Matheus Vieira Lessa Ribeiro

Proposta de Local Binary Pattern Coerente e Incoerente na Categorização de Cenas

Vitória

2017

Matheus Vieira Lessa Ribeiro

Proposta de Local Binary Pattern Coerente e Incoerente na Categorização de Cenas

Dissertação apresentada ao programa de Pós-graduação em Engenharia Elétrica da Universidade Federal do Espírito Santo como requisito parcial para a obtenção do grau de Mestre em Engenharia Elétrica.

Universidade Federal do Espírito Santo – UFES

Departamento de Engenharia Elétrica

Programa de Pós-Graduação em Engenharia Elétrica

Orientador: Prof. Dr. Evandro Ottoni Teatini Salles

Vitória

2017

Dados Internacionais de Catalogação-na-publicação (CIP)
(Biblioteca Setorial Tecnológica,
Universidade Federal do Espírito Santo, ES, Brasil)
Sandra Mara Borges Campos – CRB-6 ES-000593/O

Ribeiro, Matheus Vieira Lessa, 1989-
R484p Proposta de Local Binary Pattern coerente e incoerente na categorização de cenas / Matheus Vieira Lessa Ribeiro. – 2017.
85 f. : il.

Orientador: Evandro Ottoni Teatini Salles.
Dissertação (Mestrado em Engenharia Elétrica) – Universidade Federal do Espírito Santo, Centro Tecnológico.

1. Visão por computador. 2. Algoritmos. 3. Classificação de cenas. 4. Padrões binários locais (LBP). 5. Algoritmo Color Coherent Vector (CCV). I. Salles, Evandro Ottoni Teatini. II. Universidade Federal do Espírito Santo. Centro Tecnológico. III. Título.

CDU: 621.3

Matheus Vieira Lessa Ribeiro

Proposta de Local Binary Pattern Coerente e Incoerente na Categorização de Cenas

Dissertação apresentada ao programa de Pós-graduação em Engenharia Elétrica da Universidade Federal do Espírito Santo como requisito parcial para a obtenção do grau de Mestre em Engenharia Elétrica.

Trabalho aprovado. Vitória, 11 de outubro de 2017:

Prof. Dr. Evandro Ottoni Teatini Salles
Universidade Federal do Espírito Santo

Profa. Dra. Aura Conci
Universidade Federal Fluminense

Profa. Dra. Kelly Assis de Souza Gazolli
Instituto Federal do Espírito Santo - *Campus* Serra

Vitória
2017

Agradecimentos

Agradeço a Deus pelo dom da vida e por me ensinar cada dia mais sobre o verdadeiro Amor e à Virgem Maria Santíssima pela grandiosa proteção e intercessão nos momentos de dificuldade.

Aos meus pais por terem paciência e me perguntarem só duas vezes por mês quando eu terminaria o mestrado. Agradeço a eles também por terem me dado toda as condições para que eu realizasse esta pós-graduação e pelo incentivo em querer que eu conquiste meus sonhos.

Ao meu orientador Evandro por não ser somente um educador com imensa disponibilidade em me explicar pacientemente todo tipo de questão, como também um psicólogo e amigo, me ajudando e me compreendendo nos momentos difíceis em que estivemos trabalhando juntos.

À minha namorada Barbara pela paciência e compreensão por entender muitas vezes que passaríamos pouco tempo juntos em virtude desta pesquisa, sobretudo nos finais de semana. Agradeço também ao enorme carinho e amor que me impulsionaram a sempre acreditar ainda mais nos meus objetivos.

Aos meus amigos do Cisne, pelos momentos de descontração, desabafos, estudos e companheirismo que fizeram com que esse período que estive na Ufes fosse de grande valia e de grande aprendizado, cada um com sua característica me inspirou e me ensinou algo em especial. Ao Gabriel por sua sabedoria, Douglas por sua dedicação e esforço, Thaís por sua motivação, determinação e empatia, Bruno pelo companheirismo (sobretudo nos trabalhos da minha graduação), Thalles pelas inesquecíveis conversas no RU e pelos desafios que acabavam com as tardes de produtividade do laboratório, Daniel por sua paciência, Fernando pela sua inteligência e disponibilidade em me ajudar, Demuth pela sua responsabilidade e carinho, e ao meu novo amigo Messi por seu bom humor e alegria.

À Capes pela oportunidade da bolsa concedida.

À Ufes pela estrutura e pelos profissionais que nela estão e me ajudaram bastante, em especial ao Patrick e à Raquel.

"O amor é a única força capaz de mudar o mundo."
(Papa Bento XVI)

Resumo

Este trabalho propõe um novo descritor visual de cenas a partir da técnica *Local Binary Pattern* (LBP) e explorando a informação espacial utilizando o algoritmo *Color Coherent Vector* (CCV).

O LBP se caracteriza por ser uma técnica não linear e não paramétrica, dispensando conceitos intermediários no processo de descrição da imagem, tornando uma alternativa para usuários leigos com pouco conhecimento na área. Já a representação CCV mostrou ser uma técnica que busca mitigar o problema da falta de informação espacial pelos histogramas, expressando a imagem em pixeis coerentes e pixeis incoerentes sem que aumente a dimensionalidade dos dados.

Nesse sentido, uma primeira abordagem foi a proposta das técnicas LBP Incoerente e LBP Coerente na classificação de cenas. Resultados preliminares, empregando-se k-nn como classificador, demonstraram que o LBP Incoerente apresenta um bom compromisso entre acurácia e dimensão de representação dos dados.

Em seguida, no intuito de se incluir o conceito de contexto, para mitigar o problema da localidade do LBP, foi proposto o *Contextual Modified Local Binary Pattern* Incoerente (CMLBP Incoerente), que modela a distribuição das estruturas locais através do LBP, adicionando informação contextual, inspirado no algoritmo *Contextual Modified Census Transform* (CMCT). Entre outras características, o CMLBP Incoerente demonstrou capacidade em descartar regiões homogêneas, representadas pelos pixeis coerentes através do algoritmo CCV.

Em experimentos realizados com bancos de dados consagrados na literatura, o CMLBP apresentou resultados melhores que as técnicas originais que não descartam os pixeis coerentes, em quase todas as situações. Para cenas com muitos detalhes e informações os resultados foram satisfatórios e com um maior destaque, superando técnicas conhecidas na literatura. Os resultados obtidos foram encorajadores para a busca de um descritor com boa capacidade discriminante e baixa dimensionalidade na representação de imagens.

Palavras-chave: *Local Binary Pattern*, *Color Coherent Vector*, Classificação de Cenas, Visão Computacional.

Abstract

This master dissertation proposes a new visual descriptor of scenes from Local Binary Pattern (LBP) and exploring a spatial information utilizing the Color Coherent Vector (CCV) algorithm.

LBP is characterized by a non-linear and non-parametric technique, it does not use intermediate concepts in the process of image description, becoming an alternative for lay users with low knowledge in this area. In contrast, CCV has proved to be a safe technique in minimizing the problem of lack of information by histograms, it expresses an image by coherent and incoherent pixels with no needs to improve the dimensionality of data.

In this way, a first approach was the proposal of LBP Incoherent e LBP Coherent techniques in the scene classification. Preliminary outcomes, with k-nn classifier, indicated that LBP Incoherent performs a good compromise between accuracy and dimension of data representation.

Afterwards, with the purpose to including the concept of context, to minimizing the problem of location from LBP, the Contextual Modified Local Binary Pattern Incoherent (CMLBP Incoherent) was proposed, which models the distributions of local structures through LBP, by adding contextual information, inspired in Contextual Modified Census Transform (CMCT). The CMLBP Incoherent, among others characteristics, has demonstrated competency in discarding homogeneous regions, represented by coherent pixels, through CCV algorithm.

In experiments carried with important datasets in literature, CMLBP achieved better results than original techniques which do not discard the coherent pixels, in almost all over tests. For scenes with much detail and information the results were satisfactory and better than the results of known techniques in the literature. The results obtained by CMLBP Incoherent has been encouraging to finding of a descriptor with good discriminant performance and low dimensionality in image representation.

Keywords: Local Binary Pattern, Color Coherent Vector, Scene Classification, Computer Vision.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

BoW	*Bag of Words* - Bolsa de Palavras Visuais
CBIR	*Content Based Image Retrieval* - Recuperação de Imagem Baseado no Conteúdo
CBoW	*Contextual Bag of Words* - Bolsa de Palavras Visuais Contextuais
CCH	*Cell Color Histogram* - Histograma de Células de Cor
CCV	*Color Coherent Vector* - Vetor de Cores Coerentes
CENTRIST	*Census Transform Histogram* - Transformada Census
CMCT	*Contextual Modified Census Transform* - Transformada Census Modificada Contextual
CMLBP	*Contextual Modified Local Binary Pattern* - Padrão Binário Local Modificada Contextual
COV	*Concept Co-occurrence Vector* - Vetor de Co-ocorrência de Conceitos
CT	*Census Transform* - Transformada Census
EDCV	*Edge Direction Coherence Vector* - Vetores Coerentes de Direção de Bordas
GLCM	*Gray Level Co-occurrence Matrix* - Matriz de Co-ocorrência de Níveis de Cinza
GPS	*Global Positioning System* - Sistema de Posicionamento Global
HLBPH	*Haar Local Binary Pattern Histogram* - Histograma de Padrão Binário Local Haar
HOG	*Histogram Oriented Gradients* - Histograma de Orientação ao Gradiente
HSV	*Hue, Saturation and Value* - Cor, Saturação e Valor
JSEG	- J *Segmentation* - J Segmentação
K-NN	*K Nearest Neighbor* - K Vizinhos Mais Próximos
LBP	*Local Binary Pattern* - Padrão Binário Local

LBPHF	*Local Binary Pattern Histogram Fourier* - Padrão Binário Local e Histograma de Fourier
LBP^{riu}	*Local Binary Patter Rotation Invariant Uniform* - Padrão Binário Local Uniforme Invariante à Rotação
LC_1C_2	*Luminance, Chromatic 1, Chromatic 2* -Luminância, Cromática 1, Cromática 2
LDA	*Latent Dirichlet Allocation* - Alocação Latente Dirichlet
MCT	*Modified Census Transform* - Transformada Census Modificada
MCT8	MCT com 8 Bits
Ncut	*Normalized Cut* - Normalização Cut
PCA	*Principal Component Analysis* - Análise das Componentes Principais
PLSA	*Probabilistic Latent Semantic Analysis* - Análise Probabilística Latente Semântica
RCVW	*Region Contextual Visual Words* - Palavras Visuais Contextuais por Região
RGB	*Red, Green and Blue* - Vermelho, Verde e Azul
SIFT	*Scale Invariant Feature Transform* - Transformada de Características Invariantes à Escala
SPM	*Spatial Pyramid Matching* - Correspondência Espacial em Pirâmide
SUN	*Scene Understanding* - Entendimento de Cenas
SVM	*Support Vector Machine* - Máquinas de Vetores de Suporte

Sumário

1 Introdução

Com a crescente evolução da tecnologia digital, a utilização de aparelhos celulares se tornou mais popular. Segundo o jornal The Independent, da Inglaterra, em 2014 o número de aparelhos móveis ativos no mundo ultrapassou o número de pessoas. Outro dado a ser considerado é que estes aparelhos crescem a uma taxa cinco vezes maior que a raça humana [1].

Desse modo, os celulares deixaram de ser apenas uma ferramenta capaz de se comunicar com outra pessoa realizando uma chamada ou mandando mensagem de texto. Usuários comuns estão procurando dispositivos móveis capazes de oferecer vários recursos como câmeras, operações financeiras, jogos, GPS *(Global Positioning System)*, transmissão de dados, entre outros.

Não obstante, aliado ao uso crescente das redes sociais, pode-se destacar a utilização de diversos aplicativos para estes dispositivos móveis como Facebook, Instagram e Snap. Nesses casos, o usuário pode interagir através de fotografias, seja para registrar um determinado momento ou para divulgar um ambiente em que achou bonito. Uma consequência deste fenômeno foi a popularização da palavra *"selfie"*, que em 2013 se tornou a palavra do ano para o Oxford Dictionary [2].

Com tantos avanços no armazenamento de dados e na aquisição de imagens, o número de cenas disponibilizadas na internet cresceu consideravelmente, tornando a anotação manual do conteúdo de cada imagem uma tarefa laboriosa. A informação do conteúdo de uma cena pode ser importante para inúmeras atividades relacionadas ao processamento de imagens, dentre elas é possível citar a busca por imagens semelhantes ou por uma determinada classe, além de operações específicas, como balanceamento de uma cor, por exemplo, caso já tenha-se um conhecimento prévio da cena.

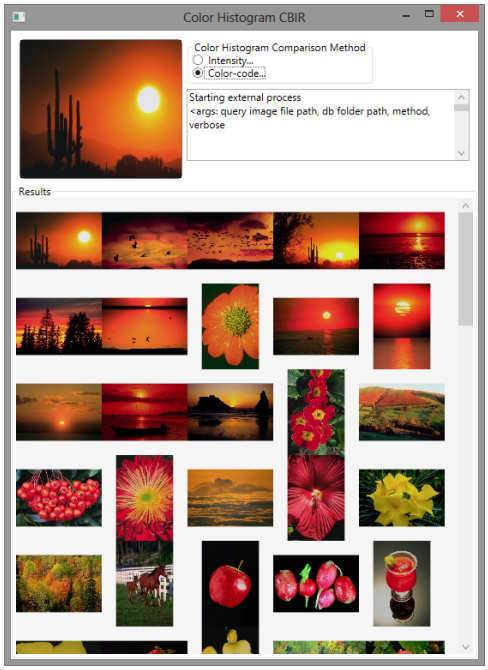
Há na internet alguns programas capazes de recuperar uma cena ou encontrar imagens com *layout* parecido, dentre eles pode-se citar: Photobook (PENTLAND; PICARD; SCLAROFF, 1996) , Visualseek (SMITH; CHANG, 1997), Netra (MA; MANJUNATH, 1999), dentre outros. Um *layout* é a disposição espacial de informações como cor, textura e linhas em uma imagem (VAILAYA; JAIN; ZHANG, 1998). A Figura 1 mostra uma aplicação denominada *Content Based Image Retrieval* (CBIR) capaz de recuperar imagens parecidas com o conteúdo da imagem de referencia [3]. Neste exemplo se considera apenas a

[1] http://www.independent.co.uk/life-style/gadgets-and-tech/news/there-are-officially-more-mobile-devices-than-people-in-the-world-9780518.html acessado em 20 de agosto de 2017

[2] http://newsfeed.time.com/2013/11/18/and-oxfords-word-of-the-year-is/ acessado em 20 de agosto de 2017

[3] http://pckujawa.github.io/portfolio/mm-cbir/ acessado em 25 de agosto de 2017.

Figura 1 – Software de recuperação de imagens.



Fonte: pckujawa.github.io

informação de cor. Observe que, no primeiro momento, as sugestões dadas pelo CBIR são as imagens que tiveram maior grau de correspondência no algoritmo e possuem a mesma classificação que a imagem de referência.

A classificação de cenas também pode ser importante no campo da robótica modelando quais atividades um determinado robô poderá realizar, já que dependendo do ambiente em que ele esteja inserido algumas ações poderão ser reprovadas, como ocorre com os seres humanos. O comportamento humano, como gestos e posturas, em uma praia é diferente de quando se está assistindo uma peça de teatro, por exemplo. Além disso, a busca por um determinado objeto pode ser facilitada sabendo a classe da imagem em que o robô esteja visualizando.

1.1 Caracterização do Problema

Classificação de cenas tem sido uma tarefa importante e desafiadora no campo da robótica e da visão computacional. Há várias aplicações nesta área, desde recuperação de imagens e balanceamento de cores até navegação autônoma de robôs. O principal objetivo da classificação de cenas é rotular automaticamente uma imagem, de acordo com a sua distribuição semântica, através de um conhecimento prévio adquirido e de um determinado banco de dados (CHU; ZHAO, 2014).

Pode-se considerar que uma cena é uma visão real de um ambiente, com múltiplas superfícies podendo conter os mais variados tipos de texturas e cores, com objetos dispostos de uma maneira organizada e em escalas aceitáveis (HOLLINGWORTH; HENDERSON, 1999). A Figura 2 ilustra um quadro no qual não pode ser considerado um tipo de cena em virtude da má distribuição dos seus objetos e de suas diferentes escalas na imagem [4].

Figura 2 – Exemplo de imagem não considerada como cena (Personal Values-Rene Magritte).

Fonte: www.renemagritte.org

Na literatura é possível encontrar diversos algoritmos de diferentes níveis de complexidade e abordagens. Entretanto, não obstante os avanços e o grande esforço na obtenção de melhores resultados, a classificação de cenas ainda se depara com muitos problemas. A principal razão deve-se à grande variabilidade entre as cenas. Uma única classe pode conter imagens com variações de iluminação, escala, oclusões e pontos de vistas diferentes, fazendo com que esta tarefa se torne ainda mais árdua e desafiadora.

Esses fatores geram nas cenas uma grande variabilidade intraclasse e uma pequena variabilidade interclasse, dificultando a etapa de classificação. Variabilidade intraclasse é a variação que duas imagens de uma mesma classe podem ter. A Figura 3 ressalta esta situação com quatro cenas da categoria Floresta. Repare que apesar dessas cenas estarem todas na mesma classe, elas possuem variações na escala, iluminação e distribuição de cor por exemplo.

[4] www.renemagritte.org/personal-values.jsp acessado em 28 de agosto de 2017

Figura 3 – Cenas da classe "Floresta".

Fonte: Oliva e Torralba (2001).

Já variabilidade interclasse é a variação que duas cenas de classes diferentes podem ter. Na Figura 4 há duas cenas de classes diferentes, na Figura 4.a é retirada uma cena da classe Rua enquanto que a cena da Figura 4.b é classificada como Estrada. Note que apesar de estarem rotuladas em classes diferentes, estas duas cenas são bastante parecidas em seu conteúdo.

Figura 4 – Cenas de classes diferentes.

(a) Rua (b) Estrada

Fonte: Oliva e Torralba (2001)

Embora para os seres humanos estas dificuldades podem muitas vezes até passar despercebidas durante a classificação de uma cena, pois é possível identificar uma grande quantidade de informação em apenas um olhar, para as máquinas esta tarefa pode ser um tanto quanto custosa.

A Figura 5 apresenta diversos exemplos de um mesmo animal representados na Figura 5.a. Estas imagens representam tradicionalmente o banco de dados de uma pessoa,

ou seja, os registros que o cérebro possui de diversos tipos de cachorros. Já na Figura 5.b há um exemplo deste mesmo animal porém com um aspecto diferente dos outros animais representados na figura ao lado. O fato de ser uma figura com características diferentes daquelas já identificadas na base de dados não atrapalha a classificação deste animal na mesma categoria que os outros por parte dos humanos. Entretanto, para uma técnica treinada com imagens segundo a Figura 5.a, quando o algoritmo se depara com uma imagem como a Figura 5.b pode ter alguma dificuldade no reconhecimento.

Figura 5 – Exemplos diferentes de "cachorros".

(a) Base de dados (b) Amostra de teste

Fonte: Próprio autor

Consequentemente, pesquisas baseadas na percepção humana de uma cena ou objeto estão cada vez mais frequentes para modelar e inspirar novos algoritmos computacionais (JIANG et al., 2010).

1.2 Objetivos

Este trabalho tem como objetivo elaborar um novo descritor visual para classificação de cenas que tenha um bom compromisso entre acurácia e a dimensão de representação dos dados. Deseja-se efetuar operações de baixo custo computacional, com poucos parâmetros a serem estimados e, ainda assim, obter resultados competitivos com os já existentes na literatura. Para isso, serão feitas operações utilizado a informação de textura proposta pelo algoritmo *Local Binary Pattern* (LBP).

Propõe-se analisar o comportamento dos resultados após acrescentar informações espaciais na imagem gerada pelo LBP, descartando as regiões homogêneas. Além disso, será investigado se o acréscimo de informação de contexto implica no aumento da acurácia do sistema.

1.3 Trabalhos Relacionados

Devido à grande facilidade com que os seres humanos classificam uma cena, ao longo das décadas foram feitos inúmeros estudos baseados em como é a percepção humana diante dessa situação, a fim de inspirar novos algoritmos computacionais mais robustos. Características como o tempo de resposta e quais as informações que são extraídas no primeiro momento são as principais questões que buscam ser respondidas com esses estudos (OLIVA; TORRALBA, 2001).

Em (POTTER, 1975) foram realizados uma série de estudos para verificar o tempo de resposta mínimo que um humano precisa para identificar uma cena. Neste trabalho foram apresentadas sucessivas imagens de cenas em um determinado tempo. Foi observado que quando o usuário conhece previamente as características de um determinado alvo presente na imagem, ele detecta este alvo mais rapidamente, com cerca de 133 ms. Porém, uma preocupação com esse estudo é que ele não traz muita compreensão a respeito das características globais de uma cena visto que a percepção é baseada em objetos descritos na imagem.

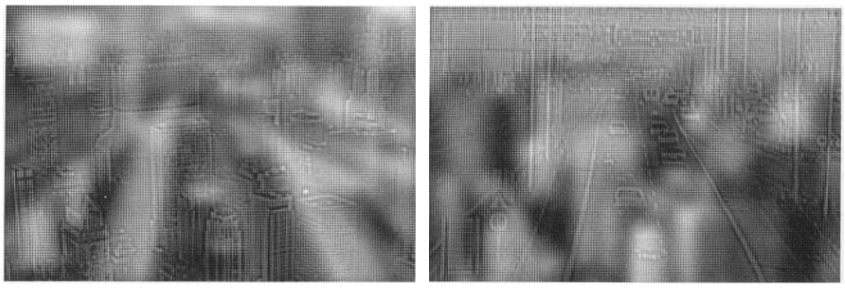
Schyns e Oliva (1994) provaram que para o ser humano identificar uma específica classe solicitada é preciso 45 a 135 ms de tempo de exposição da imagem. Estes resultados demonstraram que os seres humanos conseguem extrair informações rápidas da cena. Em um dos experimentos realizados, apresentou-se uma imagem para os pacientes e eles deveriam identificar se esta imagem pertenceria a uma determinada classe em um estimado tempo de exposição. Também foram utilizadas imagens híbridas em dois experimentos para analisar a informação de frequência na classificação das cenas.

Uma imagem híbrida pode ser vista na Figura 6, nela tem-se informações de baixa frequência de uma cena, representadas pelos borramentos e cores, e informações de alta frequência de outra cena, representadas pelas bordas de outra cena. A conclusão que Olivia e Schyns chegaram foi a de que, em uma rápida visualização, as informações de baixa frequência são mais perceptivas. Enquanto que, à medida que o tempo de exposição aumenta, as informações de alta frequência colaboram mais. Pode-se comprovar esse fenômeno fazendo um teste ao piscar o olho rapidamente sucessivas vezes olhando para a Figura 6. Perceba que as informações de borramento são mais visíveis inicialmente, enquanto que ao parar de piscar, as informações de bordas ficam mais nítidas.

Em (OLIVA; SCHYNS, 2000) investiga-se a contribuição que a cor pode proporcionar na identificação de uma cena rapidamente e conclui-se que humanos podem reconhecer, com o mesmo desempenho, cenas coloridas com até metade da resolução em relação às cenas em nível de cinza. Foi observado também que, em todas as classes, os pacientes obtiveram um tempo de resposta maior quando as cores eram alteradas na imagem.

Baseados nessas observações, em um primeiro momento, os algoritmos computa-

Figura 6 – Exemplo de imagens híbridas entre duas cenas.



Fonte: Schyns e Oliva (1994)

cionais para classificação de cenas foram voltados em extrair características globais da cena, tratando a imagem como um único objeto, com ferramentas de baixo nível como cores e histogramas de bordas. Estas características tem como ponto positivo o baixo custo computacional e poucos parâmetros a serem determinados pelo usuário, porém a performance pode ser limitada (ZOU et al., 2016).

Considerado como uma das implementações com menor custo computacional encontrada na literatura para representar as imagens, os histogramas de informações de baixo nível, como cores e bordas, tem sido, ainda nos dias de hoje, utilizado para classificação de cenas (ZHANG et al., 2016).

Renninger e Malik (RENNINGER; MALIK, 2004) concluíram que para exposição de cenas em curta duração, as características de textura podem ser uma boa alternativa para classificação. Nesse trabalho, são feitas várias comparações com o estudo de (POTTER, 1975), sobre o tempo que um ser humano necessita para identificar uma cena ou reconhecer a ausência ou presença de um determinado objeto. Para a etapa de classificação, cada imagem é representada por um histograma de textura, construído através do filtro de Gabor e da diferença de Gaussianas.

