,06	81	0,06	1,20	0,89	2,57
Refrigerator	57	0,09	63	0,10	0,90	0,88	12,69
River	112	0,21	97	0,18	1,15	1,15	8,96
Road	702	0,18	690	0,18	1,02	0,96	42,49
Rock	110	0,12	120	0,11	0,92	1,06	8,01
Rug	117	0,08	132	0,09	0,89	0,85	2,81
Sand	86	0,18	87	0,22	0,99	0,80	6,66
Screen	88	0,02	83	0,03	1,06	0,73	4,58
Sea	205	0,33	211	0,34	0,97	0,98	19,40
Seats	97	0,11	100	0,12	0,97	0,90	7,31
Shelves	168	0,24	159	0,26	1,06	0,92	9,85
Shoes	29	0,01	31	0,01	0,94	0,84	0,00
Showcase	59	0,35	50	0,34	1,18	1,02	1,47
Sign	269	0,01	268	0,01	1,00	1,58	0,73
Sink	150	0,03	144	0,03	1,04	1,03	7,16
Sky	2049	0,28	2064	0,28	0,99	1,03	62,45
Sofa	123	0,10	112	0,11	1,10	0,94	11,00
Staircase	143	0,06	137	0,07	1,04	0,84	2,48
Stand	129	0,13	136	0,10	0,95	1,32	1,32
Steps	138	0,03	128	0,03	1,08	1,04	2,02
Stone	78	0,03	77	0,03	1,01	1,18	0,42

Stool	75	0,01	82	0,01	0,91	0,99	0,00
Stove	81	0,05	87	0,05	0,93	1,05	2,49
Streetlight	195	0,00	197	0,00	0,99	1,07	0,00
Table	480	0,06	476	0,07	1,01	0,94	2,49
Television	71	0,02	59	0,03	1,20	0,87	1,46
Text	152	0,02	115	0,03	1,32	0,78	4,02
Toilet	45	0,06	43	0,06	1,05	1,08	5,01
Towel	65	0,02	61	0,02	1,07	0,76	0,00
Tower	29	0,11	22	0,14	1,32	0,80	6,56
Tray	47	0,02	48	0,02	0,98	0,94	0,00
Tree	1382	0,10	1311	0,10	1,05	0,97	13,96
Truck	86	0,05	93	0,05	0,92	0,96	7,80
Umbrella	26	0,02	28	0,03	0,93	0,67	0,00
Van	96	0,02	96	0,02	1,00	0,65	1,95
Vase	132	0,01	101	0,01	1,31	0,90	0,00
Video	10	0,02	8	0,03	1,25	0,61	0,00
Wall	2009	0,24	1939	0,23	1,04	1,01	15,72
Water	147	0,27	139	0,31	1,06	0,85	8,61
Window	1178	0,04	1142	0,04	1,03	1,12	4,01
Média	229	0,13	223	0,13	1,06	0,98	5,22

Fonte: Autor.

APÊNDICE B – ARTIGO SUBMETIDO

Pattern Recognition Letters
journal homepage: www.elsevier.com

Representation and retrieval of images by means of spatial relations between objects

Danilo N.Santos [a], Paulo E.Santos [a], AdamPease [b]

[a] *Centro Universitário da FEI, São Bernardo do Campo, Brazil*
[b] *Articulate Software, San Jose, California*

ABSTRACT

The present work addresses the semantic gap that exists in content-based image retrieval by introducing an approach to describe images by means of spatial relations. The proposed approach is called Image Retrieval using Region Analysis (IRRA) and relies on decomposing images into pairs of objects. This method generates a representation composed of *n* triples, each one containing: a *noun*, a *preposition* and another *noun*. This representation paves the way to enable image retrieval based on spatial relations.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

In this work we investigate the application of spatial relations in the image retrieval problem. The issue of representing the semantics existent in an image has been receiving great attention recently. Numerical methods (low level) are not able to fully integrate semantics (high level) due to the fact that the semantic content might be constituted by qualitative concepts. The challenge of integrating low-level information with high-level knowledge is a known problem in computer vision, often referred as the semantic gap. In this work, a multi-level approach, called *Image Retrieval using Region Analysis (IRRA)*, is proposed to retrieve images by their semantics from a bottom-up knowledge representation procedure. The method proposed here ensembles a stack of distinct neural networks in order to estimate spatial relations, expressed by a spatial preposition between pairs of objects. This procedure permits a representation of an image by the objects depicted and the relations holding between them. Thus, a sparse representation is constructed taking into account these objects and their relations. The images are indexed based on this sparse representation in order to enable fast retrieval. Finally, we extend a public ontology with this data in order to infer new relations beyond the original binary relations. This representation enables the retrieval of images based on queries with respect to spatial arrangements.

