UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
CENTRO TECNOLÓGICO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

KELLY ASSIS DE SOUZA GAZOLLI

UTILIZANDO CONTEXTO NA REPRESENTAÇÃO DE IMAGENS PARA A CLASSIFICAÇÃO DE CENAS

VITÓRIA
2014

Kelly Assis de Souza Gazolli

Utilizando Contexto na Representação de Imagens para a Classificação de Cenas

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

Universidade Federal do Epírito Santo

Centro Tecnológico

Programa de Pós-graduação em Engenharia Elétrica

Orientador: Prof. Dr. Evandro Ottoni Teatini Salles

Vitória

2014

Dados Internacionais de Catalogação-na-publicação (CIP)
(Biblioteca Setorial Tecnológica,
Universidade Federal do Espírito Santo, ES, Brasil)

Gazolli, Kelly Assis de Souza, 1976-
G291u Utilizando contexto na representação de imagens para a classificação de cenas / Kelly Assis de Souza Gazolli. – 2014.
98 f. : il.

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

1. Processamento de imagens. 2. Visão por computador. 3. Sistemas de reconhecimento de padrões. 4. Classificação de cenas. 5. Transformadas não-paramétricas. I. Salles, Evandro Ottoni Teatini. II. Universidade Federal do Espírito Santo. Centro Tecnológico. III. Título.

CDU: 621.3

Kelly Assis de Souza Gazolli

Utilizando Contexto na Representação de Imagens para a Classificação de Cenas

Tese submetida ao programa de Pós-Graduação em Engenharia Elétrica do Centro Tecnológico da Universidade Federal do Espírito Santo, como requisito parcial para a obtenção do Grau de Doutor em Engenharia Elétrica.

Aprovada em 27 de junho de 2014.

COMISSÃO EXAMINADORA

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

Profa. Dra. Aura Conci
Universidade Federal Fluminense

Prof. Dr. Thomas Walter Rauber
Universidade Federal do Espírito Santo

Profa. Dra. Raquel Frizera Vassallo
Universidade Federal do Espírito Santo

Prof. Dr. Patrick Marques Ciarelli
Universidade Federal do Espírito Santo

Aos meus pais e ao meu esposo Walter

Agradecimentos

Agradeço a Deus pela oportunidade de realizar essa conquista.

Agradeço ao meu orientador Evandro Ottoni Teatini Salles pelo comprometimento, paciência e pelo incentivo nos momentos em que me abalei.

Obrigada aos meus pais, por todo amor e confiança, e ao meu esposo, Walter, que constantemente me surpreende com sua paciência, amor e apoio incondicional.

Sou grata aos colegas estudantes da pós-graduação em Engenharia Elétrica que tive a oportunidade de conhecer, cujas presenças tornaram os desafios mais aprazíveis, em especial, ao Fernando Kentaro Inaba, por compartilhar seus conhecimentos tão generosamente, e à Karin Satie Komati, pelos sábios conselhos.

Obrigada também à minha amiga Claudinete Vincente Borges por me incentivar a me inscrever na pós-graduação em Engenharia Elétrica e por estabelecer meu primeiro contato com o orientador.

Agradeço também ao Instituto Federal do Espírito Santo, em especial aos colegas da Coordenadoria de Informática do Campus Serra, pela concessão do afastamento para capacitação.

"A simplicidade é o último grau de sofisticação." Leonardo da Vinci

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ARP *Angular Radial Partitioning* - Particionamento Radial Angular

BIF *Best Individual Features* - Melhores Características Individuais

BOV *bag-of-visterms* - Bolsa de Termos Visuais

CBoW *Contextual Bag-of-Words* - Bolsa de Palavras Visuais Contextual

CBSA *Content-Based Soft Annotation* - Anotação Suave Baseada em Conteúdo

CENTRIST *Census Transform Histogram* - Histograma da Transformada Census

CMVF *Combined Multi-Visual Features* - Características Multivisuais Combinadas

CMCT *Contextual Modified Census Transform* - Transformada Census Modificada Contextual

CT *Census Transform* - Transformada Census

CTDN *Census Transform of Distant Neighbors* - Transformada Census de Vizinhos Distantes

DST *Discriminant Spectral Template* - Modelo Discriminates Espectral

ECMCT *Extended CMCT* - CMCT Estendido

ECMCT-SM *Combined ECMCT and Spatial MCT* - CMCT Estendido e MCT Espacial Combinados

GBPWHGO *Gradiente Binary Pattern Weighted Histogram of Gradient Orientation* - Padrão Binário de Gradiente e Histograma Ponderado de Orientação de Gradientes

GECMCT *Gist Extended CMCT* - GistCMCT Estendido

GECMCT-SM *Combined GECMCT and Spatial MCT* - GistCMCTEstendido e MCT Espacial Combinados

GistCMCT-SM *Combined GistCMCT and Spatial MCT* - GistCMCT e MCT Espacial combinados

HIK *Histogram Intersection kernel* - Kernel de Histograma de Interseção

HSV *Hue, Saturation and Value* - Matiz, Saturação e Valor

K-NN *K-Nearest Neighbor* - K Vizinhos Mais Próximos

LBP *Local Binary Pattern* - Padrão Binário Local

LDA *Latent Dirichlet Allocation* - Alocação Latente Dirichlet

LDBP *Local Difference Binary Pattern* - Diferença Local de Padrões Binários

MCT *Modified Census Transform* - Transformada Census Modificada

MCT8 MCT com 8 bits

MSAR *Multiresolution Simultaneous Autoregressive Model* - Modelo de Multiresolução Simultânea Autoregressiva