Yiu (YIU, 1996) investigou o uso de propriedades globais como cor e textura para a classificação de uma cena em *indoor* ou *outdoor*. Para evitar o problema da constância de cor, no qual a cor pode mudar suas propriedades de acordo com a iluminação, utilizou-se o modelo de cor LC_1C_2 (*Luminance, Chromatic 1, Chromatic 2*)baseado no sistema visual humano, proposto por Lee (LEE, 1992), em que temos uma componente L responsável pela iluminação e as outras duas componentes responsáveis pela cromaticidade. Para não utilizar todos os níveis de cores dispostos e assim ter um histograma de cor demasiadamente grande, Yiu quantizou o espaço de cores em 34 níveis. Para informação de textura foi utilizado o modelo proposto em (GORKANI; PICARD, 1994). Ao final empregou-se uma técnica para redução de dimensionalidade conhecida como Análise de Componente Principal *(Principal Component Analysis* - PCA).

As informações de frequência também podem ser úteis quando é necessário descrever

a cena como um todo, sem o uso de segmentação ou divisões na imagem. O estudo de (TORRALBA; OLIVA, 2003) discute como o espectro de frequência de uma imagem pode estar relacionado diretamente com o tipo de cena e com sua escala. Os autores ressaltam que sabendo previamente o conteúdo de uma cena, pode-se identificar mais rapidamente a presença ou ausência de um determinado objeto. Neste artigo, as cenas são separadas em dois tipos: as que possuem estruturas construídas pelo homem, denominadas de cenas artificiais e as cenas compostas, em sua grande maioria, por estruturas da natureza, denominadas de cenas naturais.

Observou-se para esses dois tipos de cena que o comportamento do espectro de frequência da imagem varia conforme se modifica a escala também. A escala de uma cena é definida por (TORRALBA; OLIVA, 2003) como a distância entre o observador e os principais elementos que compõe essa cena.

Cecchi et al. (2010) investigaram a correlação espacial entre informações de cor e de orientação de bordas nas imagens naturais. Notou-se que as informações de linhas decaem rapidamente quando o observador se distancia da cena, enquanto que a informação de cor não se altera tanto. Portanto, esses dois canais precisam estar separados. Os resultados alcançados por meio de propriedades estatísticas da imagem estão de acordo com o sistema biológico humano uma vez que os neurônios responsáveis pela cor possuem campos receptivos maiores do que os campos receptivos dos neurônios responsáveis pela percepção da orientação das bordas (CECCHI et al., 2010).

Shimazaki e Nagao (2013) empregaram informações de cor e borda separadamente para classificar alguns tipos de cenas. Posteriormente, com uma ajuda de um classificador *Adaboost* combinaram essas duas características, colocando determinado peso para cada uma delas. A classificação foi feita de modo binário, rotulando se determinada imagem pertencia ou não à uma classe específica. Para informações de borda aplicou-se a transformada Hough (DUDA; HART, 1972), uma transformada que detecta a direção das principais linhas da imagem. A motivação principal se deu pelo fato de que uma cena que possui muitos componentes naturais obtém muitas linhas horizontais, ao contrario das cenas que possuem grande conteúdo composto por estruturas artificiais como prédios e casas.

É importante destacar que os métodos utilizando informações de baixo nível são mais indicados quando existem poucas classes a serem separadas. Não obstante, as imagens preferencialmente devem ter algumas particularidades como pequena quantidade de objetos e pouca variação de cores e orientações de bordas (GORKANI; PICARD, 1994). Desse modo, para representar uma imagem é mais comum a utilização de regiões de interesse para extrair suas principais características. Estas regiões podem ser retangulares ou circulares, divididas na imagem, sobrepostas uma às outras ou envoltas a um ponto de interesse (AKBAS; AHUJA, 2010). Um algoritmo bastante conhecido na classificação de cenas que usa regiões de interesse para extrair características locais é o *Scale Invariant Feature*

Transform (SIFT) (LOWE, 2004).

Através de vários estudos envolvendo a percepção humana, Oliva e Torralba (OLIVA; TORRALBA, 2001) propuseram uma abordagem holística baseada em algumas dimensões perceptuais da cena como naturalidade, abertura, irregularidade, expansão e rugosidade, formando o chamando "envelope espacial". Tais dimensões foram estimadas a partir de técnicas estatísticas utilizando informações de bordas e cor. Através destas dimensões é gerado um espaço multidimensional e as cenas da mesma classe são então agrupadas nesse espaço. Estas dimensões formam o *gist* (essência) da cena, ou seja, é a informação significativa que uma pessoa extrai através de uma rápida visualização (OLIVA; TORRALBA, 2001).

As abordagens holísticas, como o *gist*, tratam acena como um único objeto e extraem dela suas principais características, desconsiderando eventuais detalhes e objetos pertencentes na cena. A fim de adicionar informações locais nas abordagens globais de uma cena, muitos trabalhos passaram a dividi-la em blocos, ou *grids*, e a partir daí aplicar o algoritmo em cada bloco. Ao final, os vetores gerados de cada bloco podem ser concatenados e, dependendo da aplicação, realiza-se uma técnica de redução de dimensionalidade deste vetor final (CHEN; LI; WANG, 2006).

Lazebnik, Schmid e Ponce (2006) propuseram uma técnica de repartição da imagem, em que a quantidade de blocos divididos está em função do nível de escala escolhido, esta técnica é conhecida como *Spatial Pyramid Matching* (SPM). As características são extraídas em cada bloco com um determinado peso para cada nível e o vetores são então concatenados, formando o vetor final.

Uma outra alternativa, para evitar que se tenha um vetor demasiadamente grande, é classificar cada bloco e posteriormente determinar a classificação final da imagem. Gorkani and Picard (GORKANI; PICARD, 1994), por exemplo, classificaram as fotos obtidas em dois tipos de cenas: Cidade ou Subúrbio A abordagem utilizada foi de dividir a imagem em 16 blocos e classificar cada bloco de acordo com orientações de texturas utilizando uma pirâmide multiescalar, ao final venceria a classe com maior número de blocos rotulados.

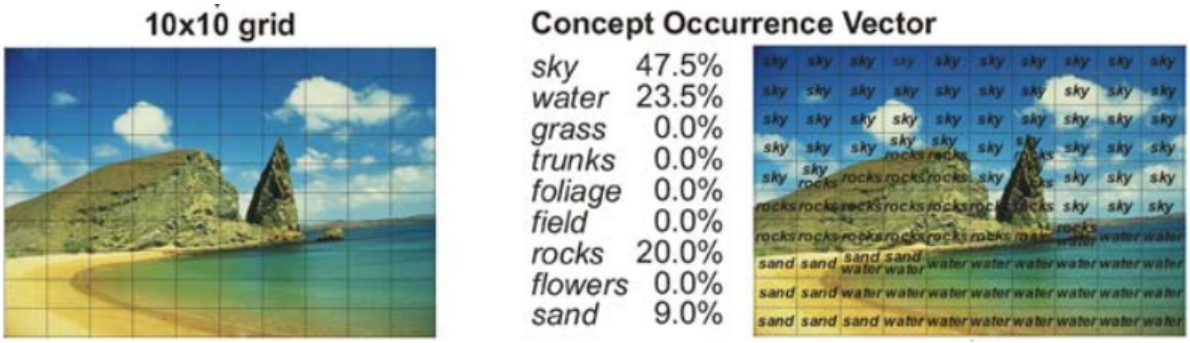
Utilizando informações de textura e cor, Serrano, Savakis e Luo (2002) dividiram a imagem em 16 blocos e aplicaram dois estágios de classificação para rotular a cena em *indoor* ou *outdoor*. As informações de cor são transformadas do modelo RGB (*Red, Green and Blue*) para LST , em que L é a luminância e S e T os espaços de cromaticidade, parecido com o espaço de cor Ohta (OHTA; KANADE; SAKAI, 1980), com exceção do fator de escala. Para cada canal é utilizado um histograma de 16 valores totalizando um vetor final de 48 posições. As características de textura da imagem são extraídas através da transformada Wavelet em dois estágios, totalizando um vetor de 7 posições. A primeira etapa de classificação consistiu em utilizar essas informações de cor e textura para classificar cada um dos 16 blocos em uma das duas classes. A resposta desse estágio

se torna a entrada para o classificador final da imagem.

Buscando simular mais o entendimento humano sobre uma imagem, na literatura se tem empregado em alguns casos o conceito semântico de uma determinada região como atributo para a descrição de uma cena. A ideia principal é modelar uma imagem em regiões locais conforme suas características como água, areia, céu, entre outros conceitos semânticos. Dependendo da qualidade da imagem, do método de classificação e da utilização apropriada desses conceitos, bons resultados podem ser obtidos (RAJA; ROOMI; KALAIYARASI, 2012).

Vogel e Schiele (2007) modelaram uma imagem em função da distribuição dos seus conceitos semânticos, estipulados pelo próprio autor. Inicialmente a imagem é dividida em 100 *grids* e cada *grid* classificado como um entre nove conceitos semânticos, conforme características de cor, textura e energia presentes. O vetor gerado pela frequência da ocorrência das classes semânticas na imagem representa o processo de descrição da imagem, conforme mostra a Figura 7.

Figura 7 – Representação descritiva de uma cena por conceitos semânticos.



Fonte: Vogel e Schiele (2007)

Embora uma partição rígida da imagem em tamanhos de blocos fixos adiciona informação espacial ao processo descritivo, uma região pode eventualmente ser partida em vários blocos ou ainda um único bloco conter várias regiões diferentes, prejudicando o reconhecimento. Por conseguinte, a abordagem local que a principio seria para extrair características locais da imagem pode ter baixa utilidade (CHEN; LI; WANG, 2006). Uma alternativa para solucionar essa deficiência é segmentar as regiões que possuem algumas características em comum.

Em (RAJA; ROOMI; DHARMALAKSHMI, 2013) é realizado um algoritmo de recuperação de imagens utilizando características de baixo nível como cor e textura, além dos mesmos conceitos semânticos utilizados em (VOGEL; SCHIELE, 2007). O modelo de cor empregado foi o HSV (*Hue, Saturation e Value*), através das componentes H e S. Para textura empregou-se o algoritmo *Gray Level Co-occurrence Matrix* (GLCM) (HARALICK; SHANMUGAM et al., 1973) e *Local Binary Pattern* (LBP) (OJALA; PIETIKÄINEN; HARWOOD, 1996). A segmentação é feita baseada em um algoritmo denominado *Norma-*

lized Cut (Ncut), em que o número de regiões que devem ser segmentadas é estipulado de acordo com o numero de picos do histograma de cor. Após este processo, através das características extraídas, uma classe semântica é atribuída para cada região segmentada. A imagem final contém um vetor de classes semânticas, denominado COV (*Concept Co-occurrence Vector*).

Além do Ncut, há na literatura outros algoritmos para segmentação de imagens, dentre eles é possível citar o JSEG *J Segmentation* (DENG; MANJUNATH, 2001), um método não supervisionado que utiliza informações de cor e textura. Neste algoritmo, a segmentação consiste em três parâmetros. O primeiro parâmetro é um limiar para quantização de cor, o segundo é um fator de escala desejado e o terceiro parâmetro um limiar para informação espacial. A Figura 8 mostra um exemplo da segmentação pelo JSEG, note que as regiões estão segmentadas de acordo com sua categoria semântica como madeira, céu e grama.

Figura 8 – Algoritmo de segmentação Jseg.

Fonte: Deng e Manjunath (2001)

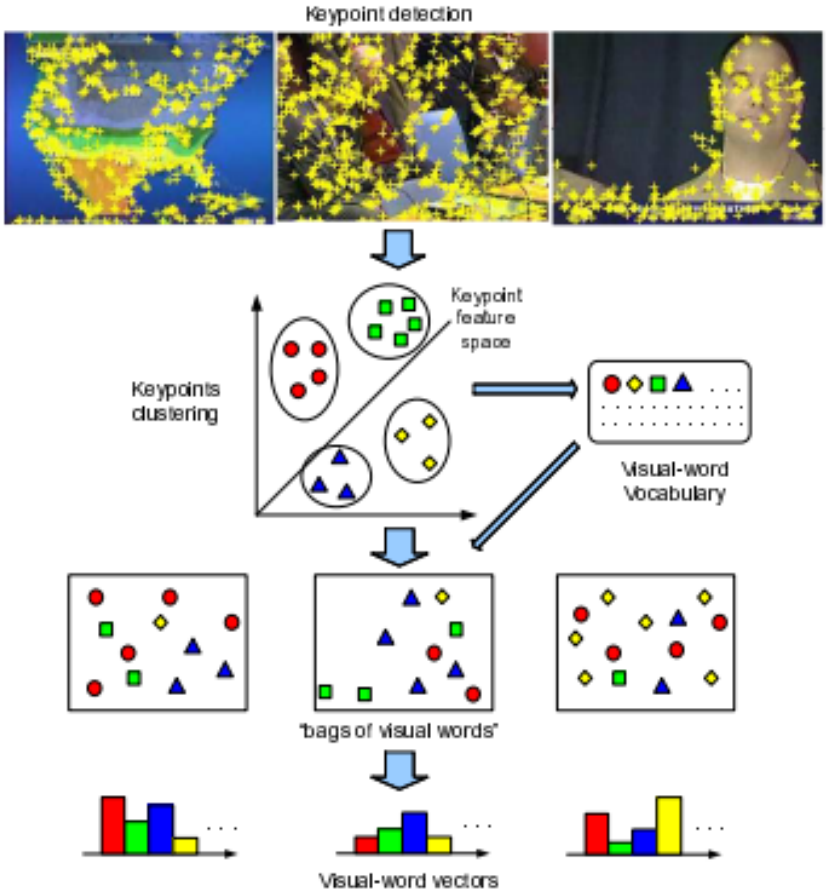
Embora os conceitos semânticos sejam bastante usados na literatura, a grande variabilidade intraclasse que as cenas possuem devido à variações de escala, iluminação e oclusão de objetos, fazem com que estes conceitos fiquem vagos e interpretativos dependendo da aplicação e do algoritmo, dificultando a implementação deste. Uma boa alternativa é a técnica *bag-of-words* (BoW).

O principal objetivo do *bag-of-words* é rotular um conjunto de descritores locais denominados de "palavras visuais", as quais descrevem toda a imagem. A representação final da imagem seria a distribuição da frequência dessas palavras sobre a cena. Ao contrário da noção padrão de linguística em que as palavras já possuem um certo significado que pode ser encontrado em um dicionário por exemplo, as "palavras visuais"não possuem um significado pré-definido, dependendo portanto do banco de dados de imagens a ser treinado (LAZEBNIK; SCHMID; PONCE, 2006).

Antes de descrever a cena de acordo com a distribuição dessas palavras detectadas,

é preciso saber agrupá-las no espaço de dimensão de características. Esse agrupamento é feito na etapa de treinamento e frequentemente usa-se o algoritmo de clusterização denominado *kmeans*, em que cada centro de *cluster* corresponde à uma palavra visual (YANG et al., 2007). Assim, para um ponto de interesse identificado na imagem, este ponto pertencerá ao *cluster* mais próximo à ele no espaço de dimensões. O conjunto de palavras visuais que uma imagem pode possuir é denominado de *codebook*. Este processo está ilustrado na Figura 9.

Figura 9 – Abordagem "Bag-of-Words".



Fonte: Yang et al. (2007)

Fei-Fei e Perona (2005) utilizam o *bag-of-words* e LDA (*Latent Dirichlet Allocation*) para classificar 13 tipos de cenas de um banco de dados construído pelo autor. LDA é um modelo probabilístico generativo não-supervisionado que modela uma coleção de dados discretos utilizando técnicas bayesianas (BLEI; NG; JORDAN, 2003).

Bouachir et al. (BLEI; NG; JORDAN, 2003) propuseram um método de segmentação para cenas externas capturadas por câmeras de segurança. Cada região homogênea é estimada como uma das classes semânticas já pré-definidas anteriormente, baseadas na informação de cor. Ademais, o método proposto emprega outro tipo de extração de características com o *bag-of-words* para os pontos de interesses detectados pelo SIFT.

Uma deficiência dos conceitos semânticos é que a classificação errada de uma determinada região pode prejudicar na classificação da cena. Quelhas et al. (2005) apresentam alguns pontos negativos da utilização desses conceitos como polissemia (uma mesma palavra pode representar diferentes termos) e sinônimos (vários palavras representarem um mesmo conteúdo). Além disso, métodos abordando conceitos intermediários exigem do usuário um conhecimento específico acerca do tema, pois é preciso adicionar parâmetros para serem estimados, proporcionando algoritmos mais complexos.

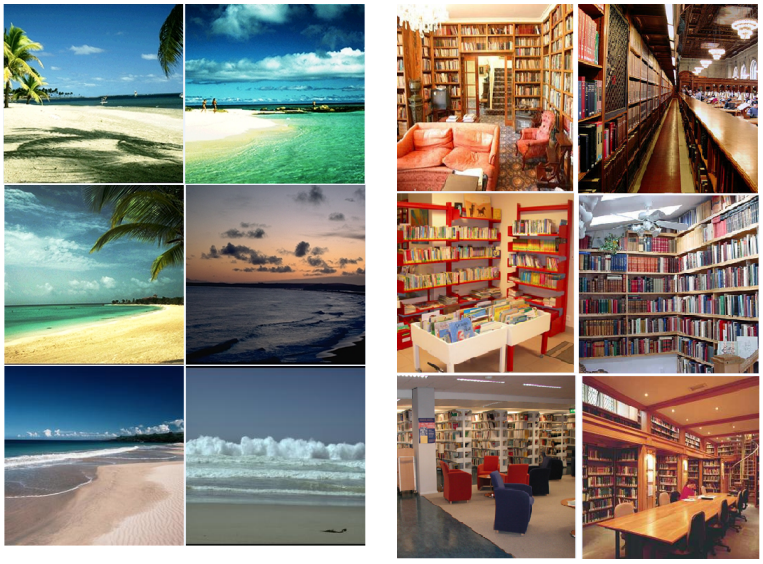
Tanto para abordagens globais quanto locais, existem pontos positivos e negativos que deverão ser analisados dependendo da aplicação que se deseja implementar. Por exemplo, as abordagem holísticas são indicadas para classificar cenas que possuem poucos detalhes, muitas regiões homogêneas e poucas variações de direções de bordas. Pois estas abordagens não exigem muito custo computacional e tratam a cena de forma rápida. Já as abordagens locais são indicadas para cenas com muita riqueza de detalhes, contendo vários objetos em que a ausência ou presença de algum deles não podem influenciar na classificação, estas abordagens possuem um alto custo computacional já que a cena é dividida em blocos e as características locais são tratadas (QUATTONI; TORRALBA, 2009).

Uma boa alternativa para entender melhor estas características é associá-las com o sistema visual humano. As abordagens globais são semelhantes à percepção rápida da cena pelos humanos, em que mesmo em um tempo extremamente curto já é possível classificar a imagem. Já as abordagens locais se compara ao comportamento humano quando analisa-se mais detalhadamente uma cena, identificando os objetos presentes nela. Nestas cenas necessita-se geralmente de um maior tempo de exposição para uma classificação correta . Na Figura 10 mostra-se exemplos de cenas da classe Praia e Livraria. Observe que para a classe Praia, na Figura 10.a, é possível o reconhecimento ocorrer rapidamente, pois estas imagens possuem poucos detalhes e várias regiões de cores uniformes, portanto estas cenas são ideais para a abordagem holística. Já para a classe Livraria, na Figura 10.b, é necessário um tempo maior para identificar corretamente seu conteúdo, analisando os objetos e estruturas presentes na imagem. Este procedimento se assemelha às abordagens locais.

Testes a respeito da importância da informação de contexto dos pontos de interesse na imagem foram realizados ao longo dos anos, tanto para a percepção humana, quanto para a implementação de algoritmos de classificação de cenas. Olivia e Torralba (OLIVA; TORRALBA, 2006) realizaram um estudo para identificar a importância do *layout*, isto é, da organização espacial das estruturas de uma imagem, na classificação da cena. A Figura 11.a ilustra uma cena com borramentos em que pode-se facilmente concluir que se trata de uma cidade, e que na imagem contém céu, carros e prédios, quando na verdade estes são móveis como é ilustrado na Figura 11.b. A conclusão errada de que se trata de

Figura 10 – Cenas de "Praia"e "Livraria".

(a) Praia (b) Livraria



Fonte: Oliva e Torralba (2001) e Quattoni e Torralba (2009)

prédios se justifica pelo fato de que analisa-se o contexto em que esta parte está inserida na imagem. Portanto, segundo Olivia e Torralba, o processamento da estrutura global e relações espaciais entre os componentes é mais importante que as informações locais detalhadas.

Dessa maneira, muitos trabalhos para reconhecimento de objetos utilizam a informação do contexto em que ele está inserido para aprimorar mais os resultados. Outrossim, a informação espacial de contexto sobre os pontos de interesse de uma imagem também vem sendo bastante utilizada para classificação de cenas, tornando-a uma poderosa ferramenta neste campo (FAN et al., 2016).

Figura 11 – Importância do contexto para identificação de cenas e objetos.

(a) Imagem com borramentos. (b) Imagem original.

Fonte: Oliva e Torralba (2006)

Para alcançar maior robustez na classificação de cenas, Chu e Zhao (CHU; ZHAO, 2014) acrescentaram ao *gist* os pontos de interesses detectados pelo SIFT na cena e a

informação contextual desses pontos. Para cada ponto de interesse detectado considera-se também três regiões adjacentes situadas nas direções horizontal, vertical e diagonal. A motivação de usar informação contextual se justifica pelo fato de que uma cena abrange uma vasta área, assim a informação espacial apenas dos pontos de interesse pode não ser suficiente, requisitando informações acerca dele.

Shahriari e Bergevin (2016) investigaram a importância da informação contextual na classificação de cenas. Através dessa informação os pontos de interesse detectados pelo SIFT são separados em regiões homogêneas e não-homogêneas da imagem por intermédio da informação de contraste. A representação final é a junção do vetor descritivo das regiões homogêneas (baixo contraste) com as não-homogêneas (alto contraste).

Em (LIU; XU; FENG, 2011) é proposto o *Region Contextual Visual Words* (RCVW), um método de classificação de cenas integrando informação contextual nos pontos descritos pelo *bag-of-words*. A motivação para essa abordagem foi o fato de que o BoW pode considerar dois *patches* (trechos) iguais, todavia eles pertencem à classes distintas. Com o acréscimo da informação ao redor desses *patches*, o poder discriminativo deles aumenta, como foi observado. Os autores introduziram ainda o *Region-Conditional Random Fields*, um modelo para aprender cada palavra gerada pelo BoW em função das demais palavras que se encontram na mesma região.

Além da criação de algoritmos cada vez mais robustos e eficientes, outro fator que é primordial na classificação de cenas é o banco de dados utilizado (LI; FEI-FEI, 2007). Isto pode ser melhor explicado analisando o comportamento humano. Quando este se depara com uma grande extensão de exemplos de todos os tipos de cenas e objetos, a identificação e reconhecimento ocorre mais rapidamente. Da mesma forma acontece com as máquinas, ou seja, quando possuem um banco de dados mais completo, contendo todo tipo de variação que uma cena pode apresentar, melhores resultados são proporcionados. Torralba, Fergus e Freeman (2008) realizaram um estudo para identificar o quão importante são os bancos de dados para o reconhecimento de algumas cenas e foi constatado que os melhores resultados foram alcançados com os maiores bancos de dados.

Na literatura encontra-se alguns trabalhos responsáveis pela aquisição de banco de dados cada vez mais completos para classificação de cenas e objetos, entre eles pode-se destacar:

- *LabelMe* (TORRALBA; RUSSELL; YUEN, 2010) - uma ferramenta aberta em que o usuário anota manualmente o conteúdo da cena. O usuário pode acessar através do aplicativo para celular "*LabelMe App*"ou através do site http://labelme.csail.mit.edu.
- *Scene Undersanting* (SUN) *Database* (XIAO et al., 2010) - atualmente composto por mais de 130 mil imagens contendo 908 categorias de cenas e 4479 categorias de objetos.

- *TinyImages* (TORRALBA; FERGUS; FREEMAN, 2008) - uma espécie de dicionário visual contendo 53464 palavras distribuídas ao longo de 80 milhões de imagens de baixa resolução. Neste sistema, o usuário pode buscar uma determinada palavra e o computador oferece algumas imagens que ele julga pertencer à essa palavra, o usuário então seleciona as palavras associadas corretamente. O objetivo aqui é ensinar a máquina a ver através de uma aprendizagem supervisionada e assim, treinar um computador para que ele esteja apto a reconhecer qualquer tipo de objeto ou cena (TORRALBA; FERGUS; FREEMAN, 2008).