To semantically interpret an image it is necessary to: define the context, detect objects, define some similarity metric and, finally, apply some method for knowledge representation (Wan et al., 2014). We can organize the proposed framework in two distinct steps: the quantitative analysis, related to low-level information processing, and the qualitative analysis, which stands for high-level knowledge representation.

The quantitative analysis tackles the semantic gap problem using a hierarchy of classifiers. We represent the semantics building a top-down approach which contains a specific classifier for each of the following tasks: scene recognition, object segmentation and preposition estimation. The proposed approach decomposes an image into scenes and then, for each scene, it segments the related objects; given a pair of these objects, the method estimates a spatial preposition.

The qualitative analysis is built on top of Suggested Upper Merged Ontology (SUMO)[1] (Niles and Pease, 2001; Pease, 2011) and is constructed with the data obtained at the quantitative phase, i.e., scenes, objects and prepositions. This representation contains the distinct segmented objects and their relations, expressed by a spatial preposition.

For evaluation purposes, the method proposed is applied to image retrieval tasks. The results obtained show that our method outperforms recent approaches aiming at the retrieval of images by means of spatial relations.

2. Related work

One of the most common approaches for retrieving images is the paradigm known as query by example. Bag of Visual Words (BOVW) is a method widely used to perform image retrieval in this context. BOVW extracts local features and, with respect to

[1] www.ontologyportal.org

a sparse representation, performs image retrieval (Philbin et al., 2008). BOVW-based techniques retrieve images by their visual similarity. However, the retrieval task does not take into account the meaning (or semantics) of the information sought, since it is solely based on numerical analyses. On the one hand, the way an image is represented is crucial to enable a fast retrieval of images. On the other hand, there is poor correlation between the semantics one image might contain and its constructed representation (Hudelot et al., 2008).

One possible way to represent the semantics expressed in images is by following the steps described in (Hare et al., 2006): extract and describe the features of interest, perform object segmentation and create a high-level representation of the detected regions in the images. Considering non-structured regions in images, the goal is to assign a distinct label to each one of them. A number of approaches have focused in labeling regions in an image. To name some of the most relevant we can refer to those based on Conditional Random Fields (CRF) (Gould et al., 2009), approaches based on deep Neural Networks (NN) (Socher et al., 2011) or Convolutional Neural Networks (CNN) (Girshick et al., 2014), combinations of CRF and CNN have also been applied to this task (Zheng et al., 2015). These methods, however, do not take into account relations between objects in images, as considered in this paper. It is worth noticing that, in order to reduce the semantic gap, it is important to represent the spatial relations among objects, since these relations suffer less from viewpoint changes than the object recognition itself (Bloch et al., 2007). Some work has investigated the inclusion of high-level knowledge focused on spatial relations in image analysis. For example, we can cite the work of (Bloch et al., 2007) and (Hudelot et al., 2008). Both of them establish high-level assumptions to enhance the low-level processing. And more recently, we refer to (Lu et al., 2016) and (Dai et al., 2017) both of them establish relations between objects in an image based on a set of distinct predicates. Similarly, (Malinowski and Fritz, 2014) focus on image retrieval by using queries built on spatial relations between objects in images and (Mai et al., 2017) that retrieves images based on spatial arrangement of an example input image. In the present paper, we propose a new paradigm that combines quantitative processing and also qualitative analysis in an end-to-end architecture.

In order to establish a spatial relation, it is necessary to define a reference system. For instance, considering the relation x is ***in front of*** y, three concepts should be defined: the target object, the reference object and the reference system (Hudelot et al., 2008). The reference system is, in general, categorized from the observer's viewpoint (relative or absolute), or with respect to the way that the relation is used: intrinsic, extrinsic or deictic. This work uses intrinsic relations, which specify a relation under the perspective of an observer. It is important to reinforce that these relations are not constant in time and can also change their status with respect to the adopted perspective. The present work uses spatial relations as defined by a general ontology in order to accomplish the automatic extraction of the semantic content present in an image.

There are distinct definitions of ontology found in the literature (Sankat et al., 2016). In this work, ontology is understood similarly to the definition proposed by (Lehmann and Vöelker, 2014) which defines ontology as a formal knowledge representation, which may or may not be restricted to a specific domain. This representation is expressed in a manner that might be understood by a computational process, specially, in the field of artificial intelligence (Sankat et al., 2016). Ontology might be referred as the commonsense knowledge with respect to a domain of interest, and can be expressed by: concepts, relations between concepts, functions and instances. SUMO extends these with axioms in higher-order logic that attempt to define each concept.