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

SIFT *Scale Invariant Feature Transform* - Transformada Invariante à Escala de Característica

spatial PACT *Spatial Principal Component Analysis of Census Transform* - PCA Espacial da Transformada Census

SPM *Spatial Pyramid Matching* - Correspondência Espacial em Pirâmide

SVM *Support Vector Machine* - Máquina de Vetores de Suporte

VE Vetor Estatístico

WaveLBP *Wavelet and Local Binary Pattern* - *Wavelet* e Padrão Binário Local

WDST *Windowed Discriminant Spectral Template* - Modelo Espectral Discriminate em Janela

WHGO *Weighted Histogram of Gradient Orientation* - Histograma Ponderado de Orientação de Gradientes

Resumo

A classificação de cenas é um campo bastante popular na área de visão computacional e encontra diversas aplicações tais como: organização e recuperação de imagem baseada em conteúdo, localização e navegação de robôs. No entanto, a classificação automática de cenas é uma tarefa desafiadora devido a diversos fatores, tais como, ocorrência de oclusão, sombras, reflexões e variações nas condições de iluminação e escala.

Dentre os trabalhos que objetivam solucionar o problema da classificação automática de cenas, estão aqueles que utilizam transformadas não-paramétricas e aqueles que têm obtido melhora no desempenho de classificação através da exploração da informação contextual. Desse modo, esse trabalho propõe dois descritores de imagens que associam informação contextual, ou seja, informação advinda de regiões vizinhas, a um tipo de transformada não-paramétrica. O objetivo é propor uma abordagem que não eleve demasiadamente a dimensão do vetor de características e que não utilize a técnica de representação intermediária *bag-of-features*, diminuindo, assim, o custo computacional e extinguindo a necessidade de informação de parâmetros, o que possibilita a sua utilização por usuários que não possuem conhecimento na área de reconhecimento de padrões.

Assim, são propostos os descritores CMCT (Transformada Census Modificada Contextual) e ECMCT (CMCT Estendido) e seus desempenhos são avaliados em quatro bases de dados públicas. São propostas também cinco variações destes descritores (*Gist*CMCT, GECMCT, *Gist*CMCT-SM, ECMCT-SM e GECMCT-SM), obtidas através da associação de cada um deles com outros descritores. Os resultados obtidos nas quatro bases de dados mostram que as representações propostas são competitivas, e que provocam um aumento nas taxas de classificação, quando comparados com outros descritores.

Abstract

Scene classification is a very popular topic in the field of computer vision and it has many applications, such as, content-based image organization and retrieval and robot navigation. However, scene classification is quite a challenging task, due to the occurrence of occlusion, shadows and reflections, illumination changes and scale variability.

Among the approaches to solve the scene classification problems are those that use non-parametric transform and those that improve classification results by using contextual information. Thus, this work proposes two image descriptors that associate contextual information, from neighboring regions, with a non-parametric transforms. The aim is to propose an approach that does not increase excessively the feature vector dimension and that does not use the bag-of-feature method. In this way, the proposals descrease the computational costs and eliminate the dependence parameters, which allows the use of those descriptors in applications for non-experts in the pattern recognition field.

The CMCT and ECMCT descriptors are presented and their performances are evaluated, using four public datasets. Five variations of those descriptors are also proposed (*Gist*CMCT, GECMCT, *Gist*CMCT-SM, ECMCT-SM e GECMCT-SM), obtained through their association with other approaches. The results achieved on four public datasets show that the proposed image representations are competitive and lead to an increase in the classification rates when compared to others descriptors.

Sumário

1 Introdução

Classificar uma cena consiste em associar automaticamente uma imagem a um rótulo considerando o seu conteúdo visual. Assim, neste contexto, uma classe ou categoria é definida como um grupo ou uma divisão que apresenta características semelhantes. A classificação de cenas é um campo bastante popular na área de visão computacional e encontra diversas aplicações tais como: organização e recuperação de imagem baseada em conteúdo, tratamento de imagens e navegação de robôs.

Com a popularização das câmeras digitais, as bibliotecas visuais têm crescido rapidamente, fazendo com que a organização e recuperação manual desse tipo de informação se torne mais difícil a cada dia. Desse modo, conhecer a categoria semântica de uma cena pode ser muito útil para organizar e acessar essa grande quantidade de dados tanto na Internet quanto nos computadores pessoais, *tablets* ou *smartphones*. Desse modo, quando se tem, por exemplo, o conhecimento de como se constitui uma cena de floresta, fica mais fácil encontrar as fotos "caminhada na floresta".

Métodos tradicionais de recuperação indexam imagens através da associação manual de rótulos, o que é bastante ineficiente para grandes bases de dados. Assim, se o dado visual puder ser categorizado automaticamente de acordo com o seu conteúdo a eficiência da aplicação pode ser aumentada.