- *Places2 Database* (ZHOU et al., 2017) - contém 10 milhões de imagens distribuídas em 434 classes, foi desenvolvido com o objetivo de construir um núcleo de conhecimento visual para treinar sistemas artificiais com técnicas de alto nível em aprendizagem de máquina. Este banco de dados permite que a utilização de características e técnicas mais aprofundadas para tarefas de reconhecimento de cenas como contexto, identificação de objetos e previsão de eventos (ZHOU et al., 2017).

Quando utiliza-se abordagens holísticas, com algoritmos de baixo custo computacional, sem acrescentar muita informação, o programa pode ter um vetor representativo da cena com poucas dimensões, porém os resultados podem não ser satisfatórios. Ao passo que, ao adicionar informações como contexto e cores, por exemplo, o vetor de características que representa a cena pode ficar demasiadamente grande. Portanto, é necessário identificar se esse acréscimo irá contribuir para um aumento da acurácia dos experimentos. O mesmo deve ser avaliado quando abordagens locais são utilizadas para extrair características pontuais dividindo a imagem em blocos.

A escolha da metodologia empregada também deve ser estudada dependendo da aplicação. Quando algoritmos de custo computacional baixo são utilizados, os resultados finais podem ser prejudicados. Todavia, o emprego de técnicas com mais parâmetros a serem estimados como o BoW, apesar de proporcionarem melhores resultados, trazem com elas um custo computacional mais elevado. Além disso, o acréscimo de parâmetros e informações específicas relacionadas ao algoritmo limita a utilização do sistema apenas para pessoas com conhecimento específico na área, excluindo assim as pessoas leigas no assunto.

Visando desenvolver um sistema para classificação de cenas aberto para a comunidade em geral, tem-se aqui um desafio: produzir um algoritmo utilizando informações importantes identificadas na literatura como contexto e espaço, sem que isso proporcione um vetor demasiadamente grande para representar a imagem, com operações de baixo custo computacional, poucos parâmetros a serem estimados para que leigos possam usufruir e ainda assim, obter resultados competitivos.

1.4 Visão Geral do Sistema

A Figura 12 ilustra o fluxograma do algoritmo proposto neste trabalho. Utilizando o LBP, um descritor que extrai padrões de textura, uma nova imagem é gerada. Os pixeis nesta imagem são então classificados entre coerentes e incoerentes de acordo com a presença de pixeis de mesma intensidade na sua vizinhança [5]. Esta operação é inspirada na técnica *Color Coherence Vector* (CCV) (PASS; ZABIH; MILLER, 1997). Por fim, é gerado um histograma apenas com a informação dos pixeis denominados incoerentes.

Esta é a primeira metade do trabalho, a segunda parte se refere ao emprego da informação de contexto, em que aplica-se o algoritmo MCT8 (GAZOLLI; SALLES, 2014) por duas vezes seguidas. Posteriormente, é realizada a mesma separação entre pixeis coerentes e incoerentes na nova imagem gerada, calculando o histograma dos pixeis incoerentes.

O vetor descritivo final é a concatenação entre esses dois histogramas gerados pelos pixeis incoerentes. Por fim, o vetor passa por um classificador treinado por uma certa base de dados.

1.5 Organização do Trabalho

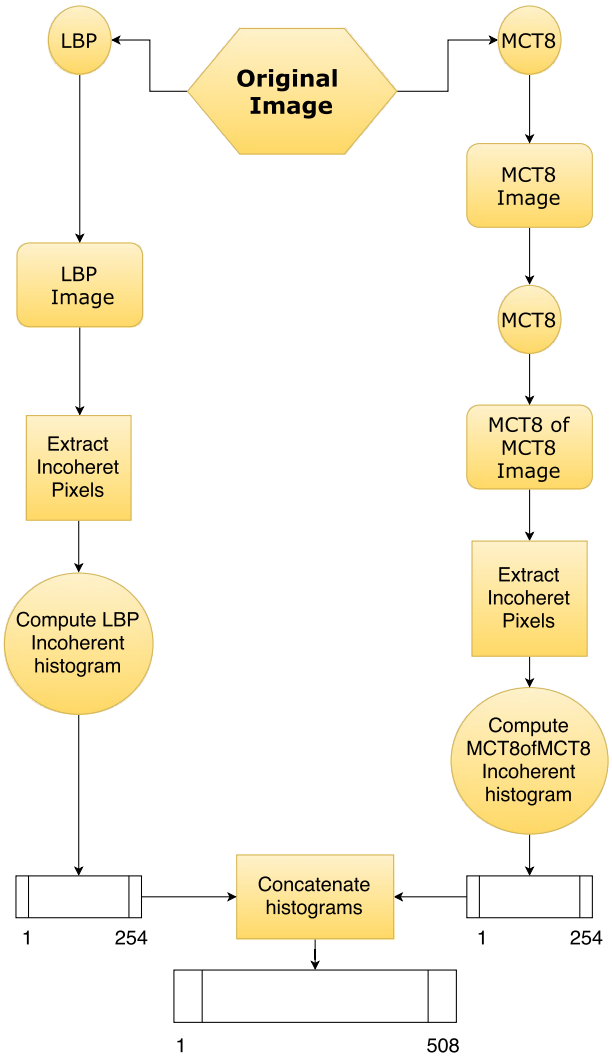
Este trabalho é organizado da seguinte maneira: no Capitulo 1 foram apresentados alguns problemas referentes à classificação de cenas e técnicas já implementadas, no Capitulo 2 é descrita a metodologia proposta para o LBP Incoerente, especificando os algoritmos já existentes na literatura que ajudaram a inspirar a técnica proposta, no Capitulo 3 acrescenta-se a informação de contexto sobre o algoritmo proposto e os resultados adquiridos, além da técnica de classificação e validação usada, e finalmente o Capítulo 4 encerra este trabalho com as principais conclusões e caminhos futuros de investigação.

1.6 Contribuições

- Proposta de um descritor de imagens que utiliza operações de baixa complexidade quando comparado a outros descritores e, ainda assim, obter resultados competitivos.
- Utilização da informação contextual como um método importante para a classificação de cenas.
- Exploração da informação espacial pelo algoritmo CCV integrado ao LBP de modo que obteve-se uma melhor representação do vetor descritivo da imagem.

[5] Pixeis coerentes fazem parte de regiões com um certo número de pixeis de mesma intensidade conectados entre si, ao contrário dos pixeis incoerentes

Figura 12 – Fluxograma do sistema proposto.



Fonte: Próprio autor

1.7 Trabalho Publicado

RIBEIRO, M. V. L. e SALLES, E. O. LBP Incoerente como uma nova proposta para descrição de cenas. In: *Simpósio Brasileiro de Automação Inteligente*, Porto Alegre, Brasil, p. 477-482, 2017.

2 LBP Incoerente

Neste capítulo é proposto o descritor LBP Incoerente, descrevendo as duas técnicas utilizadas para a construção do vetor descritivo final. Por fim, são apresentados os resultados adquiridos com as bases de dados e classificador utilizados.

O LBP Incoerente visa descrever a imagem como um único objeto, se trata portanto de uma abordagem holística. Este algoritmo captura os pixeis gerados pelo LBP com poucos valores iguais conectados entre si e calcula o seu histograma. LBP Incoerente é uma implementação com poucas operações, retirando informações de regiões homogêneas da imagem, sem que o custo computacional fique elevado em relação ao LBP original e, ainda assim, conseguir resultados competitivos.

2.1 Local Binary Pattern

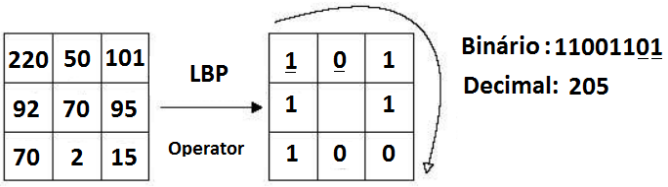
Local Binary Pattern é uma técnica desenvolvida por Ojala (OJALA; PIETIKÄINEN; HARWOOD, 1996) cujo principal objetivo é modelar as estruturas locais da cena. Trata-se de uma poderosa ferramenta para classificação de texturas em virtude de possuir poucas operações e, ainda assim, possibilitar um bom poder discriminativo (DAS; JENA, 2016). Estes fatores proporcionam ao LBP uma vasta faixa de aplicação nas mais diversas áreas, como recuperação de imagens (SINGH; AGRAWAL, 2016), reconhecimento de expressão facial (ZHANG; LIU; XIE, 2016) e sensoriamento remoto (DAN et al., 2014), em que a informação de textura é importante na representação da imagem.

O LBP se caracteriza por ser um operador não-paramétrico, ou seja, os pixeis não são avaliados de acordo com o seu valor absoluto. Esta informação é importante visto que certas variações na intensidade dos pixeis podem não alterar o seu valor final produzido pelo LBP, tornando esta ferramenta invariante à iluminação (OJALA; PIETIKAINEN; MAENPAA, 2002). Além disso, o fato deste algoritmo não acrescentar informações de contraste aumenta seu poder discriminativo na identificação de texturas.

Supondo uma janela de 3 x 3 pixeis, o LBP consiste em rotular o valor do pixel central fazendo comparações com cada um dos seus 8 pixeis vizinhos. Cada comparação produz um número binário em que o valor 0 será atribuído quando o pixel vizinho for menor que o pixel central e 1 caso contrário. Ao final de todas as comparações, os 8 valores são concatenados e um número binário de 8 dígitos é formado, a concatenação de cada bit é feita à esquerda dos bits já atribuídos. Este número passado para a forma decimal gera um valor de 0 a 255 que será atribuído ao pixel central. A Figura 13 ilustra esse processo, perceba que a ordem de comparação com os pixeis vizinhos não é feita de forma aleatória

sendo, neste caso, em sentido horário. Os dois primeiros números binários comparados estão sublinhados para melhor percepção do processo de concatenação e formação do número binário de 8 dígitos.

Figura 13 – Local Binary Pattern.



Fonte: Próprio autor

Há na literatura algumas técnicas parecidas utilizando transformada não-paramétrica, dentre elas pode-se citar o *Texture Spectrum Operator* (HE; WANG, 1990) que captura três níveis de valores diferentes para cada comparação gerando 3^8 valores possíveis. Outra técnica é a transformada *Census*, cuja principal diferença está na ordem de comparação bit a bit do pixel central com seus vizinhos (ZABIH; WOODFILL, 1994).

O valor decimal resultante das operações do LBP descritas anteriormente, pode ser representado algebricamente através da seguinte equação:

$$LBP(x,y) = \sum_{n=0}^{7} s(i_n - i_c)2^n, \tag{2.1}$$

em que x e y representam as coordenadas do pixel central na imagem original, i_c representa o valor em nível de cinza do pixel central e i_n representa um de seus 8 pixeis vizinhos. A função $s(z)$ é definida como:

$$s(z) = \begin{cases} 0, z < 0 \\ 1, z \geq 0. \end{cases} \tag{2.2}$$

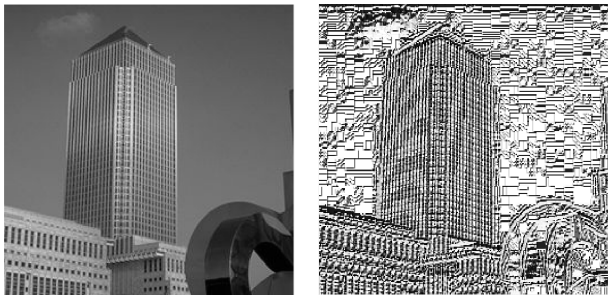
A Figura 14 exemplifica uma imagem gerada quando aplica-se o operador LBP em todos os pixeis (com exceção dos pixeis na borda da imagem). Observe que as bordas das estruturas praticamente não são alteradas. A permanência das bordas é um fator importante, visto que essa informação pode ajudar na classificação das cenas.

Após a aplicar o LBP para todos os pixeis da imagem original, calcula-se o histograma H_{lbp} deste valores gerados, isto é, a distribuição da frequência destes valores, representado algebricamente por:

$$H_{lbp}(n) = \sum_{i=1}^{a}\sum_{j=1}^{b} \delta(n = LBPimage(i,j)), \tag{2.3}$$

em que n é o n-ésimo valor de 0 a 255 e a e b os números de linhas e colunas respectivamente de $LBPimage$[1].

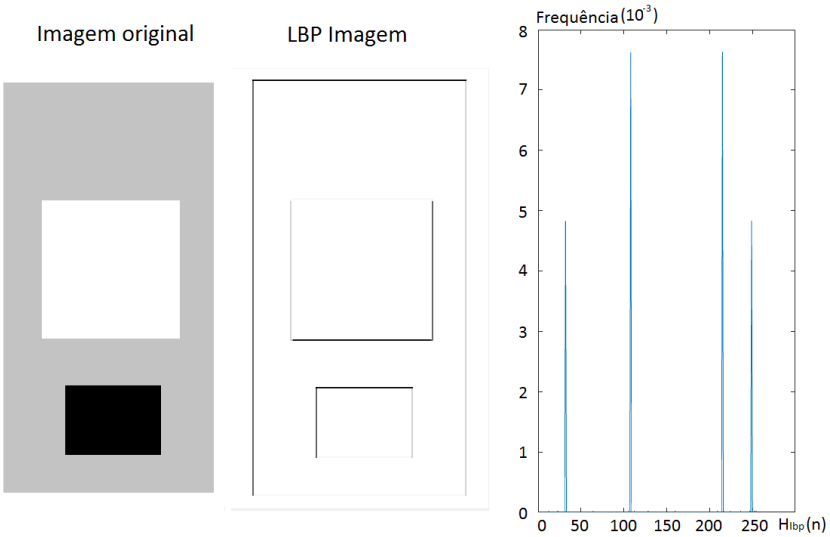
Figura 14 – Imagem formada pelos valores obtidos com a operação do LBP em todos os pixeis.



Fonte: Próprio autor

Uma vez que o LBP captura as propriedades estruturais locais da imagem, como direções de bordas e superfícies planas, seu histograma pode oferecer informações importantes já que modela a distribuição dessas estruturas na cena. A Figura 15 exemplifica como o histograma pode ajudar na descrição de uma imagem. Observe que a região uniforme preta, por ser uma região homogênea, é suprimida pelo LBP. Isto acontece porque independente do nível da intensidade de um pixel, se este for igual a todos os seus vizinhos, a ele será atribuído o valor de 255, de acordo com a equação 2.1 (lembre-se que o LBP é um operador não-paramétrico). Repare que na imagem gerada pelo LBP que as regiões homogêneas de cores brancas, pretas e cinzas irão possuir este mesmo valor após a operação.

Figura 15 – Histograma de uma imagem formada pela aplicação do operador LBP em todos os pixeis da imagem original.



Fonte: Próprio autor

[1] É importante ressaltar que o algoritmo LBP não gera uma nova imagem, porém como a sua aplicação sobre os pixeis de uma imagem original resulta em um algoritmo com valores de 0 a 255, pode-se transformar estes valores em uma nova imagem desde que se preserve a posição dos pixeis

Outro aspecto importante a ser observado é a intensidade das bordas na imagem gerada pelo LBP. Observe que, tanto no quadrilátero branco quanto no quadrilátero preto, cada lado possui um nível de intensidade diferente gerado pelo LBP. Isto acontece porque nessas bordas, a direção entre um nível de intensidade para outro interfere no valor atribuído final, já que a posição do bit comparado em relação ao pixel central influencia no valor final resultante da operação do LBP. Assim, uma borda que possui um nível de cinza maior no seu lado direito que no lado esquerdo possuirá um valor diferente de uma borda com seus lados nos valores inversos. Observe que as bordas do lado esquerdo e superior do quadrado preto tem valores parecido das bordas do lado direito e inferior do quadrado branco.

Se houvesse um tom de cinza mais escuro na região cinza, a imagem gerada pelo LBP e o seu histograma seriam os mesmos, pois para o LBP não importa o valor absoluto do pixel. O que é levado em consideração é se ele é maior, igual ou menor em relação ao pixel vizinho.

Como a imagem possui 4 níveis de cinza diferentes, além do nível de cinza 255 gerado pelos pixeis nas regiões homogêneas, obtêm-se um histograma com 5 valores diferentes de zero. No histograma da Figura 15 descarta-se o nível 255 pois representa um valor muito alto comparado aos demais, prejudicando sua análise.

Note que esta é uma característica peculiar para formas retangulares de intensidade uniforme , ou seja, a presença deste polinômio irá originar 4 níveis de cinza no histograma (além do nível de cinza 255) e os valores de frequência estarão em função do tamanho do quadrilátero. Se o quadrilátero estivesse inclinado na imagem, os números binários formados pelo LBP seriam diferentes e consequentemente o histograma assumiria outro comportamento. Entretanto, devido ao efeito de serrilhamento, o histograma não contará com apenas 4 níveis distintos como aconteceu na Figura 15

O histograma gerado pelo LBP herda deste algoritmo algumas características importantes como invariância à iluminação, custo computacional baixo e nenhum parâmetro a ser estimado pelo usuário. Além disso, caso haja pequenos ruídos na imagem, o histograma não sofrerá grandes variações, conferindo a ele maior robustez (WU; REHG, 2011). Estas características desejadas fazem com que o histograma do LBP seja utilizado como vetor descritivo em vários trabalhos nas mais diversas áreas.

Visando um sistema para detecção humana na imagem Wang et al. (2015) construíram um descritor de textura aliando o histograma do LBP com a transformada de Fourier ,formando o LBPHF(*Local Binary Pattern Histogram Fourier*), e com o Histograma de Orientação ao Gradiente (HOG). Os autores ressaltam que o algoritmo proposto produz um custo computacional menor que outras técnicas na literatura e os resultados são satisfatórios em relação à elas. Também buscando adicionar a transformada de Fourier junto ao LBP para extração de características, Bharathi, Reddy e Srilakshmi (2014) implementaram

um sistema de recuperação de imagens médicas.

O LBP é frequentemente empregado no reconhecimento de faces para a descrição e representação de imagens, pois este algoritmo serve como um bom descritor de textura, além de ser apropriado para sistemas em tempo real em virtude do seu baixo custo computacional.

Ahonen, Hadid e Pietikäinen (2004) implantaram um sistema de segurança para reconhecimento e identificação de faces utilizando o LBP e seu histograma. Os autores ressaltam o fato de que, em alguns casos, ladrões podem tentar burlar o sistema visual utilizando máscaras de alguma pessoa específica que possui permissão para entrar. Como o LBP é um bom descritor de textura ele não reconhece a pele humana nesses casos, impedindo a passagem do infrator.

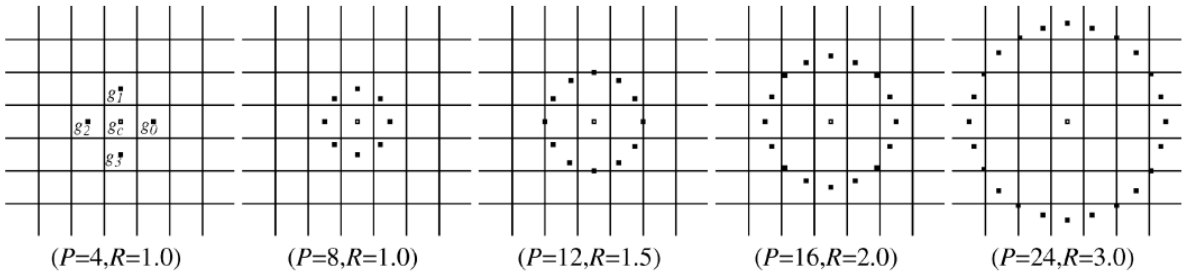
Em (TANG et al., 2010) é proposto uma nova abordagem para representar a face humana. Antes de usar o histograma do LBP para extrair as características da imagem, esta é dividida em 4 blocos através da transformada Wavelet Haar. Em seguida, o LBP é aplicado em cada um destes blocos, formando o HLBPH *Haar Local Binary Pattern Histogram*. Os autores destacam nos resultados a vantagem do sistema proposto ser invariante à iluminação.

Desta forma, o LBP em virtude do seu baixo custo computacional, permite que outras técnicas possam ser incorporadas a ele sem que gere um vetor descritivo com dimensões elevadas, proporcionando um novo tipo de descritor mais completo. Outrossim, encontra-se na literatura algumas extensões do LBP visando uma maior generalidade e robustez para este algoritmo.

Buscando adicionar invariância à rotação e escala, Ojala, Pietikainen e Maenpaa (2002) criaram o $LBP^{riu2}_{(P,R)}$ *Local Binary Pattern Rotation Invariant Uniform*, em que P é a quantidade de bits vizinhos que o pixel central deve comparar e R a distância dessa vizinhança. Tem-se portanto, dois parâmetros que permitem extrair o LBP em várias escalas diferentes como mostra a Figura 16. Os autores identificaram que em regiões uniformes, o número de variações entre bits consecutivos no valor binário gerado pelo LBP é pequeno, assim valores binários com até duas variações de um bit para o outro seriam rotulados em um mesmo grupo no histograma. Esta detecção de pixeis em regiões uniformes proporcionou uma desejada invariância à rotação ao vetor descritivo final além de uma redução do número de valores possíveis gerados pelo LBP.

Por ser um operador não-paramétrico, o LBP tem como uma de suas principais deficiências descartar as informações a respeito dos valores absolutos dos pixeis. Visando adicionar essas informações ao valor gerado pelo LBP, Li, Li e Kurita (2015) construíram o SOFT LBP, em que a função de limiarização, que no LBP original possui o valor do pixel central, é uma função logística. Assim, o limiar de comparação pode variar de acordo

Figura 16 – Diferentes valores de P e R para o $LBP^{riu2}_{(P,R)}$.



Fonte: Ojala, Pietikainen e Maenpaa (2002)

com a diferença de magnitude entre o pixel vizinho e o pixel central.

Outro ponto negativo é a falta de informação espacial que um histograma traz, pois se analisa apenas a frequência com que os níveis de cinza ocupam na imagem, independente de sua distribuição espacial. Para extrair estas informações espaciais Nosaka, Ohkawa e Fukui (2012) introduziram o conceito de *Co-ocurrence* ao LBP. *Co-ocurrence* é frequentemente utilizado para extrair informações globais em várias regiões locais. Os autores descrevem um método de obter a ocorrência de todas as combinações possíveis do LBP usando matriz de auto-correlação entre dois valores específicos.

Um método parecido com o histograma do LBP e inspirado na Transformada *Census* é o CENTRIST (*Census Transform Histogram*), criado em (WU; REHG, 2011). O CENTRIST é um descritor visual de cenas obtido através do histograma da imagem gerada pela Transformada *Census*. Assim como o histograma do LBP, este descritor captura as propriedades estruturais gerais da imagem suprimindo detalhes de textura em regiões homogêneas.

2.2 Color Coherence Vector

Baseado nas vantagens e eficiência que o histograma de cor pode oferecer para representar uma imagem em diversas aplicações e visando aperfeiçoar esta técnica, (PASS; ZABIH; MILLER, 1997) criaram o *Color Coherent Vector* (CCV). Este algoritmo foi inspirado no fato de que, como mencionado na seção anterior, os histogramas não acrescentam informação espacial ao vetor descritivo. Portanto, duas imagens completamente distintas podem possuir o mesmo número de pixeis de uma determinada cor.

A Figura 17 mostra um exemplo desta situação em que há o mesmo número de pixeis vermelhos nas duas imagens. Observe que a distribuição espacial desses pixeis para cada imagem é diferente. Enquanto que na imagem da esquerda os pixeis vermelhos se encontram em diversas regiões, representadas pelas flores, na imagem da direita há apenas uma única região uniforme em que eles se concentram, representada pela camisa do jogador (PASS; ZABIH; MILLER, 1997).

Figura 17 – Duas imagens com números de pixeis de cor vermelha iguais.

Fonte: Pass, Zabih e Miller (1997)

Assim, o CCV tem como objetivo separar um pixel em duas classes: coerente ou incoerente. Esta classificação será de acordo com a região em que ele estiver situado. Após essa separação calcula-se o histograma para cada uma das classes e constrói-se o vetor descritivo final. Estas operações acrescentam informação espacial ao histograma de cor, conferindo melhores resultados em relação ao histograma final de acordo com os testes feitos em (PASS; ZABIH; MILLER, 1997).

O processo de construção dos histogramas de pixeis coerentes e incoerentes podem ser divididos em 3 etapas: quantificação, construção do histograma de regiões e separação das classes.