3. Image Retrieval using Region Analysis (IRRA)

In this work we address the problem of the semantic gap in image retrieval by representing an image by the spatial relation among the objects existing in the image itself. Our aim is to combine information obtained from the pixel level (quantitative analysis) with information provided by specialists (qualitative analysis), in this work the specialist knowledge is defined in terms of spatial prepositions in a general ontology. The method proposed in this work is called Image Retrieval using Region Analysis (IRRA) and is summarized in Figure 1.

Considering an input image, IRRA applies a neural network is applied in order to segment the objects in the image, that are further combined into pairs of objects. In the sequence, for each pair, a preposition between them is estimated. Thus, we decompose the image into n-triples containing pairs of objects and a spatial preposition. In the proposed work we use these n-triples for knowledge representation and to do automated reasoning about the domain. Section 3.1 describes the quantitative analysis of IRRA in more detail..

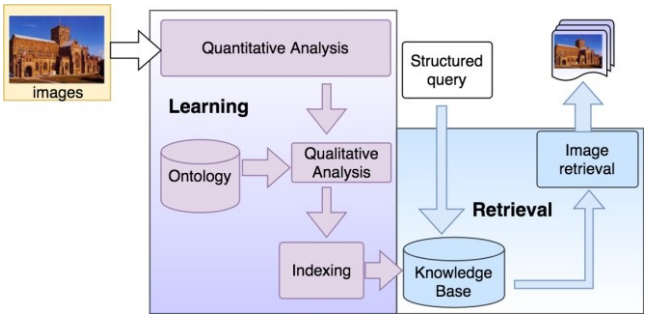


Fig. 1: Overview of IRRA method numerical analysis phase.

3.1. Quantitative analysis

The proposed pipeline of the quantitative analysis is the following:

1. Identify the context in which objects are inserted;
2. Detect regions in the images occupied by each object;
3. Combine the detected objects in pairs;
4. Estimate a spatial relation for every detected object pair.

In order to perform all these tasks, a stack of neural networks is applied. We apply three distinct neural networks to hierarchically detect context, objects and spatial relations.

Context identification: The first neural network is applied in order to estimate context. In this work, context is understood as the scene class an image belongs to. Two scenes classes are used in this work: *indoor* and *outdoor*. We created a model whose architecture is identical to that described in AlexNet (Krizhevsky et al., 2012) in order to perform this classification. The purpose of this step is to reduce the number of target objects, dismissing inconsistent objects with respect to the scene; objects such as building and cars will not be part of the segmentation of an indoor model, for instance.

After assigning specific context to an image, each image serves as input to another set of neural networks, each one of these neural networks were trained with distinct classes of objects, as described below.

Object segmentation: Object (semantic) segmentation is the second step to represent high-level information in an image. This task is challenging since we want to classify an image pixel-wise, besides object identification is dependent on the context that the objects are immersed in. All the neural networks that are performing semantic segmentation in this work have the same architecture, although each one has its own set of weights. The main difference between them is the target classes that they use to construct the model. Strictly speaking, each neural network has the goal of segmenting objects within a related scene.

The segmentation neural networks generate a set of class proposals for each pixel in the image. Additionally, similar pixels are grouped together in order to represent objects. In this work we focus on assigning spatial relations between pairs of these segmented objects. In order to define a spatial region, we map pairs of detected objects aiming at establishing a representation that might be expressed in topological terms. Then, another neural network takes place and estimates a spatial preposition from the previous (topological) classification. Each image I_x generates a $C^{I_x}_{\frac{n^{I_x}}{2}}$, where n is the number of objects detected for image I_x and C^{I_x} is the combination of n^{I_x} in pairs.

Preposition estimation: A sparse representation of C^{I_x} is created and is provided as input to a neural network whose task is to assign English prepositions to images. It is important to notice that object information is crucial at this point to disambiguate multiple (possible) assignments. Besides, this sparse representation provides a notion of the spatial topology with respect to a pair of objects. Finally, this sparse vector combined with the estimated spatial preposition enables the creation of an index that can be accessed with a structured query in order to retrieve relevant images. This index is similar to those used by BOVW applications. Thus, images can be retrieved given a feature, which in the present case is a spatial relation.

3.2. Qualitative analysis

The second part of the proposed method aims at improving the semantic knowledge about an image by means of a spatial ontology. In order to represent an image with respect to spatial arrangements a high-level ontology is used along with the numerical information automatically extracted, as described in Section 3.1.

Knowledge representation is built with respect to the spatial arrangement of the identified structures in images. The interpretation of spatial relations contributes to reducing the semantic gap in images since relations tend to suffer less with variations than the objects in their arguments (Hudelot et al., 2008).