O conhecimento sobre a categoria de cena pode ajudar também no tratamento de imagens (SZUMMER; PICARD, 1998). O balanceamento de cores em uma foto de pôr-do-sol, onde o excesso da cor amarela é bem-vindo, deve ser diferente do balanceamento realizado na imagem de um quarto capturada sob uma luz incandescente e sem a utilização de flash, onde a redução do aspecto amarelado melhora a foto. Dessa forma, conhecendo a categoria da cena, o algoritmo pode realizar um tratamento diferenciado para cada classe, ao invés de ajustar todas as imagens usando um único padrão, obtendo assim resultados mais satisfatórios.

A identificação da categoria semântica pode ainda oferecer novas capacidades para aplicações de movimentação de robôs autônomos (WU; REHG, 2008). Ser capaz de classificar um ambiente em uma categoria torna possível para um robô determinar a sua utilização, além de fornecer um forte indicativo de que objetos poderão ser encontrados no mesmo. Assim, um robô que executa serviços domésticos será capaz de identificar uma cozinha, mesmo sem conhecer a casa previamente, e pode prever que nesse ambiente

encontrará pratos, por exemplo.

1.1 Dificuldades na Classificação de Cenas

Para o ser humano, o entendimento de uma cena do mundo real ocorre de forma rápida e precisa. Nós podemos determinar rapidamente a classe de uma cena, mesmo sem assimilar todos os detalhes presentes. Para os sistemas automáticos, no entanto, a classificação de cenas é uma tarefa desafiadora devido a diversos fatores.

Há a chance de ocorrência de sombras e reflexões. Não há uma certeza sobre a presença e disposição dos elementos nas cenas de uma determinada categoria e, ainda há a possibilidade de oclusão parcial dos elementos presentes, fatos que geram uma grande variedade intraclasse. Por exemplo, na Figura 1, são apresentadas duas cenas da classe "Sala de Estar" e, apesar dos dois exemplos pertencerem à mesma classe, a disposição dos objetos neles presentes são distintas. Além disso, nem todos os elementos que aparecem em uma cena aparecem na outra. Ainda, a mesa de centro, que fornece uma pista forte sobre o ambiente, está oclusa na cena da esquerda sendo difícil identificá-la. Um outro motivo para o aumento da variabilidade intraclasse é a variação na perspectiva da foto, uma vez que a possibilidade de retratar uma cena através de diversos pontos de vista influencia nos elementos presentes e em suas disposições na cena. Existe também a variabilidade interclasse, já que existem classes que são bastante similares, podendo causar confusão entre cenas de categorias diferentes, como acontece na Figura 2, que pertence à classe "Autoestrada", mas que poderia pertencer à classe "Montanha", por exemplo.

Adiciona-se a essas dificuldades o problema da variação nas condições de iluminação, já que a iluminação pode destacar alguns elementos enquanto esconde outros. É o caso das imagens da classe "Costa" apresentadas na Figura 3, onde a iluminação na imagem à esquerda torna o céu preto e sem nenhum detalhe, mas destaca as nuvens na imagem à direita. A variação nas condições de escala também traz transtornos à classificação, pois altera a apresentação dos elementos em cena. A Figura 4 exemplifica essa situação exibindo imagens da classe "Prédio Alto". Na imagem à esquerda, há prédios tão próximos que não é possível observar os seus topos, já na imagem à direita, retratada a uma distância maior, é possível visualizar os topos de todos os prédios. Um outro fator que também dificulta a classificação de cenas é a rotação da câmera, uma vez que essa ação gera apresentações diferentes dos elementos presentes nas cenas.

Além dos problemas citados, há ainda outros fatores relacionados com a percepção humana: a ambiguidade e subjetividade do observador. A acurácia do classificador depende da consistência e da acurácia da anotação manual. Todos esses fatores representam um sério desafio mesmo para as estratégias de reconhecimento mais sofisticadas.

Figura 1 – Cenas da classe “Sala de Estar” retirados da base de dados de 15 categorias de cenas (LAZEBNIK; SCHMID; PONCE, 2006).

Figura 2 – Cena da classe “Autoestrada” retirada da base de dados de 15 categorias de cenas (LAZEBNIK; SCHMID; PONCE, 2006).

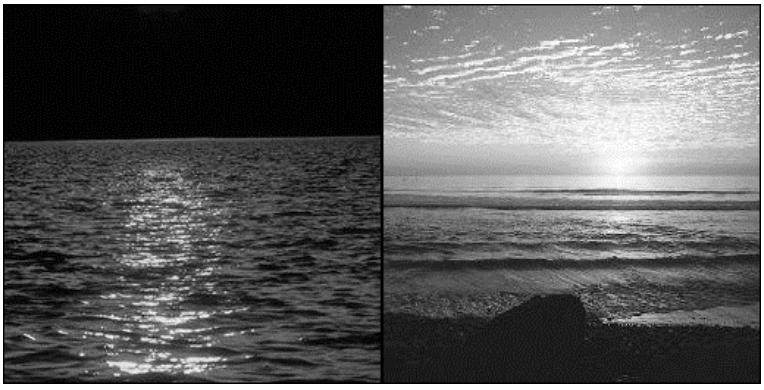


Figura 3 – Imagens da classe “Costa” retiradas da base de dados de 15 categorias de cenas (LAZEBNIK; SCHMID; PONCE, 2006).