2.2.1 Discretização

A etapa de discretização consiste em quantizar os pixeis em determinados intervalos, denominados *buckets*, igualmente espaçados de acordo com a faixa de valores que um pixel pode possuir. Assim, para o número de *buckets* igual a 16 por exemplo, serão proporcionados 16 intervalos igualmente divididos em uma faixa de 256 valores possíveis. Deste modo, bits de 0 a 15 pertencerão ao *bucket* 0, os bits de 16 à 31 pertencerão ao *bucket* 1 e assim sucessivamente até a formação do *bucket* 15 composto pelos bits com valores de 240 a 255. A figura 18 representa este processo para 16 *buckets*. A informação da quantidade de *buckets* é importante pois o histograma final terá como tamanho final o número de *buckets* escolhido, pois cada *bucket* representa um nível de cinza possível na imagem.

2.2.2 Construção do Histograma de Regiões

O próximo passo é rotular uma região para cada pixel, sendo que os pixeis vizinhos com o mesmo valor de intensidade pertencerão à mesma região. A vizinhança aqui definida é a de 8 conectada. Após atribuir a região correspondente para cada pixel, é feito um

Figura 18 – Etapa de discretização.

(a) Janela original

6	10	9	33	42	50
7	25	20	35	40	45
10	1	13	13	60	56
29	36	35	20	61	0
31	30	20	23	5	1

(b) Janela discretizada

0	0	0	2	2	3
0	1	1	2	2	2
0	0	0	0	3	3
1	2	2	1	3	0
1	1	1	1	0	0

Fonte: Próprio autor

histograma dessas regiões obtidas, ou seja, a frequência com que cada região rotulada aparece na imagem. Esse histograma é importante pois, a partir dele, será definido se todos os pixeis de determinada região serão classificados como coerentes ou incoerentes.

A Figura 19 demonstra este processo de rótulo de regiões e o seu respectivo histograma. Observe que pixeis com mesmo valor de intensidade não necessariamente pertencerão à mesma região, visto que a informação espacial também é considerada, isto ocorre entre as regiões G e A por exemplo.

Figura 19 – Processo de atribuição das regiões.

(a) Janela discretizada.

0	0	0	2	2	3
0	1	1	2	2	2
0	0	0	0	3	3
1	2	2	1	3	0
1	1	1	1	0	0

(b) Regiões rotuladas.

A	A	A	C	C	D
A	B	B	C	C	C
A	A	A	A	E	E
F	H	H	F	E	G
F	F	F	F	G	G

Fonte: Próprio autor

2.2.3 Separação de classes

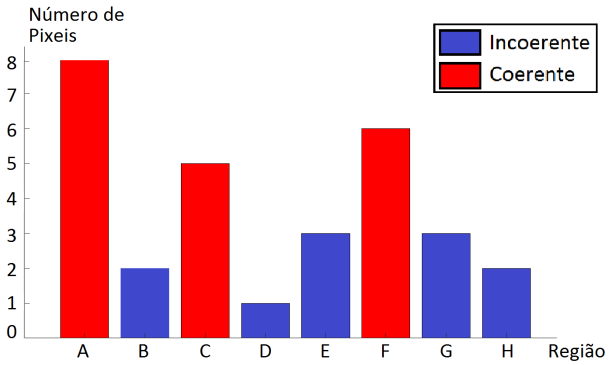
Nesta etapa os pixeis são classificados como coerentes ou incoerentes de acordo com a região na qual ele foi atribuído. Um pixel coerente faz parte de uma região em que, após a etapa de quantização, existe um grande número de pixeis com a mesma intensidade que ele conectados entre si. Como pertencem à mesma região, serão todos classificados como coerentes. Já um pixel incoerente é um pixel situado em uma região com pouca quantidade de pixeis de igual valor conectados entre si, após a etapa de quantização. Assim, quando o histograma dessas regiões é construído, os pixeis pertencentes às regiões com maiores frequências pertencerão à classe coerente, pois as regiões coerentes possuem grande quantidade de pixeis. Caso contrário pertencerão à classe incoerente.

O valor mínimo de pixeis para que uma região seja classificada como coerente é estimado através de um limiar. Deste modo, uma região com frequência maior ou igual a

esse limiar terão todos os seus pixeis classificados como coerentes e abaixo deste limiar classificados como incoerentes.

O valor deste limiar varia, como será explicado mais adiante. No trabalho de (PASS; ZABIH; MILLER, 1997) por exemplo, este valor foi de 1% da imagem. No histograma de regiões formadas na Figura 19, adotou-se um limiar de 4 pixeis. Por conseguinte, regiões com frequência abaixo deste valor no histograma terão todos os seus pixeis pertencentes à classe incoerente e acima ou igual à este valor, classificados como coerentes, como se observa na Figura 20.

Figura 20 – Separação das regiões em coerentes e incoerentes.



Fonte: Próprio autor

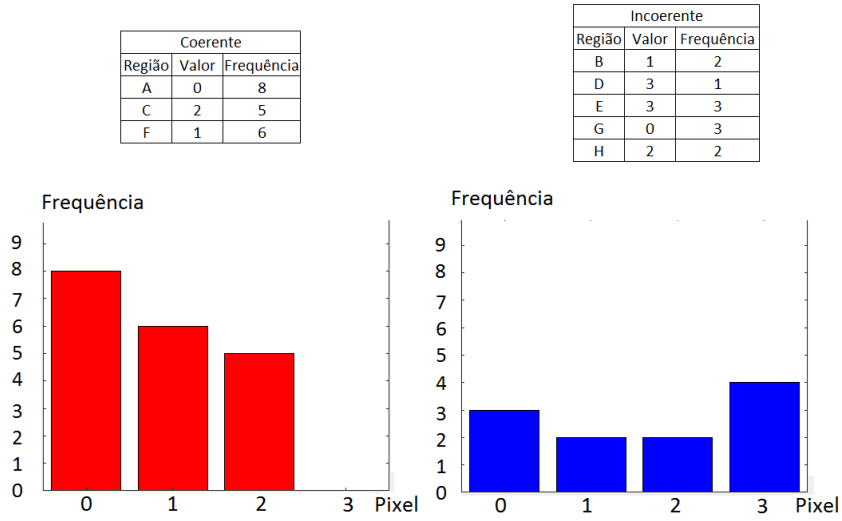
Por fim, constrói-se os dois histogramas utilizando a informação de cada região para a obtenção do vetor descritivo final. A Figura 21 ilustra esse processo. Inicialmente captura-se o valor do pixel na qual aquela região pertence, em seguida a quantidade de pixeis daquela região é inserida no histograma coerente ou incoerente, de acordo com a classificação da respectiva região.

Com os experimentos realizados em (PASS; ZABIH; MILLER, 1997) foi possível concluir que o CCV alcançou melhores resultados que o histograma de cor para a classificação de cenas . Este trabalho motivou outras pesquisas com o mesmo objetivo de adicionar informação espacial junto aos histogramas adquiridos. Além da informação de cor, podemos ver esta técnica incorporada a outras diferentes informações.

Vailaya, Jain e Zhang (1998) utilizaram esta técnica não só para cores como também para as bordas, criando o "*Edge Direction Coherence Vector*" (EDCV). O objetivo era separar as cenas em classes como Cidade, Montanha, Floresta e Pôr-do-Sol. Tanto o EDCV como o CCV tiveram desempenho melhor que as técnicas originais de histogramas de borda e de cor.

Em 2014, Salmi e Boucheham (2014) propuseram o Cell-CCV, uma nova abordagem utilizando informação de cor para recuperação de imagens. Cell-CCV é um fusão de duas técnicas que empregam conteúdo de cor: *Cell Color Histogram* (CCH) e CCV. Este novo vetor descritivo supera os resultados adquiridos pelos histogramas de cor.

Figura 21 – Construção dos histogramas coerentes e incoerentes.



Fonte: Próprio autor

2.3 LBP Incoerente

Nesta seção é descrito o algoritmo *Local Binary Pattern* Incoerente (LBP Incoerente), um vetor descritivo de imagens no qual utiliza informações a respeito das estruturas locais da cena, descartando as regiões com muitos pixeis de mesmo valor conectados entre si. Este vetor foi inspirado nas duas técnicas descritas na seções anteriores: LBP e CCV. A construção do LBP Incoerente não emprega técnicas que possibilitam um alto custo computacional ao sistema e adiciona apenas alguns conceitos em relação ao LBP como o tamanho do limiar de região para classificar as regiões em coerentes e incoerentes, como será visto neste capítulo.

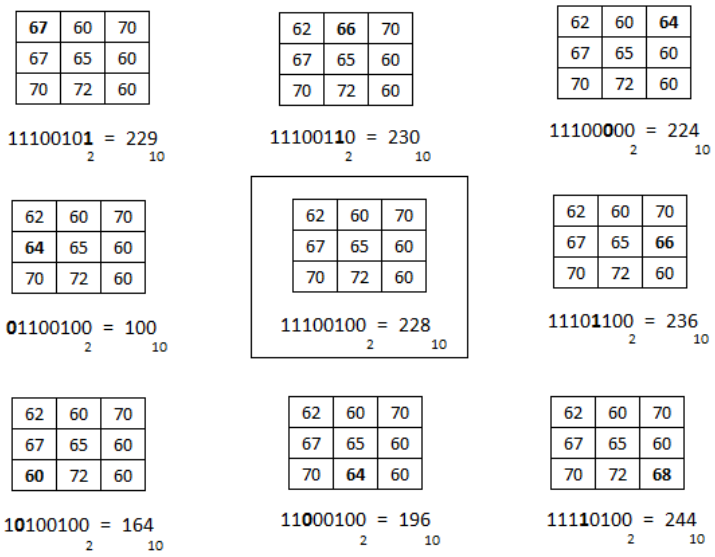
Apesar de utilizar o LBP, a informação a respeito do seu histograma não será necessária, visto que serão construídos outros dois histogramas distintos para a descrição da imagem. O algoritmo é descrito da seguinte forma: a partir de uma cena qualquer em nível de cinza, uma imagem é gerada pelo LBP, posteriormente separa-se os pixeis desta nova imagem em coerentes ou incoerentes de acordo com a relação deles com seus vizinhos e após essa separação, dois histogramas são construídos, o LBP Coerente e LBP Incoerente.

Inicialmente os valores da imagem gerada pelo LBP são discretizados em intervalos de 1, 2, 4 e 8 níveis, formando imagens com 256, 128, 64 e 32 níveis de cinza respectivamente, os *buckets* já mencionados anteriormente. Esta informação é importante pois o tamanho do intervalo determinará o tamanho do vetor descritivo final. Quanto maior o intervalo, menor o número de valores possíveis e menor o tamanho do histograma.

É importante ressaltar que o LBP, é um operador não linear, ou seja, pequenas variações no valor do pixel podem acarretar grandes variações no valor final gerado pelo

LBP dependendo da posição do pixel em relação ao pixel central de comparação, como ilustra a Figura 22. Observe que isto não acontece com as cores no CCV visto que a técnica não aplica um operador não linear como o LBP. Dessa forma, a quantização realizada através dos *buckets* aqui realizada é apenas um método para diminuir o tamanho final do histograma da imagem.

Figura 22 – Pequenas variações na intensidade do pixel podem provocar diferentes variações no valor final gerado pelo LBP.



Fonte: Próprio autor

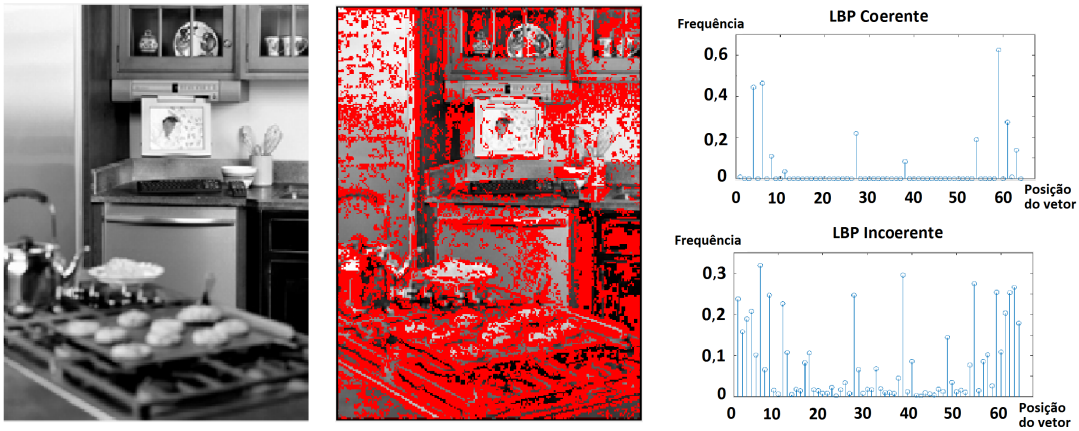
Posteriormente, constrói-se o histograma de regiões da mesma forma que descrito no CCV. Com essa informação os pixeis são separados em coerentes e incoerentes. Por fim, calcula-se os histogramas de cada um e o vetor descritivo final é formado.

A Figura 23 mostra um exemplo de uma cena da classe Cozinha e os histogramas coerente e incoerente produzidos para essa imagem. Note que é empregado um intervalo de 4 níveis na quantização, gerando histogramas de 64 posições. Na imagem, usa-se pixeis salientados em vermelho como uma forma ilustrativa para destacar as posições dos LBP coerentes. Já os pixeis restantes, que estão no mesmo nível de cinza que a imagem original, pertencem ao grupo dos LBP incoerentes.

Perceba que os pixeis coerentes se situam nas regiões mais homogêneas da imagem, como as paredes por exemplo, porém não preenchem toda a região. Isto acontece porque o LBP é um operador não paramétrico e não-linear. Assim, uma pequena diferença de valor de um pixel vizinho, pode influenciar consideravelmente no valor atribuído pelo LBP, como explicado anteriormente.

Outro fato importante de se analisar é que a parte inferior da imagem está com um borramento, o que influenciou no aumento da quantidade de pixeis coerentes na imagem. Observe que há mais pixeis vermelhos na parte inferior, mesmo com poucas

Figura 23 – Imagem com os LBP coerentes (pixels vermelhos) e incoerentes e seus histogramas.



Fonte: Próprio autor

regiões homogêneas, do que na parte superior da imagem.

Os pixeis coerentes também se situam em regiões com padrões de textura repetitivos, pois terão o mesmo valor atribuído pelo LBP, já que este algoritmo funciona como um bom identificador de textura (OJALA; PIETIKÄINEN; HARWOOD, 1996). Isto explica o fato de que algumas bordas nas imagens, apesar de não estarem em regiões homogêneas, estão classificadas como regiões coerentes.

Analisando os histogramas, é possível perceber que a frequência dos valores pertencentes aos pixeis coerentes é maior que a dos incoerentes. Por outro lado, os pixeis incoerentes apresentam um histograma mais equalizado e preenchido. Isto já era esperado, em virtude do fato de que as regiões dos pixeis coerentes possuem o mesmo valor e contém um grande número de píxeis, enquanto que as regiões incoerentes possuem valores variados e em menor número de pixeis preenchidos. Esta característica é importante pois afetará nos resultados finais, como será visto adiante.

2.3.1 Limiar de Região

A escolha da quantidade de pixeis mínima na região para que ela seja classificada como coerente é um parâmetro essencial. Por isso, necessita de um estudo mais destacado do comportamento do sistema aqui proposto em função desta variável. Para melhor compreensão, este parâmetro será denominado de limiar de região.

O limiar de região depende das características extraídas da imagem. Em (PASS; ZABIH; MILLER, 1997), por exemplo, o limiar escolhido foi de 1% da resolução da imagem para a separação das regiões em coerentes ou incoerentes. Já em (VAILAYA; JAIN; ZHANG, 1998) o valor escolhido para este parâmetro no EDCV foi de 0,1% do tamanho da imagem. Portanto, não foi adotado um valor de limiar absoluto para estas

pesquisas, pois o tamanho da imagem interfere diretamente na classificação dos pixeis em coerentes e incoerentes.

Para a escolha do melhor valor possível no qual a abordagem por ora proposta alcança os melhores resultados, vários valores para o limiar de região foram estimados. Através do banco de dados em (FEI-FEI; PERONA, 2005) os resultados obtidos são ilustrados na Figura 24. O banco de dados foi dividido em duas partes para realização dos testes: as cenas externas, composta pelas classes Montanha, Praia, Floresta, Costa, Rodovia, Cidade, Zona Aberta e Rua e cenas internas composta pelas classes Banheiro, Sala, Cozinha, Escritório e Subúrbio. Este banco de dados será melhor descrito posteriormente. Esta divisão foi feita com o objetivo de analisar o comportamento do LBP Coerente e Incoerente tendo em vista dois tipos de cenas com características diferentes, uma com mais objetos e detalhes e outra com menos bordas e mais regiões homogêneas.

Os testes foram feitos empregando-se como classificador o método do vizinho mais próximo, o k-nn, que será explicado mais detalhadamente nas seções seguintes. Para melhor observação do poder de descrição dos pixeis coerentes e incoerentes os testes foram realizados tanto para os dois histogramas concatenados, como consiste o CCV, quanto para cada um isoladamente. Assim, obteve-se três resultados para cada limiar: LBP Coerente, LBP Incoerente e LBP Coerente-Incoerente. Para melhor compreensão o LBP Coerente-Incoerente será chamado de LBP Duplo.

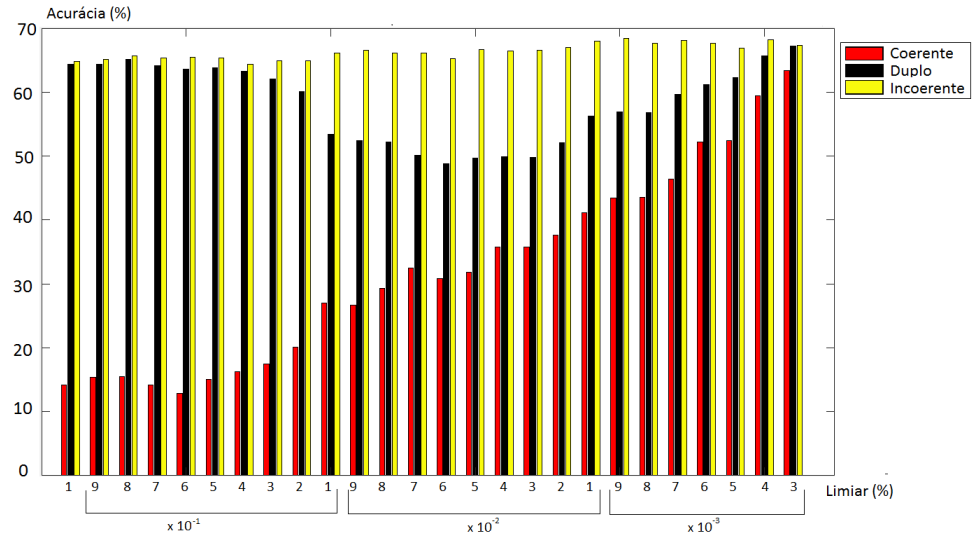
Analisando os gráficos da Figura 24 conclui-se que cada vetor tem um comportamento diferente à medida que o limiar de região é variado. Testes mostraram que o LBP coerente, para valores de limiar próximos à 1 %, apresentou acurácia insatisfatória e à proporção que se diminui o valor de limiar, seus resultados melhoram. Já o LBP Duplo tem seu reconhecimento prejudicado em um primeiro momento, à medida que o limiar de região é reduzido. Entretanto, seu desempenho volta a crescer a partir de um certo ponto. Por fim, o LBP Incoerente possui o reconhecimento mais estabilizado dentre os três vetores com baixa sensibilidade ao limiar escolhido, mas com uma acurácia ligeiramente maior para limiares mínimos.

Optou-se por reduzir este parâmetro até 0,003% da imagem pois abaixo disso, em virtude da resolução do banco de dados, todos os pixeis podem ser classificados como coerentes, já que o limiar de região pode ter valor menor que um.

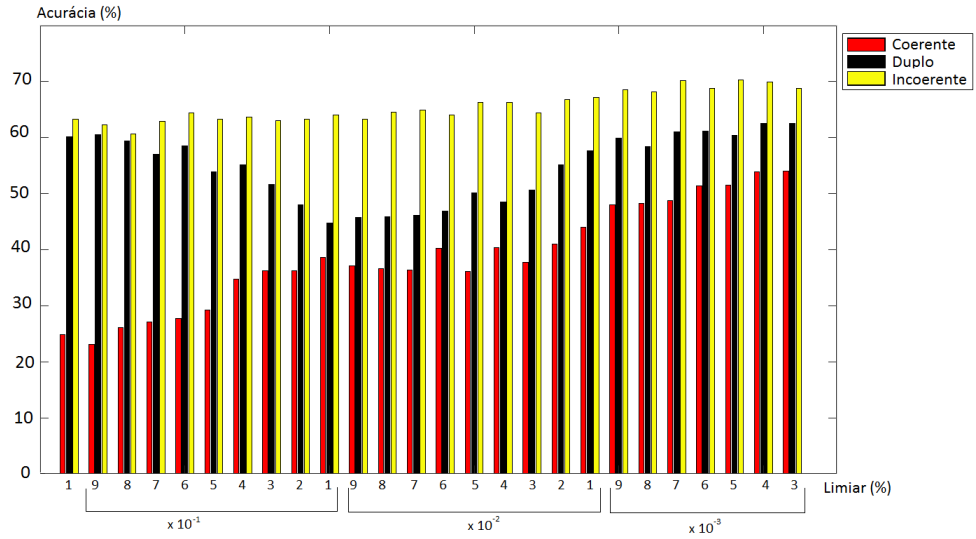
Analisando a Figura 25, em que se encontra a representação dos pixeis coerentes em uma cena externa e interna para diferentes valores de limiar, compreende-se melhor os resultados obtidos na Figura 24. Inicialmente, para o limiar com o valor de 1%, não há pixeis coerentes na imagem. Isto se explica pelo fato de que, em uma imagem gerada pelo LBP, há pouca probabilidade de se ter uma região deste tamanho com pixeis de mesmo valor conectados entre si. Portanto, este limiar alto torna a possibilidade de existir uma região coerente pequena. Isto faz com que o LBP coerente tenha pouca informação

Figura 24 – Reconhecimento dos pixeis coerentes e incoerentes para diferentes valores do limiar de região.

(a) 8 Cenas Externas.



(b) 5 Cenas Internas.

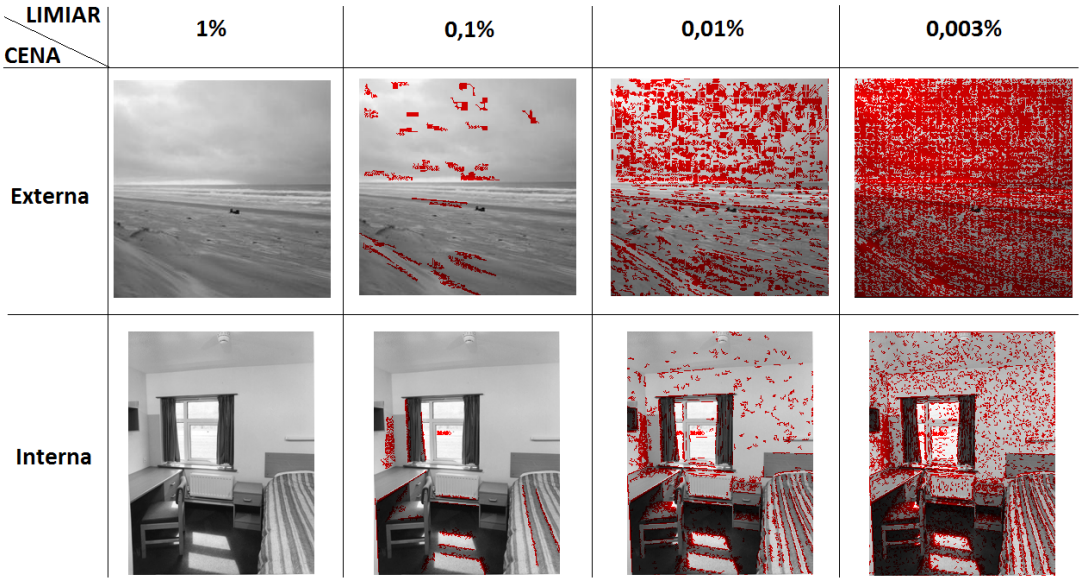


Fonte: Próprio autor

descritiva. Outrossim, em vista deste fato, os vetores LBP Duplo e LBP Incoerente possuem aproximadamente o mesmo conteúdo e, portanto, os mesmos resultados.

À medida em que se diminui o valor do limiar, a possibilidade de haver algumas regiões coerentes aumenta, sobretudo nas regiões mais homogêneas e em bordas com o mesmo padrão de textura (observe a Figura 25 para o limiar de 0,1%). No entanto, os pixeis coerentes, por se apresentarem em pequeno número e em regiões isoladas, proporcionam um poder descritivo ruim à imagem. Com efeito, é acrescentado ao LBP Duplo a informação desses pixeis coerentes, reduzindo o seu reconhecimento em um primeiro momento, como aponta os resultados da Figura 24.