Aiming at representing the knowledge acquired through the numerical analysis of images, we have opted to extend an existing ontology: SUMO (Pease, 2011). SUMO has been adopted due to the fact it is built on higher-order logic and thus it provides the flexibility to create complex constructions, not restricted to binary relations.

The initial step to represent knowledge was to create instances of the domain. In this work, each detected object is considered as an unique and independent instance.

Considering an example image I_x with the detected objects *Building*, *Floor* and *Sky*, in this work, each of these terms is identified by the suffix corresponding to the original image, in this case x. Therefore the term *Building* is referenced by $Building_x$ and so on. Formula 1 shows the instance definition.

(instance $Building_x$ Building)
(instance $Floor_x$ Floor)
(instance Sky_x Sky)

Formula 1: Instances

All the images and the detected objects are represented in a similar fashion. The next step of our framework is to construct spatial relations between the distinct objects. To execute the mapping of these relations to the segmented pair of objects, we have considered the prepositions estimated by a statistical classifier. Finally, our preposition domain contains the following relations: *above, across from, behind, below, in, in front of, inside of, left of, on, right of* and *under*.

By using an off-the-shelf ontology, the representation of these relations expressed by prepositions was simplified. In this work, only a punctual extension of SUMO was necessary.

The standard SUMO ontology provides tools to define spatial relations. In order to define a spatial relation in SUMO it is necessary to create an instance of the class *PositionalAttribute*. This class enables the stating of binary orientation relations between two objects. Additionally, it is important to mention that the semantics with respect to spatial relations might be expressed by more than one preposition, besides, distinct prepositions might be similar or complementary. For instance, two distinct objects arranged consecutively might be referenced by the following prepositions : *in front of* or *behind*, the proper term is defined according to the context and the observer's position. This characteristic allows us to represent both relations using the double implication operator <=>. Consequently, it is possible to define the preposition *in front of* based on the preposition *behind*, or vice-versa. This definition is illustrated in Formula 2. It is worth pointing out that, although *in front of* or *behind* are both relative to an observer, this paper assumes that observer is the camera point of view and, thus, it is implicit in the definitions.

(<=>

Table 1: Precision-recall for scene classification.

Scene	Precision	Recall	N
Indoor	0.86	0.84	1829
Outdoor	0.88	0.90	2439
Overall	0.87	0.87	4268

```
(orientation ?X1 ?X2 Behind)
(orientation ?X2 ?X1 InFrontOf))
```

Formula 2: Double implication to *behind* and *In front of*

According to the nature of spatial prepositions, this procedure is applied to other relations, for instance: *under* and *above* or *left of* and *right of*. Additionally, the transitivity of relations was also used in this context whenever possible. For instance, we might infer that if there is an object a above an object b and b is above a third object c, therefore a is above c.

In conclusion, through quantitative methods, qualitative information with respect to the domain was inferred. The impact of this procedure was evaluated in the tests described in the next section .

4. Experiments

This section details the experiments executed in order to evaluate the proposed method. To perform the experiments, the publicly available data set SUN09 (Choi et al., 2010) was used. This data set is composed of more than 12,000 images containing various classes of objects in distinct scenes. In order to conduct the tests, two data sets of annotations for SUN09 were used. First, the data set provided by (Lan et al., 2012) (data set 1) was used, that contains annotations in the form of structured queries ((*noun*, *preposition*, *noun*)) representing two relations *below* and *above*. Second, the annotations provided in (Malinowski and Fritz, 2014) (data set 2) were considered that includes eleven (11) distinct preposition classes: *above, across from, behind, below, in, in front of, inside, left, right, on* and *under*.

Overall 4,367 images were used for training and 4,317 images for testing. In these datasets there are 186,299 pairs of objects for training and 173,111 for testing, being 106 distinct objects considering data set 1 and 42 considering data set 2.

4.1. Scene classification

The scene classification part of the present proposal (identifying what we use as context: *indoor* or *outdoor* scenes) was evaluated using data set 2, with manual annotations for training and evaluation purposes. The results of this binary classification are shown in Table 1, where we can see that the overall precision and recall for each of the considered classes was 87%, attesting for the suitability of the method applied for this task.

As described above, scene identification provides information to refine the object segmentation procedure, whose results are shown below.

Table 2: Intersection over union.