1.2 Trabalhos Relacionados

Ao longo dos anos, várias técnicas foram propostas para resolver o problema de classificação de cenas. As abordagens mais tradicionais utilizam características de baixo

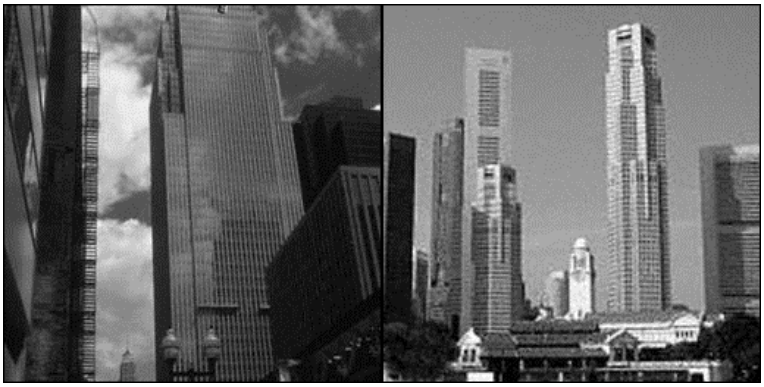


Figura 4 – Imagens da classe "Prédio Alto" retiradas da base de dados de 15 categorias de cenas (LAZEBNIK; SCHMID; PONCE, 2006).

nível, tais como cor e textura. Essas características são extraídas da imagem como um todo e cada imagem é representada por elas. Tal abordagem é chamada de representação global.

Em Gorkani e Picard (1994), as pirâmides de orientação multiescalares, uma representação de textura, são utilizadas para classificar imagens em duas classes: "Cidade" e "Subúrbio". Para isso, as informações sobre orientação são extraídas através de múltiplas escalas e, então, combinadas para encontrar a orientação perceptivelmente dominante.

Shen, Shepherd e Ngu (2005) propõem o *framework* CMVF (*Combined Multi-Visual Features*), um método híbrido não-linear para redução da dimensão de combinações de características, e considera quatro tipos de características: cores, utilizando um histograma de cores obtido no espaço de cor LUV (espaço de cor projetado para se aproximar de um espaço de cor perceptualmente uniforme); textura, extraída através de métodos baseados em filtros de Gabor para 6 frequências distintas; forma, representada através de um histograma de direção de bordas, obtidas através do operador Canny, e *layout* de cor.

Vailaya et al. (1999) tratam da classificação hierárquica de imagens de férias e utilizam representações distintas para cada nível hierárquico. As características por eles utilizadas variam de acordo com as classes abordadas. Por exemplo, para classificar uma cena como "Interna" ou "Externa", eles utilizam sub-blocos de tamanho 10 x 10 de momentos de cores no espaço LUV; já para a classes "Paisagem" e "Cidade", histograma de direções de bordas e vetor de coerência, que classifica cada pixel como coerente (parte de um grande componente conectado) ou não.

Chang et al. (2003) propõem um procedimento para atribuir rótulos semânticos, o CBSA (*Content-Based Soft Annotation*), e para representação de imagens utilizam cor e textura. Os autores trabalham com multirresolução e, para obter as características de cor, dividem as cores em 12 segmentos, 11 para as cores mais comuns em todas as culturas (preto, branco, vermelho, verde, amarelo, azul, marrom, púrpura, rosa, laranja e cinza) e

um para as demais. Nas resoluções mais grosseiras, as cores são caracterizadas por máscaras de 12 bits. Nas resoluções mais finas, são consideradas oito características de cor adicionais: histograma de cores, média das cores nos canais H, S e V (*Hue*, *Saturation* e *Value* - matiz, saturação e valor), variância das cores nos canais H, S e V e duas características de forma, ambas baseadas em momento: alongamento da cor, que caracteriza a forma da cor, e dispersão, que indica o quanto a cor se espalha dentro da imagem. Para a representação de textura, os autores optaram pela transformada discreta *wavelet*.

Ainda utilizando características de baixo nível, alguns métodos empregam abordagens locais, onde a imagem é dividida em blocos que são representados por suas características. Cada bloco é então classificado em uma certa categoria e finalmente a imagem é categorizada a partir das categorias de suas partes. Szummer e Picard (1998) trabalham com três tipos de características: cor, textura e informação de frequência. As imagens são classificadas em duas categorias: ambiente interno e externo. A característica de cor é representada através de um histograma de cores e tem 32 divisões por canal. Cada canal é representado pelo espaço de cor Ohta (OHTA; KANADE; SAKAI, 1980). As características de textura são obtidas através do MSAR (*Multiresolution Simultaneous Autoregressive model*), extraídas a partir das imagens em escalas de cinza em duas resoluções. As características de frequência são obtidas, primeiro, através do cálculo da magnitude da transformada de Fourier Discreta 2D e, depois, obtendo a transformada discreta do cosseno 2D. Todos os cálculos são feitos em blocos de tamanho 8 x 8 pixeis e, em cada região, a média dos resultados é calculada.

A grande desvantagem dos métodos que utilizam características de baixo-nível é que eles oferecem uma representação de imagem muito simples. O principal inconveniente em utilizar esse tipo de representação é que, quando há uma grande variabilidade intraclasse, ele não é suficiente para discriminar as diferentes categorias. Desse modo, esses métodos têm sido utilizados apenas para classificar as imagens em um pequeno número de categorias.