Figura 25 – Pixeis coerentes na imagem para diferentes valores do limiar de região.



Fonte: Próprio autor

Conforme o limiar de região é reduzido, mais as regiões coerentes se mostram distribuídas na imagem, aumentando o seu poder descritivo. Conclui-se desse modo que, quanto menor é o valor do limiar empregado, maior é a acurácia do LBP Coerente, como indicado nos resultados. Uma outra observação importante é que a partir de um certo ponto de limiar, a informação dos pixeis coerentes fazem com que o LBP Duplo retorne ao seu poder descritivo inicial em que o limiar de região era 1% e, à proporção que se diminui mais esse parâmetro, seu reconhecimento aumenta.

Entretanto, observe que o poder descritivo do LBP Incoerente está sempre maior em comparação com o LBP Coerente. Este comportamento independe do valor de limiar estabelecido, note que até para valores mínimos como 0,003%, em que grande parte dos pixeis estão classificados como pixel coerente, como na cena externa da Figura 25, o LBP Incoerente obteve melhor aproveitamento. Não obstante, o LBP Incoerente apresentou resultados melhores que o LBP Duplo cuja dimensão do vetor característico é duas vezes maior que o primeiro, o que torna seu poder descritivo ainda mais relevante.

Outro ponto no qual deve-se destacar através dos resultados apresentados na Figura 24 é a disparidade entre o reconhecimento feito pelo LBP Incoerente em relação ao LBP Coerente em cada tipo de cena. Observe que, quando atingi-se limiares mínimos, a discrepância entre esses dois vetores é menor para as cenas externas do que para cenas internas. Isto pode ser explicado pela Figura 25, onde é observado que as classes externas, para este banco de dados, apresentam mais regiões homogêneas, aumentando assim a quantidade de pixeis coerentes na imagem, tornando-os mais denso. Tal fato contribui no aumento da capacidade discriminativa do LBP Coerente.

Após a realização de alguns testes, foi observado que o limiar de 0,005% proporciona

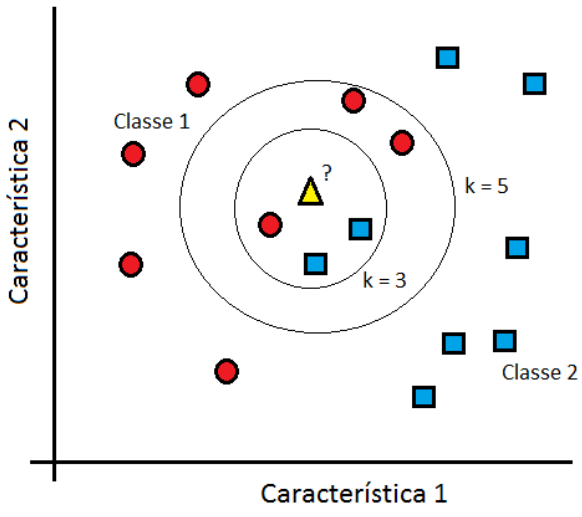
os melhores resultados. Portanto, será utilizado este valor para a implementação do LBP Incoerente e para os experimentos futuros realizados.

2.3.2 Classificador k-nn

O método do vizinho mais próximo, ou popularmente conhecido como k-nn (*k nearest neighboors*), é um método de aprendizagem supervisionada em que a classificação de uma amostra será determinada em função da maioria das classes das k amostras mais próximas dela, através de uma função de distância (COVER; HART, 1967). A escolha do parâmetro k é fundamental para o sucesso do reconhecimento e este depende da base de dados de treinamento obtida.

A Figura 26 ilustra a importância da escolha do parâmetro k na classificação, em que se observa um exemplo de algumas amostras de treinamento inseridas num espaço dimensional. Cada amostra possui uma cor relacionada à classe na qual ela pertence, lembre-se que o k-nn é um método de aprendizagem supervisionada no qual na fase de aprendizado as classes das amostras de treinamento são conhecidas. É necessário portanto, classificar a amostra de teste de acordo com a distribuição das classes mais próximas a ela. Observe na Figura 26 que para dois valores de k escolhidos, duas classificações diferentes serão obtidas.

Figura 26 – Influência do parâmetro k no classificador k-nn.



Fonte: Próprio autor

Uma das principais características da utilização do k-nn é a sua facilidade de implementação, visto que emprega conceitos matemáticos de baixo custo computacional. Outro ponto positivo a ser destacado é que sua probabilidade de erro é menor ou igual a duas vezes a probabilidade de erro de Bayes, para k=1 (RAJA; ROOMI; DHARMALAKSHMI, 2013).

Ademais, no k-nn trabalha-se com o próprio vetor descritivo sem a necessidade de outras etapas ou variáveis. Através de seus resultados é possível ter uma boa percepção a respeito da representação da imagem por meio das características extraídas pelo algoritmo, já que este classificador não depende de parâmetros específicos da área aplicada. Devido à estas características, este classificador tem sido bastante usado em várias aplicações como reconhecimento de padrões, detecção de *outliers* e reconhecimento de objeto (KHASTAVANEH; EBRAHIMPOUR-KOMLEH; HANAEE-AHWAZ, 2017).

Como desvantagens pode-se citar que, com um grande banco de dados, o custo computacional do k-nn poderá ser elevado, uma vez que são calculadas as distâncias entre todas as amostras, uma por uma. Outra desvantagem deste algoritmo é a sua intolerância a ruídos, além de prover pouca informação sobre a estrutura de dados (KUNCHEVA, 1959).

A escolha da função de distância do k-nn é fundamental para o sucesso dos resultados. Na literatura, encontra-se algumas funções como *Mahalanobis*, Euclidiana, Chi-quadrado e Interseção de Histograma (KHASTAVANEH; EBRAHIMPOUR-KOMLEH; HANAEE-AHWAZ, 2017). Nesse trabalho utiliza-se como função de distância a norma Euclidiana, definida da seguinte forma:

$$d(x, y) = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2}, \tag{2.4}$$

em que x e y são duas amostras distintas e N o tamanho do vetor descritivo. Para a realização de testes com o CCV, por possuir dois vetores descritivos gerados, em (PASS; ZABIH; MILLER, 1997) foi utilizado o método do vizinho mais próximo com distância euclidiana da seguinte maneira:

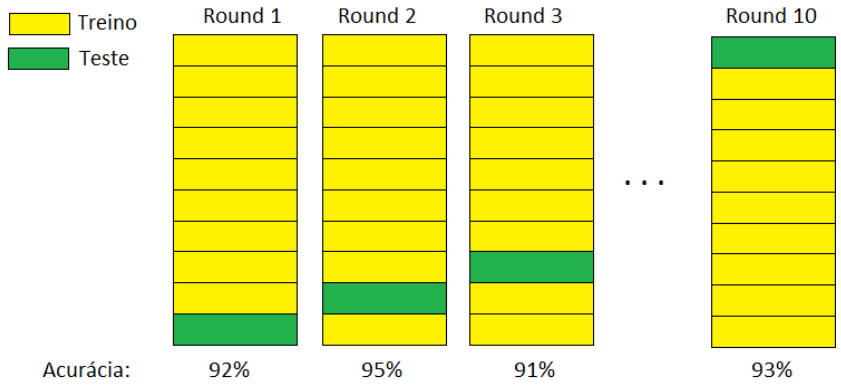
$$d(x, y) = \sqrt{\sum_{i=1}^{N}(\alpha_{x,i} - \alpha_{y,i})^2 + (\beta_{x,i} - \beta_{y,i})^2}, \tag{2.5}$$

em que α_x e β_x são os histogramas de pixeis coerentes e incoerentes respectivamente da amostra x. Assim, calcula-se apenas as diferenças entre os histogramas de mesma classe. Note que não é necessário se preocupar caso o histograma coerente de uma imagem seja parecido com o histograma incoerente de outra imagem, posto que eles não são comparados entre si. Nesta dissertação, para testes em que utilizava-se informações de pixeis coerentes e incoerentes juntos, adotou-se a equação 2.5.

2.3.3 Influência dos *Buckets*

Como explicado anteriormente o valor do número de *buckets*, ou seja, o número de intervalos de quantização dos pixeis, interfere diretamente na construção do histograma e, portanto, no tamanho do vetor descritivo do LBP Incoerente. Para compreender melhor a

Figura 27 – Método de validação *10-fold-cross-validation*.



Fonte: Próprio autor

relevância da escolha do número de *buckets* na descrição do LBP Incoerente alguns testes foram realizados.

Inicialmente, as imagens coloridas foram convertidas em níveis de cinza. Posteriormente, é aplicado o *10-fold-cross-validation*. Este método separa o banco de dados em 10 pastas aleatoriamente utilizando uma pasta para teste e as outras como treino (MANDHALA; SUJATHA; DEVI, 2014). Realiza-se, então, este procedimento para cada pasta separada, como ilustra a Figura 27. Ao final, calcula-se a acurácia média dos 10 testes obtidos e o seu desvio padrão. A acurácia é uma medida de desempenho de classificação avaliada através da seguinte equação:

$$a = \frac{v_p}{f_p + v_p}, \tag{2.6}$$

em que v_p é o número de verdadeiro positivos e f_p o número de falso positivos. O cálculo do desvio padrão é obtido através da expressão:

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(a_i - \mu_a)^2}, \tag{2.7}$$

em que N é o número de testes e μ_a é a média das amostras.

A Tabela 1 mostra os resultados adquiridos para as 5 cenas internas e 8 cenas externas de acordo com a discretização dos níveis de cinza do LBP. Observe que, mesmo reduzindo o total de possíveis valores de níveis de cinza, e consequentemente o tamanho do histograma, os resultados não sofrem uma redução considerável, principalmente para cenas internas. Isto ocorre porque um maior número de detalhes em uma imagem sugere uma menor quantidade de níveis de cinza necessária para representá-la. Ou seja, em cenas internas onde geralmente encontra-se uma grande quantidade de objetos, a redução do tamanho do histograma da imagem pode não prejudicar seu conteúdo (GONZALEZ; WOODS, 2006).

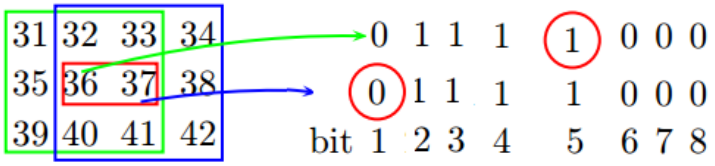
Tabela 1 – Resultados adquiridos utilizando o k-nn entre diferentes discretizações para o LBP Incoerente no banco de dados de (FEI-FEI; PERONA, 2005).

Buckets	5 Cenas Internas	8 Cenas Externas
256	70,0% ± 4,2	**69,4% ± 1,5**
128	69,2% ± 4,4	68,0% ± 1,9
64	**70,3% ± 4,1**	68,3% ± 2,8
32	68,9% ± 5,1	66,6% ± 2,6

Como na Transformada *Census*, o LBP possui uma importante propriedade, a alta correlação entre os pixeis vizinhos. Esta característica permite ao LBP perder alguma informação e seu poder descritivo não ser prejudicado. Desta maneira, em alguns trabalhos que empregam o LBP é encontrado o uso de algoritmos estatísticos para redução do dimensionamento do vetor gerado pelo LBP, como o PCA e a matriz de co-ocorrência por exemplo, e os resultados permanecerem satisfatórios (LIU; MOON, 2016).

Essa alta correlação ocorre pelo fato do LBP trabalhar na imagem com todos os pixeis e com seus vizinhos podendo, desta forma, realizar a mesma comparação duas vezes entre eles. A Figura 28 ilustra melhor este processo, observe que em um primeiro momento, quando o pixel central é o pixel 36 há uma comparação com o pixel 37 durante a formação do valor LBP. Essa mesma comparação ocorre quando o pixel 37 é o pixel central, porém o bit terá outro valor na comparação gerada, visto que se alterou o pixel central de referência da primeira comparação. O mesmo ocorre para os outros vizinhos, ou seja, caso os dois pixeis não sejam iguais, uma comparação gerará o bit 1 em uma determinada posição e outra comparação gerará o bit 0 em outra posição (WU; REHG, 2011).

Figura 28 – Alta correlação entre os pixeis na formação do valor LBP. Adaptado de Wu e Rehg (2011).



Pode-se citar também o trabalho de (OJALA; PIETIKAINEN; MAENPAA, 2002) na construção do LBP invariante à rotação, já mencionado anteriormente. Analisando as correlações entre os pixeis vizinhos na operação do LBP, a frequência em que certos valores são gerados na imagem e as alterações bit a bit no valor binário formado, reduziu-se os 255 valores possíveis do $LBP^{riu2}_{(8,1)}$ em apenas 36 valores.

2.4 Resultados com o LBP Incoerente

Nesta seção será analisado se o algoritmo apresentado contribui para uma melhor representação descritiva da cena. Para tal fim, foi utilizado o método de validação *10-fold-cross-validation*. Com os 10 resultados adquiridos, a média e o desvio padrão das acurácias são calculados, assim como foi feito para os testes de limiar explicado anteriormente.

Como o LBP não recorre à informação de cor, inicialmente todas as imagens coloridas são transformadas para nível de cinza e normaliza-se o histograma de forma que o vetor tenha norma euclidiana igual a 1. Além disso, são excluídos os valores extremos 0 e 255 gerados pelo LBP, assim como é feito na construção do CENTRIST (WU; REHG, 2011). Esta operação foi feita em virtude da grande frequência que esses dois bits aparecem na imagem, polarizando o histograma e prejudicando o reconhecimento.

2.4.1 Banco de Dados

Para os testes realizados com o LBP Incoerente e classificador k-nn, três bases de dados são utilizadas:

- 8 cenas – Publicado em (OLIVA; TORRALBA, 2001), este banco de dados dispõe de 2688 imagens coloridas com resolução de 256 x 256 pixeis, contendo cenas de ambientes abertos, no qual serão denominadas de cenas externas, separadas em 8 classes: Costa (360 imagens), Floresta (328 imagens), Montanha (274 imagens), Zona Aberta (410 imagens), Rodovia (260 imagens), Cidade (308 imagens), Prédio (356 imagens) e Rua (292 imagens).
- 13 cenas – Publicado em (FEI-FEI; PERONA, 2005), este banco de dados contém além das cenas disponibilizadas em (OLIVA; TORRALBA, 2001), cenas divididas em 5 classes em níveis de cinza: Quarto (216 imagens), Cozinha (210 imagens), Sala (289 imagens), Subúrbio (215 imagens) e Escritório (215 imagens). Essas cenas serão denominadas de cenas internas. As imagens possuem dimensão média de 250 x 300 pixeis.
- 15 cenas – Proposto em (LAZEBNIK; SCHMID; PONCE, 2006), é uma extensão dos dois bancos de dados acima. Contém, além das imagens de (FEI-FEI; PERONA, 2005), mais duas classes em níveis de cinza: Industrial (311 imagens) e Loja (315 imagens). As imagens possuem dimensão média de 250 x 300 pixeis, porém a resolução para cada imagem desta duas classes podem variar.

2.4.2 Experimentos

A Tabela 2 apresenta os resultados adquiridos com o LBP Incoerente com diversos valores de *buckets* em comparação com outros dois algoritmos conhecidos na literatura,

Tabela 2 – Resultados adquiridos utilizando o k-nn para diferentes bancos de dados.

Descrito	5 Cenas Internas	8 Cenas	13 Cenas	15 Cenas
LBP Incoerente - 256	70,0% ± 4,2	69,4% ± 1,5	64,2% ± 1,3	59,2% ± 1,8
LBP Incoerente - 128	69,2% ± 4,4	68,0% ± 1,9	63,8% ± 2,3	59,1% ± 2,7
LBP Incoerente - 64	**70,3% ± 4,1**	68,3% ± 2,8	63,6% ± 2,0	58,3% ± 2,2
LBP Incoerente - 32	68,9% ± 5,1	66,6% ± 2,6	60,6% ± 3,1	55,7% ± 2,0
LBP	64,6% ± 4,1	66,9% ± 1,8	62,3% ± 1,2	58,7% ± 2,8
Gist	60,2% ± 3,1	**77,2% ± 3,5**	**70,2% ± 1,7**	**65,8% ± 2,6**

LBP e *gist*. Note que para as cenas internas o LBP Incoerente teve um aproveitamento superior ao *gist*, mas nos outros bancos de dados o *gist* se sobressaiu. O resultado já era esperado visto que o algoritmo proposto em (OLIVA; TORRALBA, 2001) está voltado para cenas com poucos detalhes e com regiões homogêneas ocupando grande parte da imagem, como a maioria das cenas nos bancos de dados em que ele foi melhor.

O LBP Incoerente com 256 e 128 *buckets* superou em todos os casos o histograma do LBP, justificando o poder descritivo por parte dos pixeis incoerentes. Um outro ponto interessante a ser observado é que na maioria dos casos, mesmo com um vetor quatro vezes menor, o LBP Incoerente-64 se saiu melhor que o LBP original. Além disso, o LBP Incoerente-32 apesar de ter seu vetor descritivo 16 vezes menor que o *gist* atingiu melhores resultados para as 5 cenas internas.

Analisando os resultados observa-se que o LBP Incoerente sofreu uma grande queda de aproveitamento em comparação ao LBP quando as classes Loja e Indústria são adicionadas no banco de dados (15 Cenas). Um dos motivos para o baixo reconhecimento é que elas possuem resoluções variadas, prejudicando a separação de regiões coerentes e incoerentes e, consequentemente, a representação do vetor descritivo por eles. Uma solução adotada nos testes posteriores é a padronização da resolução das imagens no banco de dados,que será explicada no próximo capítulo. A Tabela 3 mostra a acurácia do LBP Incoerente-256 e do LBP para todas as classes, observe como a classe Loja prejudica o reconhecimento do LBP Incoerente em relação ao LBP original.

Tabela 3 – Resultados adquiridos utilizando o k-nn para todas as classes de (LAZEBNIK; SCHMID; PONCE, 2006)).

Classe	LBP-Incoerente 256	LBP
Costa	**77%**	73%
Floresta	86%	**89%**
Estrada	**78%**	73%
Cidade	**68%**	67%
Montanha	**64%**	52%
Zona Aberta	59%	**60%**
Rua	**70%**	61%
Prédio	38%	**44%**
Quarto	**15%**	14%
Subúrbio	**96%**	95%
Cozinha	51%	51%
Sala	57%	**60%**
Escritório	**78%**	73%
Loja	55%	**71%**
Indústria	30%	**31%**

3 Contextual Modified Local Binary Pattern Incoerente

Nesta seção é apresentado um descritor que, além das propriedades do LBP Incoerente , captura informações de contexto, ou seja, informações de estruturas vizinhas ao pixel central. Serão apresentadas as técnicas empregadas para a formação desse descritor, o classificador utilizado, além de testes obtidos com os mesmos parâmetros do LBP Incoerente como limiar de região e número de *buckets*.

O *Contextual Modified Local Binary Pattern* Incoerente (CMLBP Incoerente) é um descritor holístico que modela as propriedades estruturais da cena, através do histograma das informações locais extraídas pelo LBP, adicionando conhecimento a respeito das características em torno dos pixeis vizinhos ao pixel central. Assim, leva-se em conta o ambiente em que aquele pixel está inserido e as suas estruturas vizinhas.

Não obstante, o CMLBP Incoerente explora a informação espacial dos pixeis vizinhos e a relação entre eles na imagem gerada pelo CMLBP, com o objetivo de descartar as regiões homogêneas formadas na imagem. O processo é o mesmo que descrito anteriormente na extração das regiões coerentes e incoerentes do LBP Incoerente, com a diferença de que será trabalhado com outra imagem, gerada por outra técnica não-paramétrica.

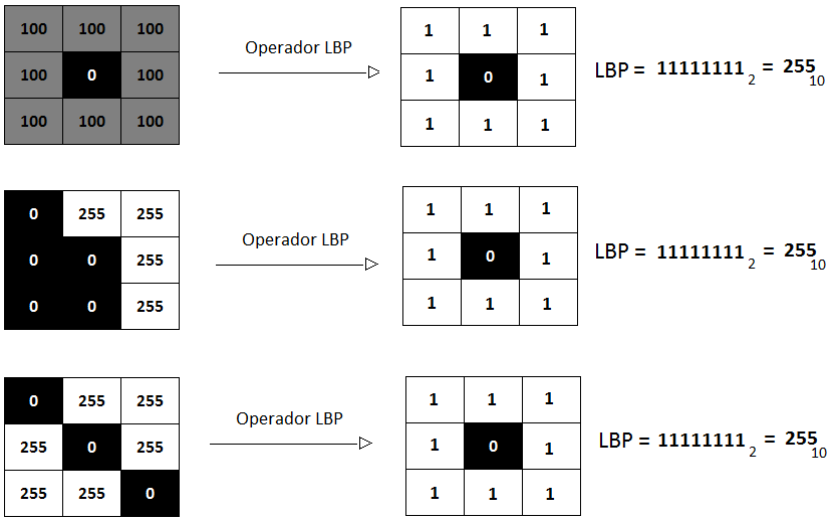
Ao decorrer deste capítulo serão apresentadas as técnicas que inspiraram a criação deste método, bem como seu poder de descrição em relação ao LBP Incoerente e outros métodos já empregados na literatura, a fim de verificar se a informação de contexto melhora o desempenho.

3.1 Modified Census Transform

Observando algumas limitações que a transformada Census pode proporcionar, Froba e Ernst (2004) criaram um descritor onde a referência para a comparação não é o valor absoluto do pixel central mas a média de valores da janela composta por ele e seus 8 pixeis vizinhos.

A Figura 29 mostra a deficiência que o LBP e a transformada *Census* possuem quando o pixel central atinge valores muito altos ou muito baixos. O pixel central nos três casos possui valor mínimo. Aos pixeis brancos é atribuído o valor 1 ao final da comparação, através da equação 2.2. Note que, em virtude do pixel central possuir o valor mínimo, após as comparações serem realizadas, o LBP sempre terá um mesmo valor final não importando a distribuição dos pixeis vizinhos. Portanto, a informação a respeito da estrutura em volta

Figura 29 – Mesmo valor gerado pelo LBP para diferentes estruturas espaciais.



Fonte: Próprio autor

do pixel central é perdida e não irá interferir no valor gerado. Esta situação sempre irá ocorrer quando o operador do LBP se deparar com o pixel central menor ou igual aos seus pixeis vizinhos na janela de operação.

Buscando uma solução para este problema que atinge as transformadas não-paramétricas locais, em (FROBA; ERNST, 2004) foi proposto o *Modified Census Transform* (MCT), onde o valor de comparação passa a ser a média do pixel central com os pixeis vizinhos. Neste algoritmo, o pixel central também é comparado com o valor de referência, gerando um número binário de 9 posições, ou seja, um número decimal com valores correspondentes a uma faixa de 0 a 511. Portanto, o tamanho do histograma gerado pelo MCT é o dobro do tamanho gerado pelo LBP.

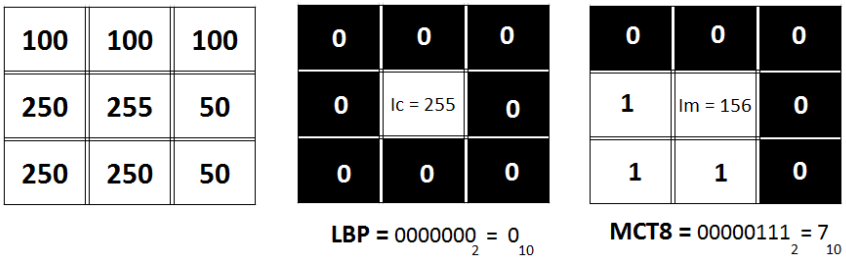
Para evitar esse aumento considerável no vetor descritivo, Gazolli e Salles (2014) modificaram o MCT e construíram o MCT8, em que o pixel central não é comparado com o valor de referência. Desta forma, tem-se novamente 8 bits no número binário gerado, como no LBP. Algebricamente, esta operação pode ser representada da seguinte forma:

$$MCT8(x,y) = \sum_{n=0}^{7} s(i_n - i_m)2^n, \tag{3.1}$$

em que i_m é a média da intensidade de todos os nove pixeis correspondentes na janela de operação. A função *s(z)* é a mesma apresentada na equação 2.2. A Figura 30 exemplifica o resultado final do MCT8 e do LBP para uma determinada situação. Tem-se uma janela de tamanho 3x3 em que os pixeis nas bordas da direita e superior possuem valores menores que os demais pixeis. Nesta figura os pixeis pretos estão representados com valor zero e os pixeis brancos com valor um. Repare que o MCT8, neste exemplo, descreveu melhor a relação do pixel central com seus pixeis vizinhos.