Scene	Object	Intersection Over Union %	
		Context	No context
	Airplane	23.62	20.24
	Armchair	13.70	23.64
	Balcony	1.45	0.00
	Bench	9.67	10.83
	Building	25.02	17.86
	Car	11.39	11.73
	Door	10.95	13.76
	Fence	21.51	20.33
	Gate	0.71	0.00
Outdoor	*Grass*	41.60	21.67
	Path	32.05	16.78
	Road	57.26	37.45
	Rock	15.60	3.28
	Sign	10.66	11.94
	Sky	78.73	71.51
	Streetlight	5.72	6.56
	Tree	21.23	21.11
	Van	17.96	14.75
	Water	59.17	29.92
Average		**24.11**	**18.60**
	Basket	5.28	8.12
	Book	1.48	0.70
	Bookcase	27.68	9.65
	Bottle	0.56	4.60
	Cabinet	16.35	0.00
	Chair	5.41	10.76
	Closet	24.35	26.60
	Curtain	31.36	33.11
	Desk	8.48	13.66
	Floor	60.94	49.30
	Ground	17.94	19.86
Indoor	*Mirror*	33.44	35.25
	Plant	15.48	10,67
	Poster	4.48	9.51
	Refrigerator	57.23	52.53
	Table	16.13	20.74
	Vase	1.46	9.87
	Wall	25.18	22.42
	Window	27.82	0.00
Average		**16.57**	**14.07**
Consolidated		**22.66**	**18.67**

4.2. Object segmentation

In this part of the system evaluation, we investigate the hypothesis of whether the information provided by the scene identification (Section 4.1) improves the segmentation. According to this premise we have manually separated the objects as indoor and outdoor. The indoor objects are: *armchair, basket, bookcase, book, bottle, box, chair, closet, cupboard, curtain, desk, floor, flower, ground, mirror, plant, poster, refrigerator, seats, table, vase, wall and window*. The outdoor objects are: *airplane, balcony, bench, building, car, door, fence, gate, grass, path, road, rock, sign, sky, streetlight, tree, van, water*. According to each scene class detected we apply one or the other segmentation models (i.e. one trained with *indoor* objects or the other trained with *outdoor objects*).

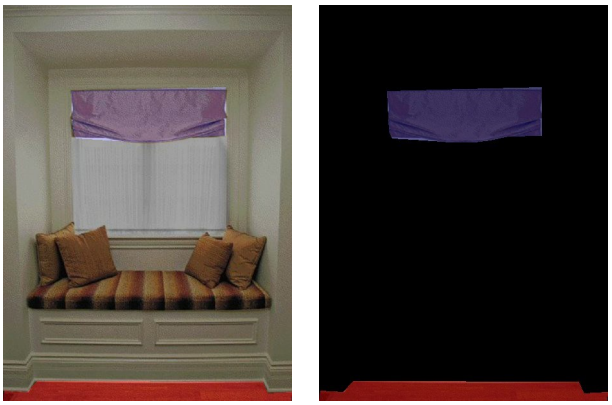
The semantic segmentation model applied in this experiment was fine tunned with the weights provided by (Zhou et al., 2016). The object segmentation results are shown in Table 2. Table 2 shows all the objects in data set 2. In the left-most column is the scene class of which the object belongs to. The sec-

ond column, from left to right, is the noun that represents the object. The following two columns represent the Intersection over union (IoU) considering object recognition using context information or without using it (column "No context"). Results show that the classification using context information had a higher IoU value overall (shown in line "Average" in Table 2). However, there were various cases in which the use of scene class information did not improve the results (such as the results related to *Armchair*, or *Bench*), this occurred due to the fact that such objects appear in both (indoor and outdoor) scenes. Thus, in these cases, assuming scene classes caused a reduction in the number of examples of some objects during the training phase.

Object segmentation is the first step in order to define the semantics of images. In this work, for each pair of segmented objects (in the same image) our aim is to estimate the spatial relations holding between them, and after that, express this spatial relation by means of a preposition, as described in the next section.

4.3. Preposition assignment

The main difficulty of assigning a preposition to a pair of objects in an image is the common overlapping of terms, i.e., there are several possible (consistent) preposition assignments to each spatial relation. In order to cope with this issue, before assigning a preposition, the topological relation between pairs of objects is classified, serving as a bridge to preposition definition. In our experiments, each image in the data set was transformed into a representation containing: *(target object, reference object)*. This representation is shown in Figure 2, Figure 2a presents the original image and the object mask is presented in Figure 2b. It is important to mention the fact that the color of the objects indicates the target object (blue) and the reference object (red).



(a) Original. (b) Object mask.

Fig. 2: Topology representation with highlighted objects.

Each image available in the data set was segmented and combined with its relative object, generating a combination of $C_{Ix} = \frac{N_{Ix}^2}{2}$ for images I^x, N_{Ix} segmented objects, N^x for the image. Every image generated from C_{Ix} was classified according to the spatial preposition related to the reference and target objects. In this test we have used data set 2 with 11 spatial prepositions.