Uma outra abordagem para representação de cenas é através do uso de conceitos intermediários. Vogel e Schiele (2007) utilizam a modelagem semântica local para representação de cenas naturais. A abordagem proposta classifica regiões da imagem em classes de conceitos semânticos, tais como, água, pedra e folhagem. A imagem é então representada através da frequência de ocorrência desses conceitos semânticos locais. Os conceitos são extraídos de *grids* regulares de regiões 10 x 10 e são classificados através do classificador SVM (*Support Vector Machine*). O melhor resultado de classificação de conceitos é obtido com a concatenação do histograma de cores no espaço HSV, com 84 divisões, do histograma de direção de bordas, com 72 divisões, e 24 características extraídas da matriz de coocorrência de níveis de cinza.

Em Fredembach, Schroder e Susstrunk (2004), a imagem é segmentada em diversas regiões e, usando características específicas, as regiões são classificadas em diversas classes

semânticas. O resultado da classificação das regiões é utilizado para obter a classificação da imagem. Para a segmentação, os autores utilizam o algoritmo *k-means* e as regiões são representadas pelos *eigenregions* obtidos através da aplicação do PCA (*Principal Component Analysis*) nas regiões segmentadas.

Em Luo, Savakis e Singhal (2005), uma abordagem híbrida foi proposta, através da fusão de características de baixo nível e semânticas. A Figura 5 exibe a estrutura geral proposta. Dois conjuntos de características são extraídos da imagem: características de baixo-nível (por exemplo, cor, textura e bordas) e objetos semânticos (como grama e céu), que podem ser extraídos automaticamente. Os vetores concatenados são, então, apresentados a uma máquina de inferência que produz predicados semânticos, utilizados na decisão entre as classes, que, no caso são, "Ambiente Interno" e "Externo".

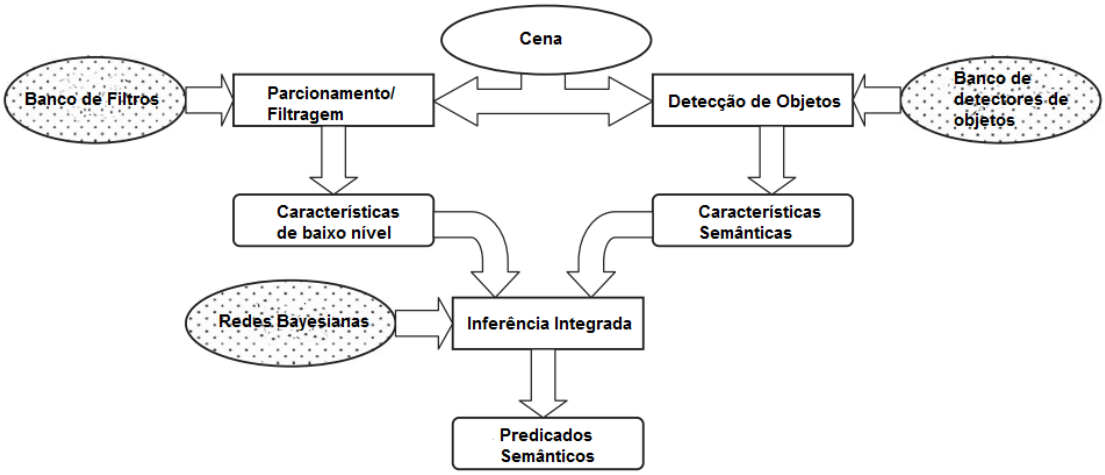


Figura 5 – Estrutura geral para o entendimento semântico da imagem (LUO; SAVAKIS; SINGHAL, 2005).

Oliva e Torralba (2001) propuseram um modelo computacional de reconhecimento de cenas do mundo real que utiliza um nível semântico intermediário, as dimensões perceptuais, para identificar o *gist* da cena, isto é, a informação significativa que um observador pode perceber a partir de uma visualização rápida da cena (POTTER, 1975). Esse modelo considera a cena como um objeto único (abordagem holística) e estima a sua estrutura através de dimensões perceptuais (naturalidade, abertura, irregularidade, expansão e rugosidade), gerando, assim, um espaço multidimensional no qual cenas pertencentes a uma mesma classe são projetadas em áreas próximas. O conjunto de dimensões perceptuais é conhecido como Envelope Espacial. No entanto, como tenta construir o "*gist* da cena", esse método é comumente referenciado como *gist* na literatura. As dimensões perceptuais podem ser estimadas através da representação espectral. Em Oliva e Torralba (2001), elas foram estimadas a partir da estatística de segunda ordem (DST - *Discriminant Spectral Templates*) e a partir do arranjo espacial das estruturas na cena (WDST - *Windowed Discriminant Spectral Template*). Em Oliva e Torralba (2006), os autores afirmam que uma

cena pode ser representada por uma combinação de características globais e que uma opção de representação é a combinação ponderada da saída de bancos de filtros multiescalares e multiorientados, como os filtros de Gabor. Assim, as saídas dos filtros são divididas em *grids* de tamanho N x N e, dentro de cada bloco, a intensidade média é calculada para representar a característica naquele bloco.