O motivo para esta melhor representação está no valor do pixel central para as comparações com os oito pixeis vizinhos. Para o LBP o valor absoluto do pixel central é maior que todos os seus vizinhos. Portanto, não importa a distribuição dos pixeis nesta janela local, o valor do LBP gerado será sempre zero. O mesmo não ocorre com o MCT8, pois para a atribuição do valor de referência é necessário primeiramente calcular a média dos pixeis nesta janela. Portanto, os pixeis vizinhos interferem no cálculo do valor central.

Figura 30 – Resultados diferentes gerados pelos operadores LBP e MCT8 sobre uma mesma janela de 3x3 pixeis.



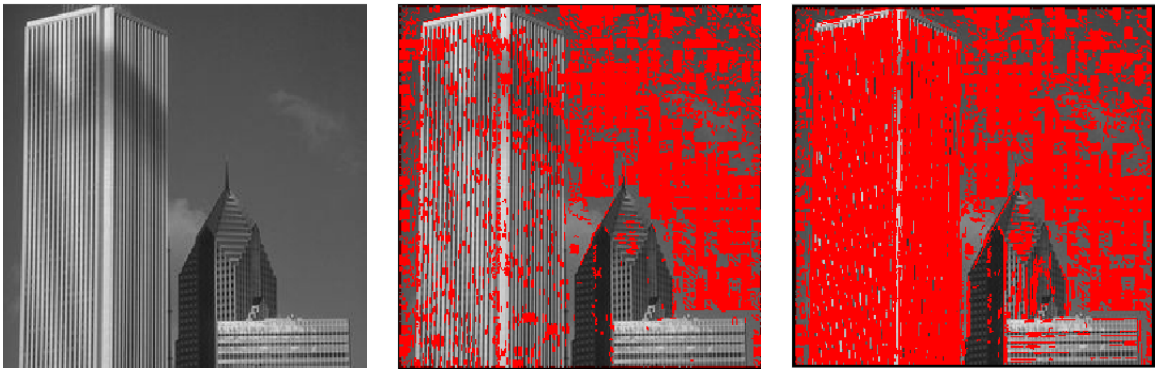
Fonte: Próprio autor

3.1.1 MCT8 Incoherent

O MCT8 se mostrou um vetor com maior poder descritivo em relação ao CENTRIST nos testes de classificação de cenas realizados em (GAZOLLI; SALLES, 2014). No entanto, os resultados adquiridos pelo MCT8 Incoerente não foram satisfatórios em comparação ao LBP Incoerente. O principal motivo para esse baixo desempenho é o fato de que as operações realizadas pelo MCT8 proporcionam mais regiões homogêneas, que posteriormente são descartadas para a construção do histograma incoerente.

Figura 31 – Pixeis coerentes gerados pelo LBP e MCT8.

(a) Image Original. (b) LBP Incoerente. (c) MCT8 Incoerente.



Fonte: Próprio autor

A Figura 31 exemplifica essa situação. Na Figura 31.a se observa uma imagem original da classe Prédio e sua representação pelos pixeis coerentes do LBP na Figura 31.b

e do MCT8 na Figura 31.c. Repare que algumas estruturas, que podem ser importante para o reconhecimento de uma cena, como o prédio à esquerda da imagem e parte das janelas do prédio à direita, são identificadas quase em sua totalidade como pixeis coerentes pelo MCT8. Isto posto, a representação da imagem feita pelos pixeis incoerentes do MCT8 perde poder descritivo.

Outrossim, regiões homogêneas que poderiam ter uma maior quantidade de pixeis coerentes com o MCT8, como o céu por exemplo, não sofrem uma grande alteração como mostra a Figura 31.c. Desta forma, o MCT8 Incoerente em relação ao LBP Incoerente descarta regiões que poderiam ser importantes para o reconhecimento da cena, ao passo que regiões não tão importantes para a descrição da cena e que poderiam ser descartadas não as são.

Analisando os resultados da Tabela 4 percebe-se que, de fato, estes fatores proporcionaram ao MCT8 Incoerente um pior desempenho em relação ao LBP Incoerente, principalmente para a classe Escritório. Enquanto que para o banco de dados em (LAZEBNIK; SCHMID; PONCE, 2006), o LBP Incoerente-256 obteve um reconhecimento de 59,2% ± 1,8, o MCT8 Incoerente-256 obteve 55,7% ± 2,2. O limiar de região para os testes com o MCT8 Incoerente foi de 0,005%.

Tabela 4 – Resultados adquiridos utilizando o k-nn para todas as classes de (LAZEBNIK; SCHMID; PONCE, 2006).

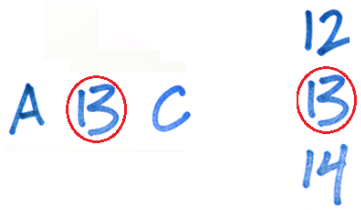
Classe	LBP-Incoerente 256	MCT8 -Incoerente 256
Costa	**77%**	73%
Floresta	86%	**90%**
Estrada	**78%**	77%
Cidade	**68%**	63%
Montanha	**64%**	58%
Zona Aberta	**59%**	56%
Rua	**70%**	68%
Prédio	**38%**	31%
Quarto	15%	**23%**
Subúrbio	**96%**	95%
Cozinha	**51%**	42%
Sala	57%	**66%**
Escritório	**78%**	35%
Loja	**55%**	50%
Indústria	30%	**34%**

3.2 Contextual Modified Census Transform

Para melhorar a representação da imagem, Gazolli e Salles (2014) propuseram o Contextual Modified Census Transform (CMCT). A percepção é motivada pelo fato de que estruturas muito similares podem estar inseridas em ambientes diferentes. Assim, identificando a estrutura da vizinhança, classifica-se melhor as estruturas locais. Para classificação de cenas por exemplo, um conjunto de pixeis brancos pode ser classificado como neve na imagem, porém caso se saiba que a sua vizinhança é composta por pixeis azuis, é possível supor que este conteúdo se trata de uma nuvem na verdade.

A Figura 32 exemplifica a importância do contexto para identificação de elementos no sistema visual humano. Observe que, analisando em torno do elemento central, a classificação é realizada em dois tipos nas duas estruturas mostradas: pela letra B ou pelo número 13. Todavia, as duas representações são absolutamente iguais, ou seja, a classificação da estrutura é baseada no contexto em que ela está inserida.

Figura 32 – Importância do contexto na identificação de elementos.



Fonte: Próprio autor

Um outro exemplo no qual é notável a importância do contexto na descrição de uma imagem é a Figura 33. Pela imagem borrada à esquerda observa-se que se trata de uma pessoa em um local de trabalho com um telefone na mão, monitor, mouse e teclado à sua frente, além de um estabilizador no canto da mesa. Entretanto, quando a imagem original é revelada, com exceção do teclado, percebe-se que todos os outros objetos citados se tratam de elementos diferentes e atípicos para essa situação. Ou seja, a suposição de determinados objetos foi influenciada pelo contexto em que eles estavam inseridos.

Portanto, utilizar a informação de contexto pode proporcionar melhores resultados na representação de cenas. Tem-se aqui um desafio, acrescentar esta informação de forma que o custo computacional não tenha um aumento elevado utilizando as transformadas não-paramétricas. Para isso, o CMCT adota uma abordagem recursiva, através da criação de novas estruturas locais formadas a partir de estruturas locais vizinhas, associando assim, a cada estrutura identificada, informações da sua vizinhança (GAZOLLI; SALLES, 2014).

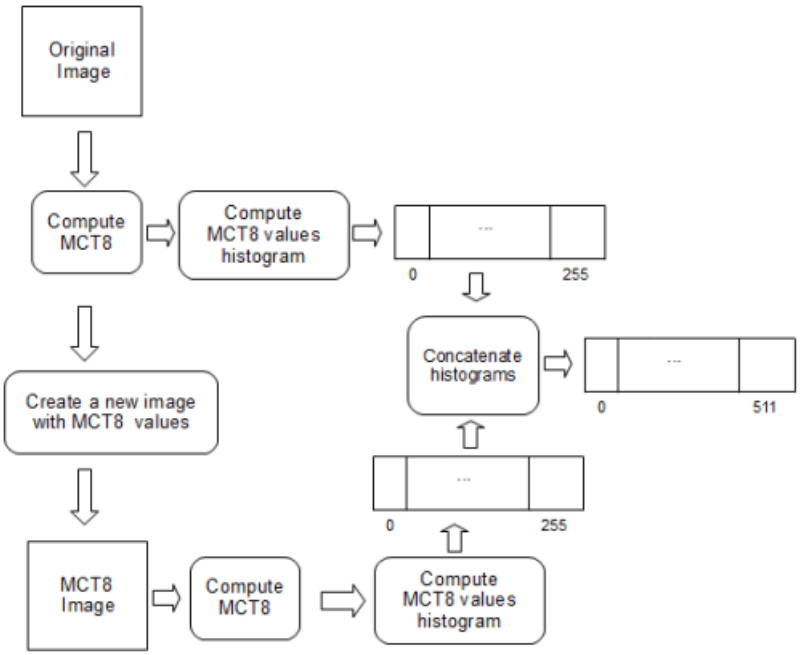
A Figura 34 ilustra o processo de construção do vetor descritivo do CMCT. Primeiramente, o MCT8 é computado sobre a imagem original, gerando uma nova imagem (MCT8Image) e um vetor descritivo pelo seu histograma. Por fim, aplica-se o algoritmo MCT8 novamente sobre a MCT8Image gerando uma outra imagem e um outro histograma.

Figura 33 – Erros de interpretação baseados na informação de contexto.

(a) Imagem borrada.

(b) Imagem original.

Figura 34 – Descrição do processo do algoritmo CMCT.

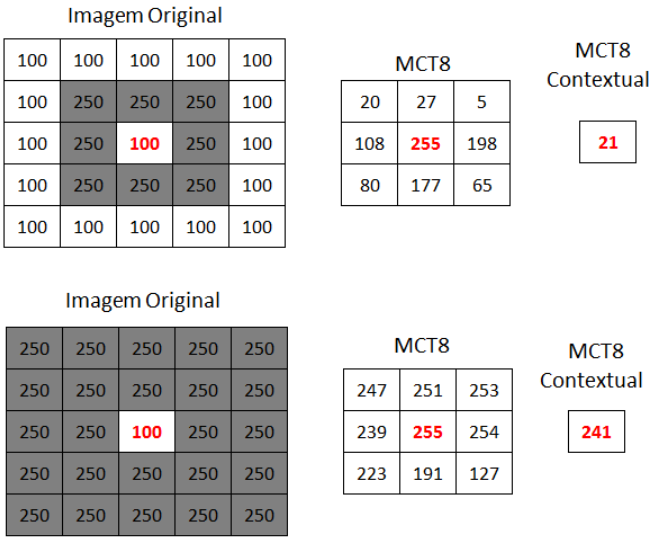


Fonte: Gazolli e Salles (2014)

O vetor descritivo final com 512 posições é a concatenação entre esses dois histogramas gerados.

Por consequência, quando o operador MCT8 é computado sobre a ”MCT8 Image”, o valor final é resultante da estrutura local formada a partir de estruturas locais da imagem original. Assim, quando há duas estruturas similares cercadas por diferentes conjuntos de estruturas locais, o valor do operador MCT8 pode ser o mesmo, porém aplicando o MCT8 novamente, este valor poderá ser diferente nas duas situações. Para melhor entendimento, ao aplicar o MCT8 duas vezes, essa operação será chamada de MCT8 Contextual. A Figura 35 exemplifica este fato, em que é ilustrado o valor resultante para o MCT8 e MCT8 Contextual computado para duas situações com estruturas vizinhas diferentes.

Figura 35 – Comparação entre MCT8 e MCT8 Contextual para diferentes vizinhanças.



Fonte: Próprio autor

Observe que, mesmo com a diferença de contexto, o valor MCT8 final do pixel central não é afetado, pois é computado apenas a informação da janela, não importando o contexto no qual ela está inserida. Não obstante, quando aplica-se o operador MCT8 novamente sobre os valores computados pelo próprio MCT8, gerando o MCT8 Contextual, a informação de contexto acerca da vizinhança desta janela se torna relevante e interfere no valor final do pixel central. Consequentemente, os dois valores finais atribuídos ao pixel central se tornam distintos.

3.2.1 CMLBP Incoerente

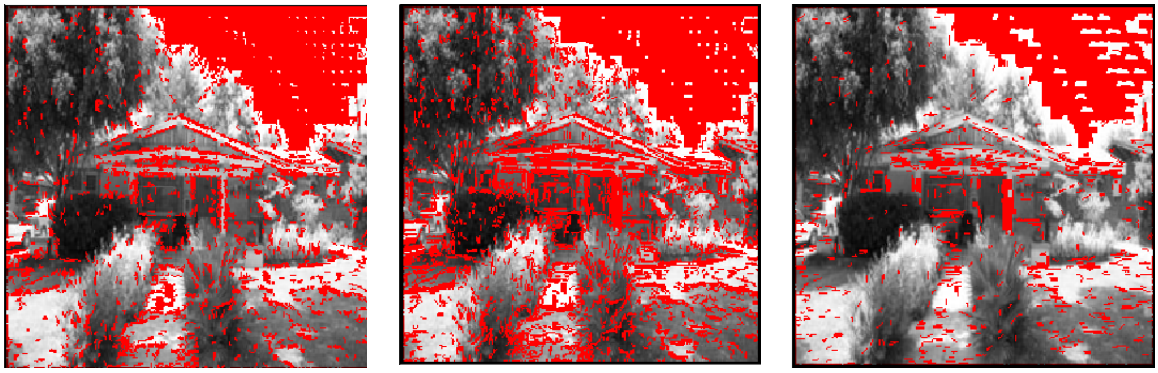
O *Contextual Modified Local Binary Pattern* Incoerente (CMLBP Incoerente) é um descritor que integra informações espaciais e de contexto sobre as estruturas locais identificadas pelo LBP. Assim como o LBP Incoerente, este algoritmo separa os pixeis das regiões coerentes e incoerentes, formando um histograma dos pixeis incoerentes na imagem gerada pelo MCT8 Contextual. O vetor descritivo final é a concatenação deste histograma com o LBP Incoerente.

A Figura 36 mostra a distribuição dos pixeis coerentes sobre as imagens geradas pelo LBP, MCT8 e MCT8 Contextual para melhor compreensão a respeito do comportamento destes algoritmos. Perceba que o MCT8 Contextual captura menos pixeis em relação ao MCT8, sobretudo nas regiões homogêneas como o céu.

Outro ponto observado é que, ao contrário dos pixeis coerentes gerados pelo LBP, que captura padrões de textura uniformes como as bordas da casa por exemplo, os pixeis coerentes do MCT8 Contextual ignoram esta informação. Perceba que isto ocorre em virtude do fato de que, ao se aplicar o operador MCT8 duas vezes sobre um pixel, a

Figura 36 – Distribuição dos pixeis coerentes pára diferentes algoritmos.

(a) LBP. (b) MCT8. (c) MCT8 Contextual.



Fonte: Próprio autor

probabilidade de seus pixeis vizinhos possuírem o mesmo valor é pequena, pois é levado em consideração o contexto nesses casos, como a Figura 35 ilustrou.

Assim, para se ter um mesmo valor MCT8 Contextual igual, o pixel central deve estar inserido em uma região na qual as estruturas locais dos pixeis vizinhos a ele sejam iguais também. Isto explica o fato de que muitos pixeis no céu na Figura 36.c não são classificados como coerentes, sobretudo os que estão próximos de outros elementos como as árvores, já que o contexto interfere no valor atribuído.

Outro ponto a ser observado é que o MCT8 Contextual recebe do MCT8 a capacidade de capturar os pixeis coerentes em regiões homogêneas que o LBP Coerente não captura, como por exemplo as paredes da casa na Figura 36.

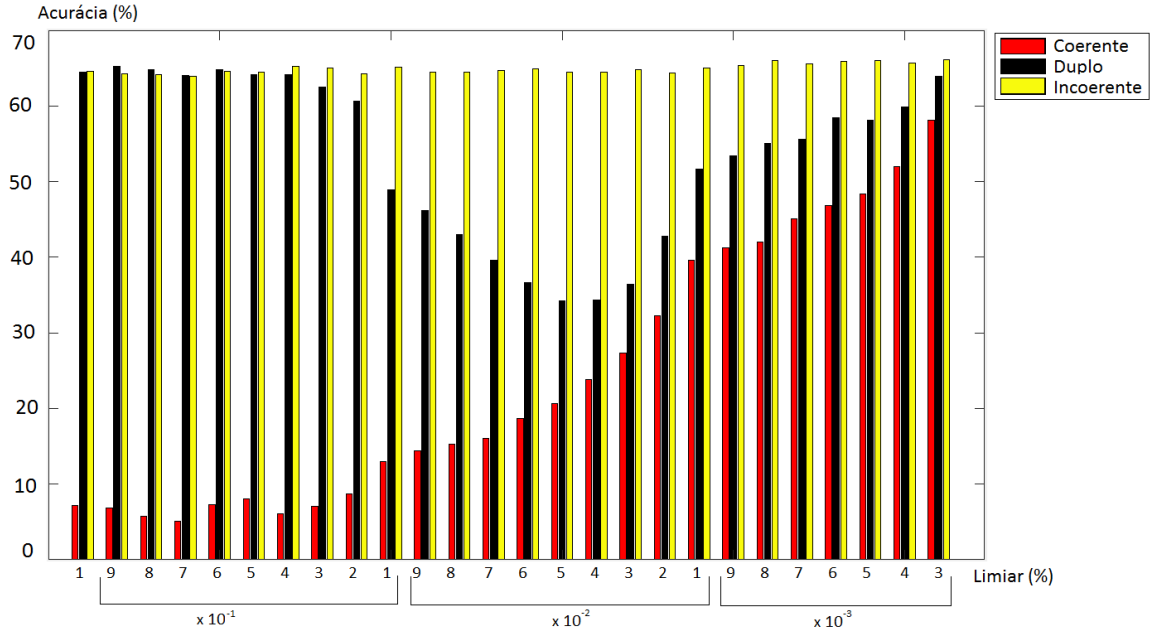
Por conseguinte, observa-se que em alguns casos o MCT8 Contextual e o LBP possuem regiões coerentes que são extraídas em situações diferentes. A concatenação dessas duas abordagens, pode ser de bastante utilidade, visto que o uso de ambas as técnicas podem complementar informações.

3.2.2 Limiar de Região

Assim como o LBP Incoerente, a escolha do limiar de região interfere diretamente na distribuição dos pixeis coerentes e incoerentes da imagem e consequentemente no poder descritivo do vetor final. A Figura 37 mostra a taxa de reconhecimento dos histogramas formados a partir da imagem gerada pela operação do MCT8 duas vezes, gerando o MCT8 Contextual. Da mesma forma que para o LBP, será analisado o comportamento dos histogramas coerentes, incoerentes e ambos juntos.

O banco de dados é o de 15 cenas de (LAZEBNIK; SCHMID; PONCE, 2006), já descrito anteriormente. Para os experimentos foi utilizado o classificador *Support Vector Machine* (SVM), no qual será explicado na próxima sessão.

Figura 37 – Reconhecimento dos pixeis coerentes e incoerentes em função do limiar de região para o MCT8 Contextual.



Fonte: Próprio autor

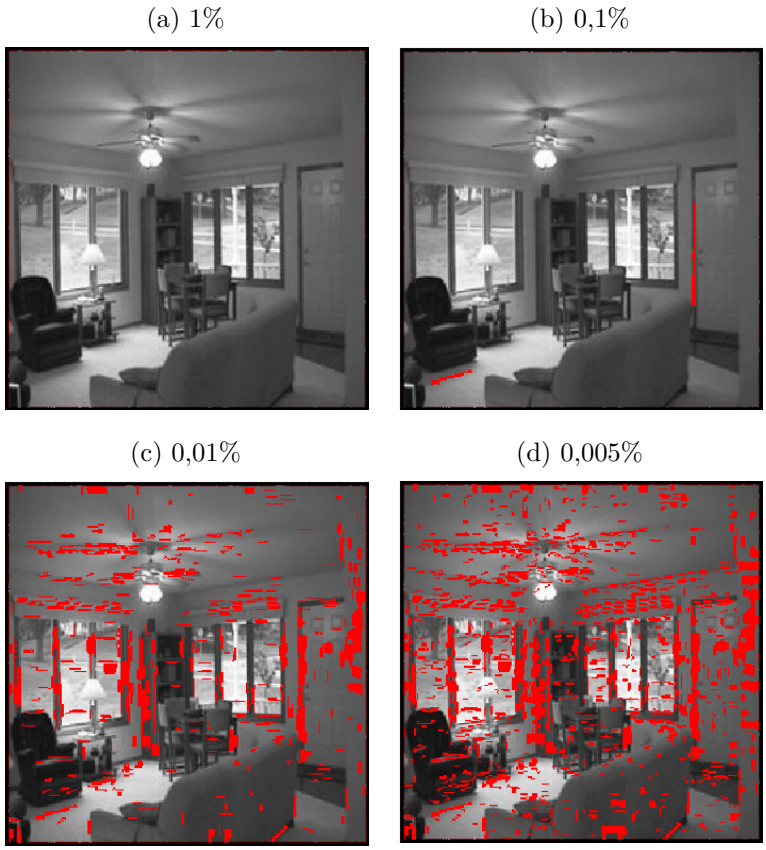
A técnica de validação empregada é a mesma utilizada em (LAZEBNIK; SCHMID; PONCE, 2006). Para cada classe captura-se um conjunto de 100 imagens aleatórias para treinamento e 70 imagens para teste. Esse procedimento é repetido cinco vezes e, ao final, calcula-se a média e o desvio padrão das acurácias apresentadas.

Analisando o gráfico da Figura 37 é possível extrair algumas conclusões. Perceba que o desempenho dos três histogramas gerados pelos MCT8 Contextual é semelhante aos três histogramas gerados pelo LBP, mostrados na Figura 24. A Figura 38 ressalta essa semelhança ilustrando a distribuição dos pixeis para diferentes valores de limiar. Observe que à medida em que se reduz o valor do limiar de região, mais pixeis coerentes são formados e maior é o poder de descrição do vetor final produzido por ele. Note, entretanto, que o histograma incoerente tem sempre um poder descritivo maior, confirmando a hipótese de que os pixeis incoerentes, em qualquer limiar, descrevem melhor a imagem tanto para o LBP quando para o MCT8 Contextual.

O vetor descritivo final desta dissertação será a concatenação dos histogramas coerentes da imagem gerada pelo LBP e da imagem gerada pelas duas operações seguidas do MCT8 (MCT8 Contextual), formando o CMLBP incoerente.

O valor escolhido para o parâmetro de limiar de região em ambos os casos foi de 0,005%, no qual obteve as melhores respostas. Para os próximos resultados com o CMLBP Incoerente serão utilizados três níveis diferentes de discretização: 256, 128 e 64 *buckets*.

Figura 38 – Distribuição dos pixeis coerentes para diferentes valores do limiar de região.



Fonte: Próprio autor

3.3 Support Vector Machine

Support Vector Machine SVM é um classificador supervisionado, cujo objetivo consiste em achar um hiperplano ótimo capaz de separar duas classes com a menor taxa de erro esperada possível (BURGES, 1998). Assim, caso haja um conjunto de dados não linearmente separáveis, através de uma função de transformação não linear, denominada *kernel*, mapeia-se o espaço de entradas em um espaço de alta ordem, porém linearmente separável. Se trata de uma técnica capaz de descrever resultados precisos mesmo com uma base de dados de treinamento pequena (MANDHALA; SUJATHA; DEVI, 2014).

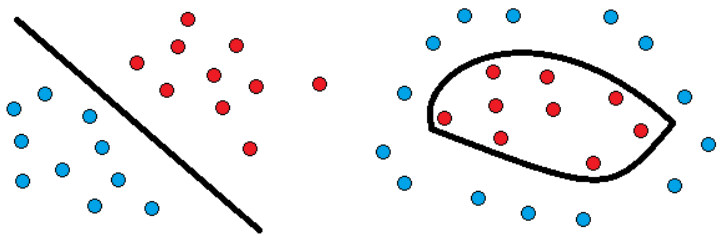
Um conjunto de dados linearmente separáveis, ou mutualmente exclusivos, é ilustrado na imagem da esquerda da Figura 39 , onde há dois tipos de classes com cores diferentes. Neste exemplo, a linha reta que separa as duas classes funciona como um classificador. Assim, valores acima desta linha pertencerão à uma classe e valores abaixo pertencerão à outra classe. Nesta mesma figura há um exemplo de dados não-linearmente separáveis, na imagem da direita. Observe que é inviável a escolha de uma reta para separar as duas classes, necessitando de um classificador mais robusto.