Table 3: Precision, recall and f-measure for the estimated prepositions.

Preposition	precision	recall	f-measure	n
Above	0.76	0.58	0.66	6
Across from	1.00	0.03	0.06	387
Behind	0.65	0.62	0.63	329
Below	0.84	0.79	0.81	361
In	0.59	0.84	0.69	475
In front of	0.55	0.69	0.61	317
Inside of	0.00	0.00	0.00	65
Left of	0.01	0.01	0.01	187
On	0.60	0.77	0.67	208
Right of	0.00	0.00	0.00	59
Under	0.87	0.98	0.93	2399
Overall	0.75	0.75	0.71	4953

Table 3 shows the precision, recall and f-measure for each of the tested prepositions, where the right-most column is the number of tested samples. The overall values are exhibited in the last line: for the 4953 relations tested, the system reached an overall precision of 0.75 with a recall of 0.75 and a f-measure of 0.71.

Table 3 also shows that the prepositions whose definitions depend on a great region overlapping are harder to detect, such as *inside of* (or *across from*), that received very low values for precision and recall. The overlapping in this context is related to the shared area of the detected objects. For instance, *inside of* should contain a complete overlap of at least one of the detected objects to be assigned (which hardly happens). The prepositions *Left of* and *Right of* also received low scores. This is due to the fact that, in situations where they could be assigned, there were other, competing, prepositions (such as *In front of* or *Behind*) that took precedence in the assignment as a result of minor variations in the object's relative positions.

Next section compares the performance of IRRA with other state-of-the-art methods in the task of image retrieval from structured queries.

4.4. Retrieval evaluation

In order to evaluate image retrieval using IRRA the annotations provided by (Lan et al., 2012) were used. To the best of our knowledge, this is the only available data set that maps spatial relations to objects detected in images. We have tested our method against all the structured query types proposed in (Lan et al., 2012). The structured queries contain a *noun*, e.g. *pedestrians*, or a *relation* set expressed by a triple in the form *(noun, preposition, and noun)*, e. g. *"car on the road"*. The available structures are represented as: *Structure a (Sa)*, which contains only a relation set, for instance, *"car on road"*; *Structure b (Sb)* contains a relation set and a noun, e.g, *"car on road, pedestrians"*; *Structure c (Sc)* contains two relation sets, e.g., *"car on road, sky above building"*; *Structure d (Sd)* contains two relation sets and a noun, e.g., *"car on road, sky above building, pedestrians"*; *Structure e (Se)* contains three relation sets, e.g., *"car on road, sky above building, books inside of bookcase"*;

Table 3 illustrates the obtained results and also displays a comparison with other approaches using the same data set. Only recall is presented since this is the measurement used in (Lan et al., 2012).

According to Table 3 it is possible to observe that IRRA outperforms the other methods in all the structured queries considered.

Analyzing IRRA results alone, we can see that in larger queries that do not include single objects (in scenarios *Sc* and *Se*, for instance), IRRA's performance is not as good as in other scenarios. This behavior occurs due the fact that, when answering a query such as *Sc* or *Sc*, segmentation or preposition detection errors are propagated to the retrieval task.

We have also evaluated the retrieval by using the second data set which has a larger set of prepositions. In this setting IRRA achieved a retrieval with mean average precision (mAP) of 53.95, outperforming the recent results reported in (Malinowski and Fritz, 2014). The superior performance of IRRA with respect to other (competing) methods is explained by the fact that IRRA uses the various relations existing in an image in the retrieval task.

4.5. Ontology expansion

The queries executed in the tests above were strictly unidirectional due to the fact that there is no high-level reasoning with respect to the spatial arrangement of the items in the document collection. To address this issue, we investigated the application of reasoning using first-order logic based on an ontology.

We extended the annotations in (Malinowski and Fritz, 2014) with the aim to infer new relations derived from the original relations that were manually annotated. Through the use of the SUMO ontology we extended the system's knowledge about the relations in order to evaluate spatial prepositions that were not obtained by the quantitative analysis processes.

In order to evaluate this method, distinct queries were proposed, but keeping the same information to be retrieved from the set of images. To accomplish this task inverse relations were applied. For instance, for two objects (x, y) and the query (x-above-y), we also evaluated the retrieval for (y-below-x) against the same annotation for above as used in the original query.

The instantiation of the whole ontology with SUN09 data generated more than $13K$ terms with respect to the images and objects, and more than $18K$ formulas referencing the created relations. The retrieval task based on the new set of queries was performed by evaluating every image using the E first-order logic theorem prover (Schulz, 2013). Although SUMO is defined in a higher-order logic, we were able to achieve our goals with just the first-order logic content of the theory in this work, which allowed us to use a first-order logic prover.