Em Liu, Kiranyaz e Gabbouj (2012), propôs-se uma melhoria nos resultados do descritor *gist* utilizando o particionamento radial angular - ARP (*Angular Radial Partitioning*). Dessa forma, ao invés de obterem a média dentro de cada bloco no *grid* N x N, como mencionado anteriormente, os autores particionam cada bloco dentro de A divisões utilizando o ARP. Como justificativa, os autores alegam que o algoritmo proposto não só delineia a estrutura de um bloco, como também oferece margem de manobra para liberdade espacial.

Uma deficiência na utilização dos conceitos intermediários é a necessidade de suas anotações manuais, o que envolve várias horas de trabalho. Além disso, na maioria das abordagens, há a necessidade da classificação de regiões em conceitos intermediários, e, como a classificação da imagem é baseada nas ocorrências desses conceitos, a classificação errada de regiões pode culminar na classificação errada da cena.

Evitando a anotação manual das imagens de treinamento, mas ainda adotando os conceitos intermediários, surgiram diversas abordagens que utilizam a técnica *bag-of-features* (também conhecida como *bag-of-visterms* ou *bag-of-words*). Inspirada nas técnicas de recuperação de informação, a abordagem *bag-of-features*, ilustrada na Figura 6, representa a imagem como uma coleção de *patches* locais que podem ser amostrados densamente ou a partir de pontos de interesse (*patches* salientes da imagem) detectados. Daí, é feita a extração de características desses *patches* através de algum descritor. As características extraídas são, então, quantizadas e geram as palavras visuais por meio de um algoritmo de agrupamento. A representação final da imagem é dada através de um histograma de palavras visuais. Um descritor bastante popular neste tipo de abordagem é o SIFT (*Scale Invariant Feature Transform*) (LOWE, 1999), que transforma a imagem em uma grande coleção de vetores locais, cada qual invariante à translação, escala e rotação e parcialmente invariante a mudanças na iluminação.

Fei-Fei e Perona (2005) propuseram uma abordagem onde a imagem é representada por uma coleção de palavras visuais, e cada palavra visual é representada como parte de um tema. A distribuição dos temas, bem como a distribuição das palavras visuais dentro dos temas são aprendidas de maneira automática (sem supervisão) através de uma modificação do modelo LDA (*Latent Dirichlet Allocation*), um modelo probabilístico generativo (isto é, que explica um conjunto de dados observados através de parâmetros não observáveis), proposto por Blei, Ng e Jordan (2003) para representar e aprender modelos de documentos.

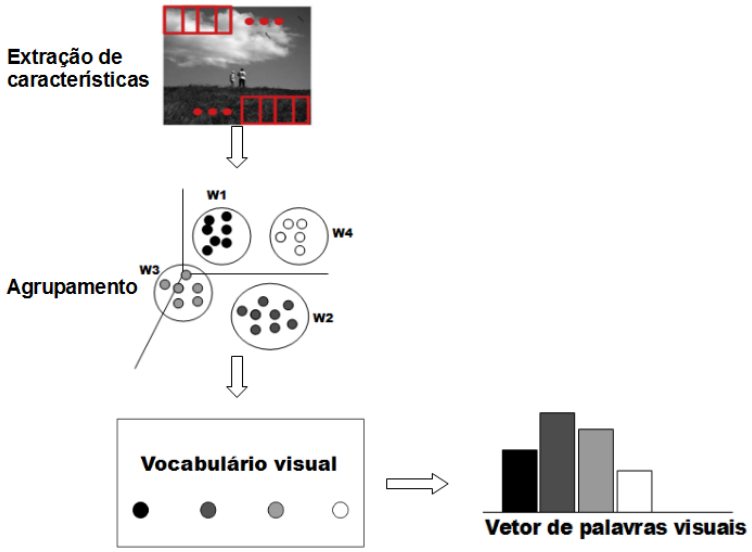


Figura 6 – Abordagem *bag-of-features*.

Em Quelhas et al. (2005) e Quelhas et al. (2007), foi proposto o *bag-of-visterms* (BOV). Nessa abordagem, primeiro, os descritores locais são quantizados em palavras visuais, as quais são chamadas de termos visuais (*visterms*). Afim de fornecer informação do *layout* espacial e tratar os sinônimos (várias palavras visuais podem representar um mesmo conteúdo de cena) e a polissemia (uma única palavra visual pode representar diferentes conteúdos de cena), os autores utilizam o PLSA (*Probabilistic Latent Semantic Analysis*) (HOFFMAN, 2001), um modelo probabilístico generativo, e, também, modelaram imagens como sendo uma mistura de tópicos aprendidos automaticamente.

A técnica de *bag-of-features* representa a imagem como uma coleção desordenada de características locais e, portanto, não considera a informação espacial. Para aproveitar esse tipo de informação, Lazebnik, Schmid e Ponce (2006) propuseram o SPM (*Spatial Pyramid Matching*) um método que realiza particionamentos consecutivos da imagem em sub-regiões cada vez mais finas e calcula o histograma dentro de cada uma delas. Os histogramas são então ponderados utilizando-se o método *pyramid matching kernel* (GRAUMAN; DARRELL, 2007). Ergul e Arica (2010) propuseram um método que combina o SPM e o PLSA. O esquema, chamado de PLSA em Cascata, executa o PLSA em um sentido hierárquico depois que o *bag-of-features* é extraído a partir de *paches* locais amostrados densamente. A informação do *layout* espacial é associada através da divisão da imagem em regiões sobrepostas em diferentes níveis de resolução e da implementação de um modelo PLSA para cada região individualmente.