Como as cenas possuem pequena variação interclasse, a escolha de um classificador capaz de trabalhar com dados não-linearmente separáveis é fundamental para a obtenção

de bons resultados.

Figura 39 – Dados linearmente e não-linearmente separáveis

(a) linearmente separáveis. (b) não-linearmente separáveis



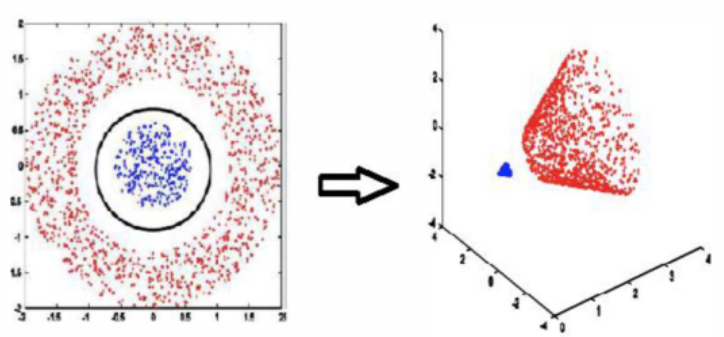
Fonte: Próprio autor

A Figura 40 exemplifica uma operação de *kernel* realizada pelo SVM para os dados não-linearmente separáveis e dispostos em um espaço bidimensional. Após a aplicação de uma função de transformação para uma dimensão maior, no caso um espaço tridimensional, os dados se tornam linearmente separáveis, tornando a classificação mais fácil. A nova dimensão na qual os dados estarão dispostos depende da distribuição das amostras de treinamento, não tendo um valor fixo ou recomendado (MANDHALA; SUJATHA; DEVI, 2014).

O método define também uma região de segurança, conhecida como margem, em torno do hiperplano de separação. A definição da margem garante uma distância mínima entre o hiperplano de separação e as amostras ,conhecidas como vetores de suportes, proporcionando uma melhor discriminação de classes.

Na literatura, observa-se que o SVM é constantemente usado como classificador de cenas, independente da complexidade do algoritmo ou das ferramentas utilizadas para a descrição da imagem, se tornando um dos principais métodos nesta área. Alguns pesquisadores inclusive, acreditam ser o melhor classificador de aprendizado supervisionado para esse tipo de classificação (MANDHALA; SUJATHA; DEVI, 2014). A principal razão

Figura 40 – Transformação de dados para um espaço dimensional de maior dimensão.



Fonte: Mandhala, Sujatha e Devi (2014)

se deve ao fato de que o SVM é robusto contra ruídos, requer um baixo custo computacional e disponibiliza um bom aproveitamento mesmo com as amostras em um espaço dimensional elevado.

Em (RAJA; ROOMI; DHARMALAKSHMI, 2013) foi proposto um classificador para cenas externas utilizando características de baixo nível como cor e textura. As informações de cor foram obtidas através de um modelo descritivo obtido por operações no modelo RGB e as informações de textura através do filtro de Gabor, LBP e PCA. Nesse trabalho, o SVM apresentou resultados melhores que o k-nn em todos os experimentos.

Em (MANDHALA; SUJATHA; DEVI, 2014) realizou-se um estudo com SVM e constatou-se que seu emprego pode simplificar bastante uma classificação de alto nível como a rotulação de cenas. Neste trabalho, as imagens foram classificadas em quatro tipos de cena: Costa, Floresta, Estrada e Rua. Mandhala, Sujatha e Devi (2014) também apresentaram um estudo a respeito do melhor *kernel* a ser utilizado e o filtro gaussiano obteve os melhores resultados em comparação aos filtros linear e polinomial.

Serrano, Savakis e Luo (2002) classificaram um banco de dados de 1200 imagens em duas classes de cenas: Internas ou Externas. Foi utilizado informação de cor através de uma transformação do modelo RGB e informação de textura adquirida pela transformada wavelet. A etapa de classificação consistia em dois estágios e em ambos foi empregado o SVM. A principal motivação para a escolha deste classificador é que o SVM tem se mostrado mais competitivo, obtendo melhores resultados e taxa de erros menores comparado à outros métodos como o k-nn.

A Tabela 5 comprova a superioridade do SVM em relação ao k-nn para a classificação de cenas. O descritor empregado foi o LBP Incoerente-64 e o para o *kernel* adotou-se a função gaussiana. O método de validação é o *10-fold-cross-validation*, o mesmo dos testes anteriores para o LBP-Incoerente. É possível perceber que para todos os banco de dados o SVM obteve os melhores resultados. Portanto, nos próximos testes a serem realizados tanto pelo LBP Incoerente quanto para o CMLBP Incoerente será utilizado este classificador.

Tabela 5 – Resultados adquiridos utilizando o k-nn e o SVM para o LBP Incoerente-64.

Banco de dados	k-nn	SVM
5 cenas	70,3% ± 4,1	**78,2% ± 4,4**
8 cenas	68,3% ± 2,8	**78,8% ± 1,4**
13 cenas	63,6% ± 2,6	**75,6% ± 2,4**
15 cenas	58,3% ± 2,2	**72,1% ± 2,2**

3.4 Resultados obtidos

Nesta seção serão apresentados os resultados obtidos para o CMLBP Incoerente e compará-los com o LBP Incoerente e técnicas conhecidas na literatura. Um estudo particular também será realizado em relação ao CMLBP Incoerente e o CMCT, para comprovar se de fato a adição de informações espaciais através do descarte das regiões coerentes pode aumentar o poder descritivo da imagem, assim como feito anteriormente com a comparação entre LBP Incoerente e LBP.

Os resultados apresentados a seguir e todo o processamento das imagens foram realizados no ambiente de desenvolvimento do Matlab®. Para a utilização do classificador SVM foi usado o pacote LIBSVM (CHANG; LIN, 2011). Para o parâmetro C, que controla os conflitos entre os erros do SVM durante o treinamento e maximização da margem, empregou-se, após testes exaustivos, C=13. Utilizou-se como operador *kernel*, a função gaussiana com fator de gamma igual a 3, também após vários testes nos diferentes bancos de dados. A estratégia adotada neste trabalho, por se tratar de um classificador multi-classes, foi a técnica *one-against-all*.

3.4.1 Banco de Dados

Além dos banco de dados de 8 classes de cenas em (OLIVA; TORRALBA, 2001) e 15 classes de cenas em (LAZEBNIK; SCHMID; PONCE, 2006), empregou-se mais dois banco de dados conhecidos na literatura.

- 8 classes de eventos de esportes - publicado em (LI; FEI-FEI, 2007), este banco de dados contém 1579 imagens de oito classes de esportes: Badminton (200 imagens), Bocha (142 imagens), Críquete (245 imagens), Polo (184 imagens), Escalada (215 imagens), Remo (254 imagens), Navegação à vela (201 imagens) e Snowboard (217 imagens). A Figura 41 mostra alguns desses eventos, onde há 3 imagens para cada classe de esporte. O número de imagens para cada categoria varia de 137 a 250 e o tamanho da imagem também possui uma grande variação. Este banco de dados tem uma característica importante de generalização de escala. O autor capturou, para uma mesma classe, várias imagens de escalas diferentes como podemos ver na Figura 41 em que tem-se três exemplos para cada classe.

- 67 classes de cenas internas - publicado por Quatonni e Torralba (QUATTONI; TORRALBA, 2009), este banco de dados contém 15616 imagens divididas em 67 classes de diferentes ambientes internos, presentes no cotidiano tais como Igreja, Livraria e Padaria. O número de imagens por classe pode variar de 120 à 790. A resolução também não é fixa, com uma extensa faixa para imagens em uma mesma classe. Este banco de dados possui uma extensa variabilidade intraclasse. Devido a

Figura 41 – Imagens de 8 classes de eventos de esportes.

Fonte: Li e Fei-Fei (2007)

esses fatores, é considerado um problema desafiador no campo do reconhecimento de cenas (QUATTONI; TORRALBA, 2009).

Apesar de haver imagens coloridas em alguns bancos de dados, todas as imagens foram convertidas para nível de cinza. Da mesma forma que o LBP Incoerente, para a construção do vetor descritivo referente ao CMLBP Incoerente descartou-se os valores 0 e 255. A normalização foi feita de forma que o vetor tenha média 0 e norma euclidiana 1, assim como é feito em (WU; REHG, 2011). Estas operações melhoraram os resultados finais para todos os bancos de dados nos experimentos.

Tabela 6 – Informações a respeito dos bancos de dados.

Base de dados	Total de Imagens	Imagens por classe	Tamanho máximo
8 classes cenas	2688	260 a 410	122 Kb
15 classes cenas	4485	215 a 410	122 Kb
8 classes eventos de esporte	1579	142 a 254	3,27 Mb
67 classe de cenas internas	15620	101 a 734	7,17 Mb

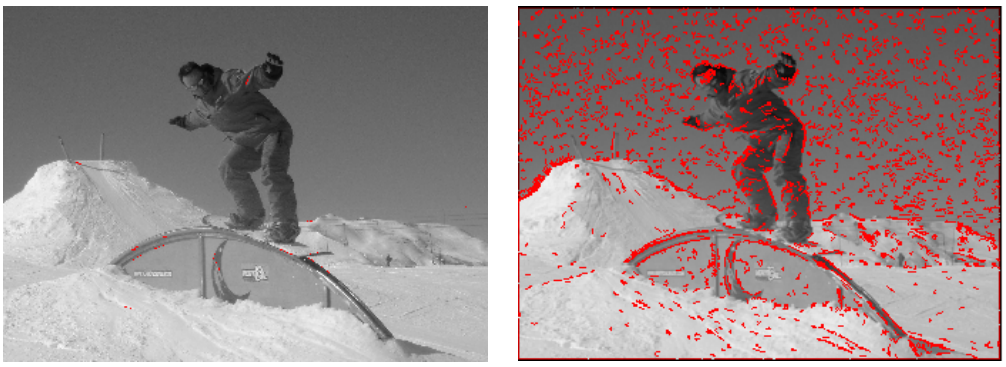
A Tabela 6 resume a descrição dos quatro banco de dados utilizados para este trabalho. Note que os dois últimos banco de dados apresentados possuem imagens com resoluções variadas. Este fator é prejudicial à extração de pixeis incoerentes visto que uma

imagem com grande resolução sugere ter um comportamento diferente da imagem com baixa resolução na formação das regiões coerentes, pois a definição do tamanho mínimo para a formação da região coerente, ou seja o limiar de região, está em função do tamanho da imagem.

A Figura 42 exemplifica melhor esta situação quando utiliza-se o LBP Incoerente para diferentes resoluções. Observe que para a imagem com alta resolução, mesmo tendo uma grande faixa de região homogênea, há poucos pixeis coerentes detectados. Isto acontece porque 0,005% da resolução desta imagem corresponde a 125 pixeis, um número elevado de pixeis com mesmo valor agrupados em uma imagem gerada pelo LBP. Ao passo que, com a resolução redimensionada para um número menor, obtemos mais pixeis coerentes, aumentando o poder descritivo de nosso algoritmo. Nesta mesma figura com a nova resolução, o limiar de região passa a ser de 4 pixeis. Observe que as regiões homogêneas são preenchidas em maior parte pelos pixeis coerentes após o redimensionamento.

Figura 42 – Influência da resolução na detecção de pontos coerentes para o LBP Incoerente.

(a) Resolução original (1852 x 1532) (b) Resolução redimensionada (329x240)



Fonte: Próprio autor

A fim de alcançar uma padronização das imagens para que o algoritmo de detecção de regiões coerentes não seja afetado, utilizou-se a função *imresize* do Matlab para que todas as imagens possuíssem aproximadamente o mesmo tamanho. Esta função utiliza a interpolação bicúbica como método padrão. As novas dimensões de uma imagem com tamanho [a x b] serão obtidas através da seguinte equação:

$$\begin{aligned} a_n &= w \\ b_n &= \frac{b.w}{a}, \end{aligned} \tag{3.2}$$

em que [a_n x b_n] é a nova resolução da imagem e w o valor da largura a ser escolhido. Neste trabalho, verificou-se que w igual 240 obteve os melhores resultados, além de estar próximo do valor de dimensão do banco de dados de 8 cenas de (OLIVA; TORRALBA, 2001).

3.4.2 8 classes de cenas

Como mencionado nos resultados da seção anterior, o referencial de desempenho é a acurácia, já explicada na Equação 2.6. Entretanto, utilizou-se o método de validação realizado em (OLIVA; TORRALBA, 2001) a fim de comparar os resultados com outras técnicas para esse banco de dados nas quais testaram com este mesmo método de validação.

Inicialmente, 100 imagens de cada classe foram capturadas aleatoriamente para compor a base de treinamento. Assim, as imagens restantes em cada classe irão preencher a base de testes. Esse processo é repetido cinco vezes, com a base de treinamento e de testes sempre sendo construída novamente. Com cinco medidas de acurácia, calcula-se a média final e o desvio padrão.

A Tabela 7 apresenta os resultados obtidos em que o CMLBP Incoerente-256 obteve 80,70% ± 1,0 de aproveitamento, enquanto que o LBP Incoerente-256 obteve um reconhecimento de 77.92% ± 0,4. Portanto, a informação de contexto proporcionou melhores resultados ao algoritmo. O *gist* foi o descritor com melhor resposta, visto que este algoritmo é mais apropriado para cenas com pouca quantidade de objetos e poucas direções de bordas, como o *layout* das imagens deste banco de dados.

O CMLBP Incoerente-256 obteve melhores resultados em comparação ao CMCT, além disso, o LBP Incoerente com 256 e 128 *buckets* obteve melhores resultados que o CENTRIST, ou seja, o acréscimo da informação a respeito do espaço em que o pixel está inserido através dos pixeis incoerentes proporcionou melhores resultados. É importante ressaltar que o LBP Incoerente-128 tem sua dimensão duas vezes menor que o CENTRIST.

Tabela 7 – Resultados com o banco de dados de 8 classes de cenas.

Descritor de características	Acurácia
gist (OLIVA; TORRALBA, 2001)	**83,70% ± 0,7**
CMLBPIncoherent-256	80,70% ± 1,0
CMCT (GAZOLLI; SALLES, 2014)	79,91% ± 1,0
CMLBPIncoherent-128	78,51% ± 1,2
LBPIncoherent-256	77,92% ± 0,4
MCT8 (GAZOLLI; SALLES, 2014)	77,07% ± 0,7
LBPIncoherent-128	76,86% ± 0,4
CENTRIST (WU; REHG, 2011)	76,49% ± 0,8

A Figura 43 apresenta a matriz de confusão dos resultados para esse banco de dados em que só foram considerados os resultados iguais ou acima de 9% para melhor visualização.

Analisando a matriz, observa-se que os piores resultados que o CMLBP Incoerente-256 apresentou foram para as classes Zona Aberta e Prédio. As confusões entre as classes

Zona Aberta, Costa e Montanha foram em virtude do seu *layout* parecido em comparação com as outras classes. Nestas três classes há uma maior presença de regiões homogêneas e bordas horizontais. Já para as classes de Prédio, Rua e Cidade há uma presença maior de estruturas artificiais como prédios e casas. O aparecimento destes objetos proporciona uma maior quantidade de bordas verticais. A Figura 44 mostra exemplos destas classes retiradas no banco de dados.

Figura 43 – Matriz de confusão para a base de dados de 8 cenas em (OLIVA; TORRALBA, 2001). Apenas resultados acima ou iguais a 9% são mostrados.

	Costa	Floresta	Estrada	Cidade	Montanha	Zona Aberta	Rua	Prédio
Costa	84%							
Floresta		91%						
Estrada			85%					
Cidade				82%			11%	
Montanha					79%	10%		
Zona Aberta	10%				9%	72%		
Rua				10%			83%	
Prédio				9%			10%	76%

Fonte: Próprio autor

Figura 44 – Semelhança entre classes.

(a) "Montanha", "Zona Aberta"e "Costa"

(b) "Rua", "Prédio"e "Cidade"

Fonte: Oliva e Torralba (2001)

Um outro ponto importante a ser observado é o maior aproveitamento para a classe Floresta em relação às demais classes. Esta classe possui algumas peculiaridades como o seu *layout* diferente das outras classes desse banco de dados. As cenas de Floresta possuem bordas com várias orientações.

3.4.3 15 classes de cenas

Para o banco de dados de 15 classes de cenas os resultados saíram piores em comparação com as técnicas de CENTRIST e CMCT. O método de validação é o mesmo usado para o banco de dados de 8 classes de cenas. O CMLBP Incoerente-256 alcançou um resultado de 76,35% ± 0,9, enquanto que para o CMCT o aproveitamento foi de 76,87% ± 0,6. Já o LBP Incoerente-256 teve um percentual de 72,16% ± 0,4 enquanto que o CENTRIST obteve 73,29% ± 1,0.

Em (LI; FEI-FEI, 2007) é proposto um método para reconhecimento de cenas integrando informação contextual com a abordagem *bag-of-words*, o *Region of Contextual of Visual Words* (RCVW). Este método é superado pelo CMLBP Incoerente, assim como o *gist*. Esta queda de aproveitamento do *gist* já era esperada em virtude do fato de que este banco de dados contém mais cenas internas.

Acredita-se que a principal razão para os resultados do LBP Incoerente e CMLBP Incoerente serem inferiores ao CENTRIST e CMCT respectivamente é devido à taxa de aproveitamento da classe Industrial. Enquanto que para o CMLBP Incoerente-256 o reconhecimento nesta classe foi de 56%, para o CMCT foi de 69%.

Tabela 8 – Resultados com o banco de dados de 15 classes de cenas de (LAZEBNIK; SCHMID; PONCE, 2006).

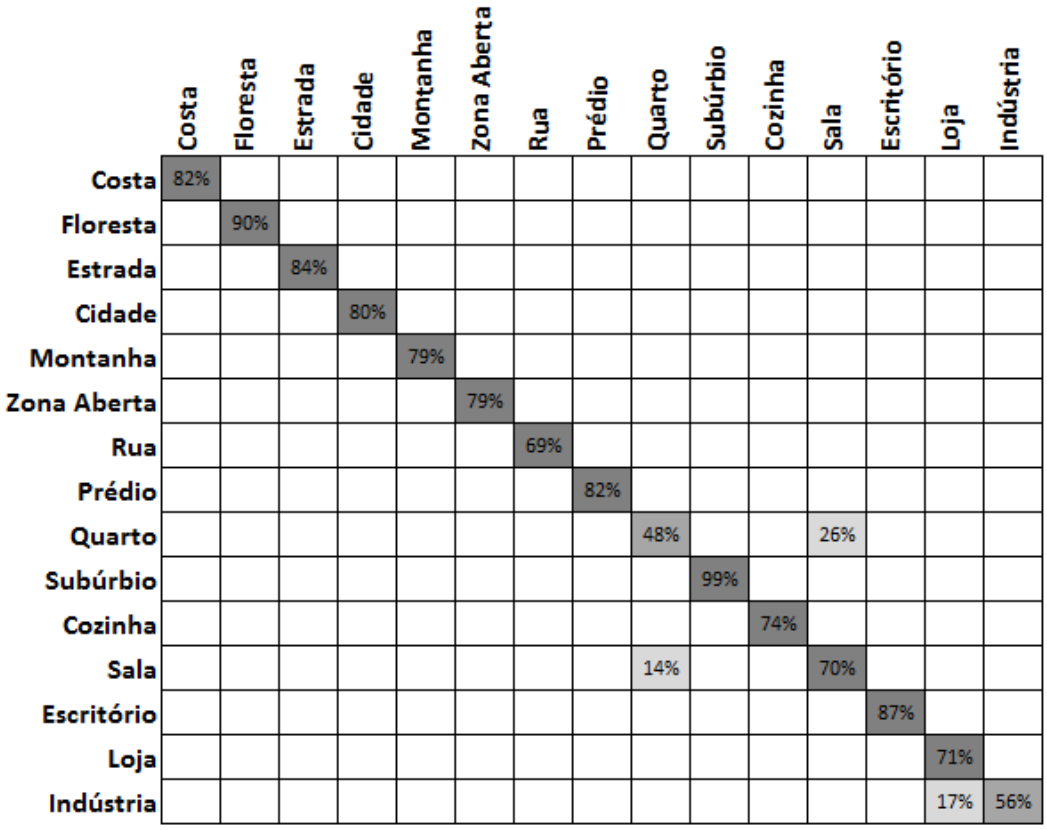
Feature descriptor	Accuracy (%)
CMCT (GAZOLLI; SALLES, 2014)	**76,87% ± 0,6**
CMLBPIncoherent-256	76,35% ± 0,9
CMLBPIncoherent-128	74,93% ± 0,5
RCVW(LI; FEI-FEI, 2007)	74,5%
MCT8(GAZOLLI; SALLES, 2014)	73,71% ± 0,8
CENTRIST (WU; REHG, 2011)	73,29% ± 1,0
gist(OLIVA; TORRALBA, 2001)	73,28% ± 0,7
LBPIncoherent-256	72,16% ± 0,4
LBPIncoherent-128	71,52% ± 0,9

A matriz de confusão apresentada na Figura 45 expõe esse baixo desempenho na classe Industrial devido a presença da classe Loja. Mas a principal confusão neste banco de dados está entre as classes Quarto e Sala. A justificativa é que estas duas classes

possuem objetos parecidos, como os móveis, além da distribuição das janelas. Até para o seres humanos o reconhecimento rápido entre essas duas classes pode gerar um pouco de confusão. A Figura 46 ilustra um exemplo de cada classe.

O alto desempenho da classe Subúrbio também deve ser destacado, pois obteve uma taxa de acurácia de quase 100%. Por se tratar de uma classe com pouca variação em suas imagens, o CMLBP Incoerente -256 foi eficiente em reconhecer esse tipo de cena. A Figura 47 ilustra alguns exemplos de imagens de Subúrbio retiradas do banco de dados de (LAZEBNIK; SCHMID; PONCE, 2006). Note a semelhança de conteúdo entre cada imagem e a pouca variação intraclasse que elas apresentam.

Figura 45 – Matriz de confusão para o banco de dados de 15 cenas de (LAZEBNIK; SCHMID; PONCE, 2006) com o CMLBP-Incoerente 256. Apenas taxas acima de 10% foram consideradas.



Fonte: Próprio autor

3.4.4 8 eventos de esportes

A Tabela 9 compara o desempenho dos métodos propostos para o banco de dados de 8 eventos de esportes. O método de validação utilizado foi o mesmo de (LI; FEI-FEI, 2007) em que captura-se 70 imagens para treino e 60 para testes em cada classe, realizando esse procedimento cinco vezes. Ao final calcula-se a acurácia média e o desvio padrão. O CMLBP Incoerente-256 obteve aproveitamento de 72,38% ± 1,1 e o LBP Incoerente-256

Figura 46 – Semelhança entre as classes "Quarto"e "Sala".

(a) Quarto. (b) Sala.

Fonte: Lazebnik, Schmid e Ponce (2006)

Figura 47 – Imagens de "Subúrbio".

Fonte: Lazebnik, Schmid e Ponce (2006)

obteve 69,06% ± 1,7. Em (LI; FEI-FEI, 2007), é proposto um método de classificação de eventos resultante da junção de duas abordagens, uma envolvendo a cena como um todo e outra envolvendo classificação de objetos. A acurácia para este método foi de 73,4% e superou todos os outros algoritmos. Todavia, é importante ressaltar que o custo computacional desta técnica é maior que o CMLBP Incoerente, caracterizado por seu baixo custo computacional, pouca memória utilizada e uso de técnicas não-paramétricas.

Em relação aos algoritmos que utilizam transformada não-paramétricas, os métodos aqui propostos tiveram desempenho melhor. O CMLBP Incoerente e LBP Incoerente, com 256 e 128 *buckets*, superaram o CMCT e Centrist. Vale ressaltar que o LBP Incoerente-128 possui o tamanho do seu vetor quatro vezes menor que o CMCT e mesmo assim atingiu resultados melhores com este banco de dados.

3.4.5 67 classes de cenas internas

Para o banco de dados com 67 classes de cenas internas utilizamos o método de validação empregado em (QUATTONI; TORRALBA, 2009) em que captura-se 80 imagens

Tabela 9 – Resultados com o banco de dados de 8 eventos de esportes (LI; FEI-FEI, 2007).