Every image was tested using the new annotated queries, the mAp achieved in this case was 41.60, whereas he original set for these prepositions obtained a mAp of 51. The decrease in performance observed with the extended set of relations (in contrast with the original) was due to the fact that, by increasing the size of the knowledge base, errors were possibly included in the process, whose detection becomes increasingly complex to perform (Pease, 2011).

5. Discussion

This paper proposed a framework for image retrieval based on spatial relations. The proposed method decomposes the image in distinct levels in order to detect objects and also the spatial relations between them. The initial step is to detect the scene class that the object is inserted, this procedure has shown an increase in 4% of the IoU metric over the method without considering the scene class.

The preposition detection, the last part of quantitative analysis, did not perform as well as expected with respect to the prepositions: *across from, inside of, left of, right of*. The problem is originated by the multiple possible distinct prepositions that could be applied to the scenes related to these cases. A deeper investigation of how to represent the object pairs in order to enhance the estimation of their spatial preposition is a task for future work. The creation of data sets aiming at this task should motivate developments in this area. Future work also should focus on improving the results of the related neural networks in order to avoid errors, mainly false positive propagation problems. It is also within our future interests to augment the set of relations covered by the system, relaxing the present restriction to binary relations.

6. Conclusion

This work investigated the semantic gap that exists in content-based image retrieval by introducing an approach that establishes relations between objects in images by means of spatial arrangements. The method proposed in this paper, called Image Retrieval using Region Analysis (IRRA), starts by decomposing images with respect to pairs of objects, where each pair is also combined with a spatial relation. Each spatial relation is related to a spatial preposition expressed in natural language. IRRA was evaluated on a public data set, whose results show that our approach outperforms previous (recent) work in the retrieval of images using spatial relations.

We believe that the proposed framework has two compelling applications. The first is to improve statistical classifiers, following an approach similar to (Chen et al., 2014), where new samples are evaluated before inserting them into the knowledge base. The second is the possibility to include abstractions (in terms of high-level relations, spatial or not) to static data sets in order to enhance image retrieval tasks.

Acknowledgements

Danilo Nunes was partially supported by CAPES. Paulo Santos acknowledges support from FAPESP-IBM (grant 2016/18792-9) and CNPq (grant 307093/2014-0).

References

Bloch, I., Hudelot, C., Atif, J., 2007. On the interest of spatial relations and fuzzy representations for ontology-based image interpretation., in: Proceedings of the 7th International Conference on Advances in Pattern Recognition, ICAPR, Kolkata, India. pp. 15–25.

Chen, X., Shrivastava, A., Gupta, A., 2014. Enriching visual knowledge bases via object discovery and segmentation. Computer Vision and Pattern Recognition .

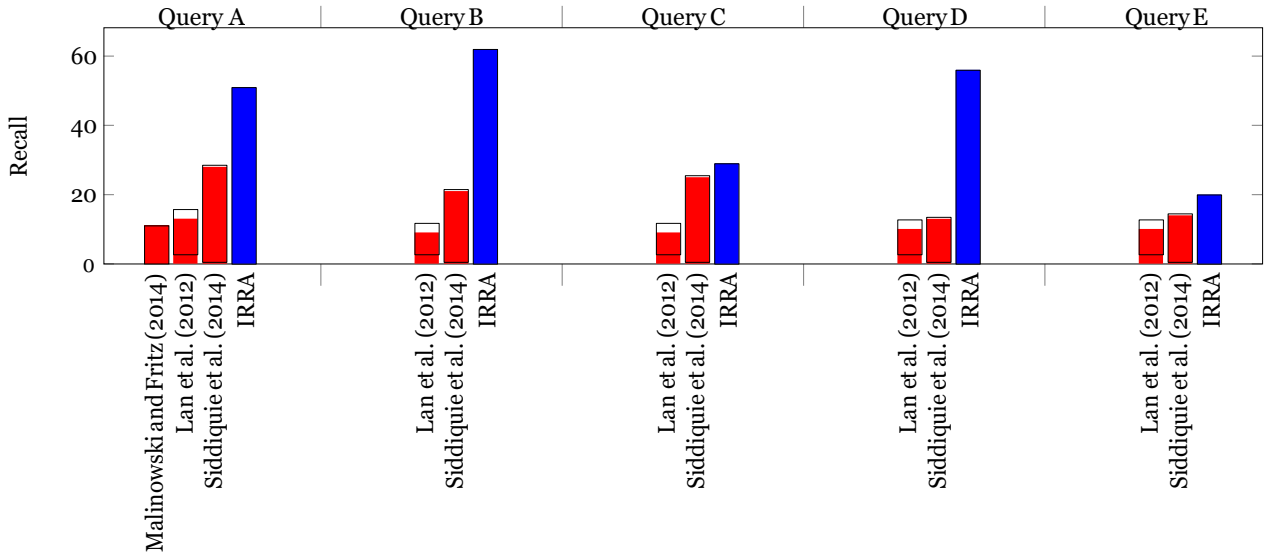


Fig. 3: Comparison with other approaches.