Partindo da hipótese de que as características da imagem em torno das regiões de interesse podem oferecer uma informação útil ou uma pista sobre a região, Qin e Yung (2010) propuseram uma extensão do *bag-of-features*. Baseado no aprendizado não-supervisionado, o método introduz informação contextual nas regiões de interesse a partir

de regiões vizinhas e de regiões em escalas de menor resolução. A informação vinda das regiões de fora da região de interesse são chamadas de contexto e as palavras visuais representadas dessas maneira são chamadas de palavras visuais contextuais. O objetivo de introduzir a informação contextual é reduzir a ambiguidade das palavras visuais utilizadas para representar as regiões locais. A Figura 7 apresenta os pontos amostrados dos *patches* em diferentes escalas e apresenta um exemplo de região de onde vem o contexto das escalas mais grossas e exemplos das regiões de onde vem o contexto das regiões vizinhas. Os *patches* são amostrados densamente em cada nível da escala.

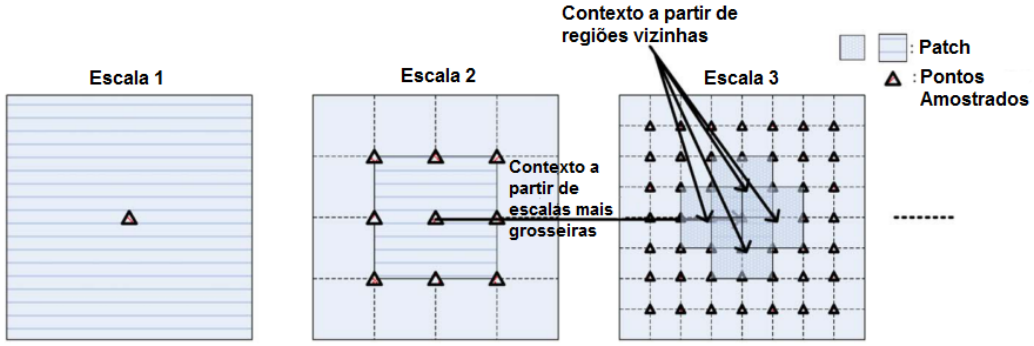


Figura 7 – Os pontos de amostragem dos *patches* de imagem em diferentes escalas e a localização das regiões que formam o contexto a partir da escala mais grosseira e da vizinhança (QIN; YUNG, 2010) .

Também utilizando a estratégia de inserir informação contextual, Liu, Xu e Feng (2011) propuseram o método *Region Contextual Visual Words*, que integra informação contextual na abordagem *bag-of-features*, e, com esse fim, apresentaram o modelo *Region Conditional Random Fields*, no qual o aprendizado de cada palavra visual depende das demais palavras visuais existentes na região. O objetivo do método é diferenciar *patches* que possuem aparência semelhante, mas conceitos semânticos distintos, visto que esses casos não são captados pelo método tradicional, pois as palavras visuais são construídas de forma independente.

Li et al. (2011) apresentaram o *bag-of-words* contextual, CBoW (*Contextual Bag-of-Words*), uma representação que modela dois tipos de relações contextuais típicas entre *patches* locais: a relação semântica conceitual e a relação da vizinhança espacial. Para modelar a relação semântica conceitual, as palavras visuais são agrupadas em vários níveis semânticos de acordo com similaridade de distribuição de classe induzidas por elas, medida pela divergência Kullback e Leibler (1951). Para explorar a relação espacial da vizinhança, uma técnica de extração automática de termos é adotada para medir a certeza de que as palavras visuais vizinhas são relevantes. Grupos de palavras com alta relevância são usados e suas estatísticas são incorporadas na representação *bag-of-features*.

Bolovinou, Pratikakis e Perantonis (2013) propuseram o método *Bag of spatio-*

visual words para a inferência de contexto na classificação de cenas. Nesse método um espaço de características de palavras visuais é construído e cada imagem é representada por um histograma de palavras visuais e, em paralelo, é obtida a representação da imagem através de um agrupamento ordenado de palavras visuais. Desse modo, um espaço de características de palavras visuais espaciais (*spatio-visual words*) é construído contendo conjuntos ordenados localmente de palavras visuais. Cada imagem é representada por um histograma de palavras visuais espaciais.

Apesar de seus resultados inspiradores, a abordagem *bag-of-features* tem alguns problemas. O primeiro deles diz respeito ao custo computacional, uma vez que é preciso obter inicialmente os descritores locais, e os melhores desempenhos ocorrem quando estes são amostrados densamente, e, depois, submetê-los a um algoritmo de agrupamento, o que gera um processamento considerável. Um outro problema é o número de palavras visuais utilizadas. Esse parâmetro influencia no desempenho da classificação e varia de acordo com o conjunto de dados. Esse problema é agravado nos modelos de aspecto latente como o PLSA e o LDA onde o número de tópicos também deve ser especificado.

Recentemente, Wu e Rehg (2011) propuseram o CENTRIST (CENsus TRansform hISTogram), que representa a imagem através de um histograma de valores obtidos através da aplicação da transformada census (ZABIH; WOODFILL, 1994), uma transformada local não-paramétrica, em janelas de pixeis de tamanho 3 x 3. Nesse mesmo trabalho, também foi proposto o *spatial* PACT (*spatial Principal Component Analysis of Census Transform*) que utiliza o método de pirâmide espacial (SPM) combinado com o CENTRIST. Meng, Wang e Wu (2012), por sua vez, propuseram uma versão estendida do histograma de transformada census, incluindo a diferença local de sinal e a informação de magnitude no cálculo da transformada.