Feature descriptor	Accuracy (%)
Local + global (LI; FEI-FEI, 2007)	**73,4%**
CMLBPIncoherent-256	72,38% ± 1,1
CMLBPIncoherent-128	70,57% ± 0,9
LBPIncoherent-256	69,06% ± 1,7
LBPIncoherent-128	68,77% ± 1,5
CMCT (GAZOLLI; SALLES, 2014)	67,41% ± 1,1
Centrist (WU; REHG, 2011)	63,91% ± 2,4

para treino e 20 imagens para teste em cada classe, realizando esse procedimento cinco vezes. Por fim, calcula-se a acurácia média e o desvio padrão. O CMLBP Incoerente-256 obteve os melhores resultados alcançando 27,70% ± 0,6 de reconhecimento enquanto que o CMLBP Incoerente-128 teve um aproveitamento de 26,64% ± 1,9. O LBP Incoerente-256 com 23,04% ± 1,3 de aproveitamento também superou o CENTRIST que obteve 22,46% ± 0,8.

Para esse experimento o *gist* obteve os piores resultados, com um aproveitamento de 21%, o que já era esperado, visto que este banco de dados é composto por cenas com grande quantidade de informação e detalhes. Outrossim, o CMLBP Incoerente superou o algoritmo proposto em (LI; FEI-FEI, 2007), mesmo com custo computacional menor.

Tabela 10 – Resultados com o banco de dados de 67 classes de cenas internas (QUATTONI; TORRALBA, 2009).

Feature descriptor	Accuracy (%)
CMLBPIncoherent-256	**27,70% ± 0,6**
CMLBPIncoherent-128	26,64% ± 1,9
CMCT (GAZOLLI; SALLES, 2014)	25,82% ± 0,7
Local + global (LI; FEI-FEI, 2007)	25%
LBPIncoherent-256	23,04% ± 1,2
CENTRIST (WU; REHG, 2011)	22,46% ± 0,8
gist (OLIVA; TORRALBA, 2001)	21%
LBPIncoherent-128	22,27% ± 0,7

Os resultados apresentados para as cenas internas sugerem que a técnica de separação dos pixeis incoerentes proporciona melhores resultados para esse tipo de cena em comparação com outra técnicas conhecidas na literatura. A Figura 48 ilustra alguns tipos de cenas que pertencem à esse banco de dados, observe que as imagens possuem como características em comum uma grande quantidade de objetos, dificultando uma percepção

a respeito da cena de forma rápida.

Figura 48 – Algumas imagens de classes do banco de dados de 67 classes internas. Da esquerda para direita, de cima para baixo tem-se:: "Padaria", "Restaurante", "Laboratório"e "Igreja". Na linha de baixo temos: "Cassino", "Hospital", "Metrô"e "Aeroporto".

Fonte: Quattoni e Torralba (2009)

4 Conclusões e Projetos Futuros

Neste trabalho foi proposto o *Contextual Modified Local Binary Pattern* Incoerente (CMLBP Incoerente), um descritor visual que captura propriedades estruturais da imagem através do operador *Local Binary Pattern* e de seu histograma. Além disso, o CMLBP Incoerente combina informação contextual destas estruturas locais e descarta as regiões homogêneas detectadas pelos pixeis coerentes, através do algoritmo de separação de regiões coerentes e incoerentes proposto em (PASS; ZABIH; MILLER, 1997).

Assim como o LBP, este método possui algumas vantagens como custo computacional baixo e poucas variáveis para serem estimadas como o limiar de região e o *gamma* para o *kernel* gaussiano do SVM. Tal fato torna atrativa a utilização do CMLBP Incoerente para usuários leigos, sem muito conhecimento técnico a respeito da classificação de cenas.

A informação incoerente tem seu poder descritivo maior que a informação coerente na imagem. Quando concatena-se o histograma das duas informações, ainda assim, o histograma dos pixeis incoerentes possui um resultado melhor. Portanto, conclui-se que para o LBP as regiões homogêneas formadas pelos pixeis coerentes podem prejudicar o reconhecimento de cenas, sendo uma boa alternativa não considerá-las na construção do histograma final da imagem.

O valor do parâmetro para decidir o tamanho limite de uma região, a ponto dela ser classificada como coerente ou incoerente, foi de 0,005% da imagem, pois este valor alcançou os melhores resultados após testes. Assim, regiões com pixeis de mesma intensidade com tamanho maior que este limiar teriam todos os seus pixeis classificados como coerente e o contrário como incoerente. Para melhor padronização desse limiar em função do tamanho da imagem, redimensionou-se as cenas para que as imagens tenham uma dimensão aproximada.

Observando os resultados apresentados nas tabelas 7, 8, 9 e 10 pode-se concluir que o CMLBP Incoerente alcançou melhores resultados que o CMCT e que o LBP Incoerente alcançou melhores resultados que o Centrist, na maioria dos casos. A exceção ficou por conta do banco de dados de 15 cenas (LAZEBNIK; SCHMID; PONCE, 2006). Acredita-se que o acréscimo das classes Loja e Industrial prejudicou as técnicas de detecção de pixeis coerentes.

Em alguns casos, mesmo com um vetor descritivo com tamanho quatro vezes menor, o LBP Incoerente atingiu melhores resultados que o CMCT, como mostrado na Tabela 9. Estes testes comprovam que o acréscimo da informação espacial, representado pela separação de pixeis incoerentes, em algoritmos que utilizam transformadas não-paramétricas, podem aumentar o poder descritivo de seu histograma final.

CMLBP Incoerente se mostrou robusto visto que obteve bons resultados tanto para cenas internas quanto para cenas externas, ao contrário de outros algoritmos conhecidos na literatura em que sua abordagem é específica para um tipo de *layout* de cena, como o *gist* e a técnica de (LI; FEI-FEI, 2007). Em alguns momentos o CMLBP foi superado por esses algoritmos, entretanto, quando os mesmos eram testados em outros tipos de banco de dados, os resultados não foram satisfatórios, ao contrário da acurácia alcançada pela presente técnica proposta neste trabalho.

Comparando os resultados entre o CMLBP Incoerente e LBP Incoerente é possível constatar que o acréscimo da informação contextual proporciona uma melhora no poder descritivo do histograma final e na representação da imagem. Esse acréscimo não teve como consequência um grande custo computacional. Outrossim, o CMLBP Incoerente-128, mesmo tendo o mesmo tamanho do LBP Incoerente-256 se saiu melhor em todos os resultados.

Como propostas para trabalhos futuros é possível citar:

- O emprego das técnicas de separação entre pixeis coerentes e incoerentes aliados à informação de cor para constatar se esta informação pode ajudar na representação da imagem pelos pixeis incoerentes.
- Utilização de técnicas de separação da imagem em *grids* para agregar informação local na construção do histograma de pixeis incoerentes da imagem.
- Expandir a informação de contexto capturando pixeis mais distantes do pixel central a fim de comprovar se estes pixeis influenciam em uma melhor descrição da imagem final.
- Experimentar a aplicação de outros classificadores com métodos probabilísticos com o objetivo de melhorar os resultados.
- Incorporar ao CMLBP Incoerente o método de segmentação JSEG (DENG; MANJUNATH, 2001) para a detecção de regiões segmentadas em que os pixeis coerentes possuem maioria. A motivação é pelo fato de que em algumas regiões homogêneas há a presença de pixeis incoerentes indesejáveis para a construção do histograma final
- Utilização de algoritmos para redução do tamanho do vetor como PCA e o método de Fisher, este último mais apropriado para situações com grande variação intraclasse e pequena variação interclasse como ocorre nos bancos de dados empregados.
- Combinação do método de construção do LBP Incoerente com o método de LBP invariante à rotação proposto em (OJALA; PIETIKAINEN; MAENPAA, 2002).

Há ainda algumas implementações mais específicas a serem feitas após um estudo bibliográfico nessa área como a utilização de outras métricas de distância como Interseção de histograma, Chi quadrado e Mahalanobis, substituindo a distância euclidiana.

Referências

AHONEN, T.; HADID, A.; PIETIKÄINEN, M. Face recognition with local binary patterns. *Computer vision-eccv 2004*, Springer, p. 469–481, 2004. Citado na página 37.

AKBAS, E.; AHUJA, N. Low-level image segmentation based scene classification. In: IEEE. *Pattern Recognition (ICPR), 2010 20th International Conference on*. [S.l.], 2010. p. 3623–3626. Citado na página 22.

BHARATHI, P.; REDDY, K. R.; SRILAKSHMI, G. Medical image retrieval based on lbp histogram fourier features and knn classifier. In: IEEE. *Advances in Engineering and Technology Research (ICAETR), 2014 International Conference on*. [S.l.], 2014. p. 1–4. Citado na página 36.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. *Journal of machine Learning research*, v. 3, n. Jan, p. 993–1022, 2003. Citado na página 26.

BURGES, C. J. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, Springer, v. 2, n. 2, p. 121–167, 1998. Citado na página 64.

CECCHI, G. A.; RAO, A. R.; XIAO, Y.; KAPLAN, E. Statistics of natural scenes and cortical color processing. *Journal of vision*, The Association for Research in Vision and Ophthalmology, v. 10, n. 11, p. 21–21, 2010. Citado na página 22.

CHANG, C.-C.; LIN, C.-J. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, ACM, v. 2, n. 3, p. 27, 2011. Citado na página 67.

CHEN, Y.; LI, J.; WANG, J. Z. *Machine learning and statistical modeling approaches to image retrieval*. [S.l.]: Springer Science & Business Media, 2006. v. 14. Citado 2 vezes nas páginas 23 e 24.

CHU, J.; ZHAO, G.-H. Scene classification based on sift combined with gist. In: IEEE. *Information Science, Electronics and Electrical Engineering (ISEEE), 2014 International Conference on*. [S.l.], 2014. v. 1, p. 331–336. Citado 2 vezes nas páginas 17 e 28.

COVER, T.; HART, P. Nearest neighbor pattern classification. *IEEE transactions on information theory*, IEEE, v. 13, n. 1, p. 21–27, 1967. Citado na página 48.

DAN, Z.; SANG, N.; HE, Y.; SUN, S. An improved lbp transfer learning for remote sensing object recognition. *Optik-International Journal for Light and Electron Optics*, Elsevier, v. 125, n. 1, p. 482–485, 2014. Citado na página 33.

DAS, S.; JENA, U. R. Texture classification using combination of lbp and glrlm features along with knn and multiclass svm classification. In: *2016 2nd International Conference on Communication Control and Intelligent Systems (CCIS)*. [S.l.: s.n.], 2016. p. 115–119. Citado na página 33.

DENG, Y.; MANJUNATH, B. Unsupervised segmentation of color-texture regions in images and video. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 23, n. 8, p. 800–810, 2001. Citado 2 vezes nas páginas 25 e 78.

DUDA, R. O.; HART, P. E. Use of the hough transformation to detect lines and curves in pictures. *Communications of the ACM*, ACM, v. 15, n. 1, p. 11–15, 1972. Citado na página 22.

FAN, H.; MEI, X.; PROKHOROV, D.; LING, H. Multi-level contextual rnns with attention model for scene labeling. *arXiv preprint arXiv:1607.02537*, 2016. Citado na página 28.

FEI-FEI, L.; PERONA, P. A bayesian hierarchical model for learning natural scene categories. In: IEEE. *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. [S.l.], 2005. v. 2, p. 524–531. Citado 5 vezes nas páginas 10, 26, 45, 51 e 52.

FROBA, B.; ERNST, A. Face detection with the modified census transform. In: IEEE. *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*. [S.l.], 2004. p. 91–96. Citado 2 vezes nas páginas 55 e 56.

GAZOLLI, K. de S.; SALLES, E. O. T. Exploring neighborhood and spatial information for improving scene classification. *Pattern Recognit. Lett.*, v. 46, p. 83–88, 2014. Citado 8 vezes nas páginas 31, 56, 57, 59, 60, 70, 72 e 75.

GONZALEZ, R. C.; WOODS, R. E. *Digital Image Processing (3rd Edition)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2006. ISBN 013168728X. Citado na página 50.

GORKANI, M. M.; PICARD, R. W. Texture orientation for sorting photos"at a glance". In: IEEE. *Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision & Image Processing., Proceedings of the 12th IAPR International Conference on*. [S.l.], 1994. v. 1, p. 459–464. Citado 3 vezes nas páginas 21, 22 e 23.

HARALICK, R. M.; SHANMUGAM, K. et al. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, Ieee, v. 3, n. 6, p. 610–621, 1973. Citado na página 24.

HE, D.-C.; WANG, L. Texture unit, texture spectrum, and texture analysis. *IEEE transactions on Geoscience and Remote Sensing*, IEEE, v. 28, n. 4, p. 509–512, 1990. Citado na página 34.

HOLLINGWORTH, A.; HENDERSON, J. M. Object identification is isolated from scene semantic constraint: Evidence from object type and token discrimination. *Acta psychologica*, Elsevier, v. 102, n. 2, p. 319–343, 1999. Citado na página 17.

JIANG, A.; WANG, C.; XIAO, B.; DAI, R. A new biologically inspired feature for scene image classification. In: IEEE. *Pattern Recognition (ICPR), 2010 20th International Conference on*. [S.l.], 2010. p. 758–761. Citado na página 19.

KHASTAVANEH, H.; EBRAHIMPOUR-KOMLEH, H.; HANAEE-AHWAZ, A. Unknown aware k nearest neighbor classifier. In: *2017 3rd International Conference on Pattern Recognition and Image Analysis (IPRIA)*. [S.l.: s.n.], 2017. p. 108–112. Citado na página 49.

KUNCHEVA, L. I. *Combining Pattern Classifiers: Methods and Algorithms*. 1959. Citado na página 49.

LAZEBNIK, S.; SCHMID, C.; PONCE, J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: IEEE. *Computer vision and pattern recognition, 2006 IEEE computer society conference on*. [S.l.], 2006. v. 2, p. 2169–2178. Citado 14 vezes nas páginas 9, 10, 23, 25, 52, 54, 58, 62, 63, 67, 72, 73, 74 e 77.

LEE, H. C. A physics-based color encoding model for images of natural scenes. In: IEEE. *Proceedings of the Conference on Modern Engineering and Technology, Electro-Optics Session, Taipei, Taiwan*. [S.l.], 1992. p. 25–52. Citado na página 21.

LI, L.-J.; FEI-FEI, L. What, where and who? classifying events by scene and object recognition. In: IEEE. *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. [S.l.], 2007. p. 1–8. Citado 9 vezes nas páginas 10, 29, 67, 68, 72, 73, 74, 75 e 78.

LI, R.; LI, X.; KURITA, T. Soft local binary patterns. In: IEEE. *Soft Computing and Pattern Recognition (SoCPaR), 2015 7th International Conference of*. [S.l.], 2015. p. 70–75. Citado na página 37.

LIU, K.; MOON, S. Robust dual-stage face recognition method using pca and high-dimensional-lbp. In: IEEE. *Information and Automation (ICIA), 2016 IEEE International Conference on*. [S.l.], 2016. p. 1828–1831. Citado na página 51.

LIU, S.; XU, D.; FENG, S. Region contextual visual words for scene categorization. *Expert Systems with Applications*, Elsevier, v. 38, n. 9, p. 11591–11597, 2011. Citado na página 29.

LOWE, D. G. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, Springer, v. 60, n. 2, p. 91–110, 2004. Citado na página 23.

MA, W.-Y.; MANJUNATH, B. S. Netra: A toolbox for navigating large image databases. *Multimedia systems*, Springer, v. 7, n. 3, p. 184–198, 1999. Citado na página 15.

MANDHALA, V. N.; SUJATHA, V.; DEVI, B. R. Scene classification using support vector machines. In: IEEE. *Advanced Communication Control and Computing Technologies (ICACCCT), 2014 International Conference on*. [S.l.], 2014. p. 1807–1810. Citado 4 vezes nas páginas 50, 64, 65 e 66.

NOSAKA, R.; OHKAWA, Y.; FUKUI, K. Feature extraction based on co-occurrence of adjacent local binary patterns. *Advances in image and video technology*, Springer, p. 82–91, 2012. Citado na página 38.

OHTA, Y.-I.; KANADE, T.; SAKAI, T. Color information for region segmentation. *Computer graphics and image processing*, Elsevier, v. 13, n. 3, p. 222–241, 1980. Citado na página 23.

OJALA, T.; PIETIKÄINEN, M.; HARWOOD, D. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, Elsevier, v. 29, n. 1, p. 51–59, 1996. Citado 3 vezes nas páginas 24, 33 e 44.

OJALA, T.; PIETIKAINEN, M.; MAENPAA, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, IEEE, v. 24, n. 7, p. 971–987, 2002. Citado 5 vezes nas páginas 33, 37, 38, 51 e 78.

OLIVA, A.; SCHYNS, P. G. Diagnostic colors mediate scene recognition. *Cognitive psychology*, Elsevier, v. 41, n. 2, p. 176–210, 2000. Citado na página 20.

OLIVA, A.; TORRALBA, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, Springer, v. 42, n. 3, p. 145–175, 2001. Citado 13 vezes nas páginas 9, 18, 20, 23, 28, 52, 53, 67, 69, 70, 71, 72 e 75.

OLIVA, A.; TORRALBA, A. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, Elsevier, v. 155, p. 23–36, 2006. Citado 2 vezes nas páginas 27 e 28.

PASS, G.; ZABIH, R.; MILLER, J. Comparing images using color coherence vectors. In: ACM. *Proceedings of the fourth ACM international conference on Multimedia*. [S.l.], 1997. p. 65–73. Citado 7 vezes nas páginas 31, 38, 39, 41, 44, 49 e 77.

PENTLAND, A.; PICARD, R. W.; SCLAROFF, S. Photobook: Content-based manipulation of image databases. *International journal of computer vision*, Springer, v. 18, n. 3, p. 233–254, 1996. Citado na página 15.

POTTER, M. C. Meaning in visual search. *Science*, American Association for the Advancement of Science, v. 187, n. 4180, p. 965–966, 1975. Citado 2 vezes nas páginas 20 e 21.

QUATTONI, A.; TORRALBA, A. Recognizing indoor scenes. In: IEEE. *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. [S.l.], 2009. p. 413–420. Citado 8 vezes nas páginas 10, 27, 28, 67, 68, 74, 75 e 76.

QUELHAS, P.; MONAY, F.; ODOBEZ, J.-M.; GATICA-PEREZ, D.; TUYTELAARS, T.; GOOL, L. V. Modeling scenes with local descriptors and latent aspects. In: IEEE. *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*. [S.l.], 2005. v. 1, p. 883–890. Citado na página 27.

RAJA, R.; ROOMI, S. M. M.; DHARMALAKSHMI, D. Classification and retrieval of natural scenes. In: IEEE. *Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on*. [S.l.], 2013. p. 1–8. Citado 3 vezes nas páginas 24, 48 e 66.

RAJA, R.; ROOMI, S. M. M.; KALAIYARASI, D. Semantic modeling of natural scenes by local binary pattern. In: IEEE. *Machine Vision and Image Processing (MVIP), 2012 International Conference on*. [S.l.], 2012. p. 169–172. Citado na página 24.

RENNINGER, L. W.; MALIK, J. When is scene identification just texture recognition? *Vision research*, Elsevier, v. 44, n. 19, p. 2301–2311, 2004. Citado na página 21.

SALMI, M.; BOUCHEHAM, B. Content based image retrieval based on cell color coherence vector (cell-ccv). In: IEEE. *ISKO-Maghreb: Concepts and Tools for knowledge Management (ISKO-Maghreb), 2014 4th International Symposium*. [S.l.], 2014. p. 1–5. Citado na página 41.

SCHYNS, P. G.; OLIVA, A. From blobs to boundary edges: Evidence for time-and spatial-scale-dependent scene recognition. *Psychological science*, SAGE Publications Sage CA: Los Angeles, CA, v. 5, n. 4, p. 195–200, 1994. Citado 2 vezes nas páginas 20 e 21.

SERRANO, N.; SAVAKIS, A.; LUO, A. A computationally efficient approach to indoor outdoor scene classification. In: IEEE. *Pattern Recognition, 2002. Proceedings. 16th International Conference on*. [S.l.], 2002. v. 4, p. 146–149. Citado 2 vezes nas páginas 23 e 66.

SHAHRIARI, M.; BERGEVIN, R. Can contextual information improve scene classification performance? In: IEEE. *Digital Image Computing: Techniques and Applications (DICTA), 2016 International Conference on*. [S.l.], 2016. p. 1–7. Citado na página 29.

SHIMAZAKI, K.; NAGAO, T. Scene classification using color and structure-based features. In: IEEE. *Computational Intelligence & Applications (IWCIA), 2013 IEEE Sixth International Workshop on*. [S.l.], 2013. p. 211–216. Citado na página 22.

SINGH, H.; AGRAWAL, D. An analysis based on local binary pattern (lbp) and color moment (cm) for efficient image retrieval. In: *2016 International Conference on Emerging Technological Trends (ICETT)*. [S.l.: s.n.], 2016. p. 1–6. Citado na página 33.

SMITH, J. R.; CHANG, S.-F. Visualseek: a fully automated content-based image query system. In: ACM. *Proceedings of the fourth ACM international conference on Multimedia*. [S.l.], 1997. p. 87–98. Citado na página 15.

TANG, H.; SUN, Y.; YIN, B.; GE, Y. Face recognition based on haar lbp histogram. In: IEEE. *Advanced Computer Theory and Engineering (ICACTE), 2010 3rd International Conference on*. [S.l.], 2010. v. 6, p. V6–235. Citado na página 37.

TORRALBA, A.; FERGUS, R.; FREEMAN, W. T. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 30, n. 11, p. 1958–1970, 2008. Citado 2 vezes nas páginas 29 e 30.

TORRALBA, A.; OLIVA, A. Statistics of natural image categories. *Network: computation in neural systems*, Taylor & Francis, v. 14, n. 3, p. 391–412, 2003. Citado na página 22.

TORRALBA, A.; RUSSELL, B. C.; YUEN, J. Labelme: Online image annotation and applications. *Proceedings of the IEEE*, IEEE, v. 98, n. 8, p. 1467–1484, 2010. Citado na página 29.

VAILAYA, A.; JAIN, A.; ZHANG, H. J. On image classification: City images vs. landscapes. *Pattern Recognition*, Elsevier, v. 31, n. 12, p. 1921–1935, 1998. Citado 3 vezes nas páginas 15, 41 e 44.

VOGEL, J.; SCHIELE, B. Semantic modeling of natural scenes for content-based image retrieval. *International Journal of Computer Vision*, Springer, v. 72, n. 2, p. 133–157, 2007. Citado na página 24.

WANG, A.; DAI, S.; YANG, M.; IWAHORI, Y. A novel human detection algorithm combining hog with lbp histogram fourier. In: IEEE. *Communications and Networking in China (ChinaCom), 2015 10th International Conference on*. [S.l.], 2015. p. 793–797. Citado na página 36.

WU, J.; REHG, J. M. Centrist: A visual descriptor for scene categorization. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 33, n. 8, p. 1489–1501, 2011. Citado 9 vezes nas páginas 8, 36, 38, 51, 52, 68, 70, 72 e 75.

XIAO, J.; HAYS, J.; EHINGER, K. A.; OLIVA, A.; TORRALBA, A. Sun database: Large-scale scene recognition from abbey to zoo. In: IEEE. *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*. [S.l.], 2010. p. 3485–3492. Citado na página 29.

YANG, J.; JIANG, Y.-G.; HAUPTMANN, A. G.; NGO, C.-W. Evaluating bag-of-visual-words representations in scene classification. In: ACM. *Proceedings of the international workshop on Workshop on multimedia information retrieval*. [S.l.], 2007. p. 197–206. Citado na página 26.

YIU, E. C. *Image classification using color cues and texture orientation*. Tese (Doutorado) — Massachusetts Institute of Technology, 1996. Citado na página 21.

ZABIH, R.; WOODFILL, J. Non-parametric local transforms for computing visual correspondence. In: SPRINGER. *European conference on computer vision*. [S.l.], 1994. p. 151–158. Citado na página 34.

ZHANG, B.; LIU, G.; XIE, G. Facial expression recognition using lbp and lpq based on gabor wavelet transform. In: *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*. [S.l.: s.n.], 2016. p. 365–369. Citado na página 33.

ZHANG, L.; VERMA, B.; STOCKWELL, D.; CHOWDHURY, S. Spatially constrained location prior for scene parsing. In: *2016 International Joint Conference on Neural Networks (IJCNN)*. [S.l.: s.n.], 2016. p. 1480–1486. Citado na página 21.

ZHOU, B.; LAPEDRIZA, A.; KHOSLA, A.; OLIVA, A.; TORRALBA, A. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, 2017. Citado na página 30.

ZOU, J.; LI, W.; CHEN, C.; DU, Q. Scene classification using local and global features with collaborative representation fusion. *Information Sciences*, Elsevier, v. 348, p. 209–226, 2016. Citado na página 21.