Choi, M.J., Lim, J.J., Torralba, A., Willsky, A.S., 2010. Exploiting hierarchical context on a large database of object categories, in: 2010 IEEE Comp. Soc. Conference on Computer Vision and Pattern Recognition, pp. 129–136.

Dai, B., Zhang, Y., Lin, D., 2017. Detecting visual relationships with deep relational networks, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proc. of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, Washington, DC, USA. pp. 580–587.

Gould, S., Fulton, R., Koller, D., 2009. Decomposing a scene into geometric and semantically consistent regions, in: 2009 IEEE 12th International Conference on Computer Vision, pp. 1–8.

Hare, J.S., Sinclair, P., Lewis, P.H., Martinez, K., Enser, P.G., Sandom, C.J., 2006. Bridging the semantic gap in multimedia information retrieval: Top-down and bottom-up approaches, in: 3rd Euro. Semantic Web Conference.

Hudelot, C., Atif, J., Bloch, I., 2008. Fuzzy spatial relation ontology for image interpretation. Fuzzy Sets Syst. 159, 1929–1951.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Proc. of the 25th International Conference on Neural Information Processing Systems, USA. pp. 1097–1105.

Lan, T., Yang, W., Wang, Y., Mori, G., 2012. Image retrieval with structured object queries using latent ranking svm, in: Proceedings of the 12th European Conference on Computer Vision - Volume Part VI, Springer-Verlag, Berlin, Heidelberg. pp. 129–142.

Lehmann, J., Vöelker, J., 2014. An introduction to ontology learning, in: Lehmann, J., Vöelker, J. (Eds.), Perspectives on Ontology Learning. AKA Heidelberg, pp. ix–xvi.

Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L., 2016. Visual relationship detection with language priors, in: European Conference on Computer Vision.

Mai, L., Jin, H., Lin, Z., Fang, C., Brandt, J., Liu, F., 2017. Spatial-semantic image search by visual feature synthesis, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Malinowski, M., Fritz, M., 2014. A pooling approach to modelling spatial relations for image retrieval and annotation. arXiv:1411.5190 [cs.CV] .

Niles, I., Pease, A., 2001. Towards a standard upper ontology, in: Proceedings of the International Conference on Formal Ontology in Information Systems - Volume 2001, ACM, New York, NY, USA. pp. 2–9.

Pease, A., 2011. Ontology: A Practical Guide. Articulate Software Press, Angwin, CA.

Philbin, J., Sivic, J., Zisserman, A., 2008. Object mining using a matching graph on very large image collections, in: Proceedings of the 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing, IEEE Computer Society, Washington, DC, USA. pp. 738–745.

Sankat, M., Thakur, R.S., Jaloree, S., 2016. Semi-Automatic Ontology Design for Educational Purposes. IGI Global, Hershey, PA, USA. pp. 124–142.

Schulz, S., 2013. System Description: E 1.8, in: McMillan, K., Middeldorp, A., Voronkov, A. (Eds.), Proc. of the 19th LPAR, Stellenbosch, Springer.

Siddiquie, B., White, B., Sharma, A., Davis, L.S., 2014. Multi-modal image retrieval for complex queries using small codes, in: Proc. of International Conference on Multimedia Retrieval, ACM, New York, NY, USA. pp. 321–328.

Socher, R., Lin, C.C.Y., Ng, A.Y., Manning, C.D., 2011. Parsing natural scenes and natural language with recursive neural networks, in: Proceedings of the 28th International Conference on International Conference on Machine Learning, Omnipress, USA. pp. 129–136.

Wan, J., Wang, D., Hoi, S.C.H., Wu, P., Zhu, J., Zhang, Y., Li, J., 2014. Deep learning for content-based image retrieval: A comprehensive study, in: Proceedings of the 22Nd ACM International Conference on Multimedia, ACM, New York, NY, USA. pp. 157–166.

Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H.S., 2015. Conditional random fields as recurrent neural networks, in: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), IEEE Computer Society, Washington, DC, USA. pp. 1529–1537.

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A., 2016. Semantic understanding of scenes through the ade20k dataset, in: arXiv preprint arXiv:1608.05442.