Como dito anteriormente, uma das necessidades a serem atendidas com a classificação automática de cenas é a organização de imagens, seja na internet ou em computadores pessoais. Os métodos propostos na literatura, visando o alcance de taxas de reconhecimento cada vez mais altas, tornaram-se bastante complexos, exigindo que os operadores dos sistemas de classificação tenham conhecimentos específicos, como o conceito de palavras visuais ou tópicos. Esses métodos, portanto, são inviáveis para o uso em aplicações voltadas para computadores ou dispositivos pessoais, dos quais grande parte dos usuários é totalmente leiga nos assuntos ligados ao reconhecimento de padrões. Além disso, há um custo computacional relevante associado a tais métodos. Os métodos mais simples, por sua vez, são capazes de classificar em poucas classes, o que também não é viável diante do volume de imagens criadas a cada dia.

A utilização de transformações locais não-paramétricas para a representação de imagens, como proposto por Wu e Rehg (2011), leva a resultados de classificação competitivos e diminui a necessidade de um conhecimento avançado por parte dos usuários

do sistema de classificação, visto que a quantidade de parâmetros necessários é pequena. Além disso, esse método apresenta baixa complexidade computacional.

No entanto, o uso da abordagem de pirâmide espacial realiza a aplicação do PCA, que é uma técnica utilizada para diminuir a dimensionalidade dos dados enquanto retém a variação presente nos dado originais. Essa estratégia não leva em consideração os grupos dentro dos dados, visto que é uma técnica não supervisionada. Assim, apesar de ser possível que grupos separados, apareçam nos dados reduzidos, isso nem sempre é o caso, e a redução de dimensionalidade pode reduzir a existência de grupos separados, ou seja as escolhas feitas pelo PCA nem sempre serão as melhores para a discriminação (WEBB, 2002).

Diante do exposto, pode-se notar que a tendência atual dos trabalhos é classificação por contexto (QIN; YUNG, 2010; LIU; XU; FENG, 2011; LI et al., 2011; BOLOVINOU; PRATIKAKIS; PERANTONIS, 2013), o que melhora os resultados, mas implica em algoritmos complexos de classificação, pois, além de utilizarem representações intermediárias, através da técnica de *bag-of-features*, demandam um esforço para identificar as relações contextuais.

Sendo assim, tem-se aqui um problema a ser tratado: como acrescentar informação contextual utilizando uma técnica que não utilize conceitos intermediários sem elevar demasiadamente a dimensão do vetor de características, o que acarretaria no uso de técnicas de redução de dimensionalidade?

1.3 Objetivos do Trabalho

O objetivo geral deste trabalho é propor um descritor eficiente de imagens de cenas que possibilite a criação de sistemas de classificação em apenas uma classe, incorporando a informação de contexto, sem ter que elevar demasiadamente a dimensão do vetor de características. O dito descritor não deve empregar representações intermediárias como aquelas baseadas em *bag-of-features*, evitando assim um aumento da complexidade computacional. Concomitantemente, a presente abordagem não deve necessitar de ajustes de parâmetros de difícil compreensão por parte de usuários leigos na área de reconhecimento de padrões, o que favorece sua aplicabilidade.

Para atingir esse objetivo geral, os seguintes objetivos específicos são definidos:

- averiguar se é possível encontrar, para o problema da classificação de cenas, descritores, que, quando juntos, alcancem taxas maiores de acerto na classificação;
- o modelo proposto deve ser voltado aos usuários de computadores pessoais, ou seja, deve-se propor um descritor que não necessite que parâmetros complexos sejam informados e que não se exija um grande volume de processamento.

Para atender o objetivo geral mencionado acima, duas hipóteses precisam ser avaliadas:

(i) se é possível extrair a informação contextual da imagem sem que para isso tenha que se apelar para o uso de técnicas que adotem representações intermediárias ou que exijam entrada de parâmetros.

(ii) se a adição da informação contextual melhora a representação da imagem de forma que resultados competitivos de classificação sejam alcançados, mesmo quando mais de duas classes estão envolvidas.

1.4 Contribuições

- Proposta de um descritor de imagens de baixa complexidade, por fazer uso de cálculos simples, de melhor usabilidade, quando comparado com os métodos existentes na literatura, e que apresenta resultados de classificação competitivos;
- Aumento da informação sobre textura na representação de imagens através da introdução de informação contextual;
- Exploração da informação de contraste associada à informação espacial com intuito de obter uma representação de imagem mais eficiente para a classificação;
- Obtenção de taxas de classificação competitivas através da combinação de técnicas de representação de imagens distintas.

1.5 Trabalhos Publicados

GAZOLLI, K. e SALLES, E. A contextual image descriptor for scene classification. In: *Online Proceedings on Trends in Innovative Computing*, São Carlos, Brasil, p. 66:71, 2012.

GAZOLLI, K. e SALLES, E. Combining holistic descriptors for scene classification. In: *Proceedings of the International Conference on Computer Vision Theory and Applications (VISAPP 2013)*, Barcelona, Espanha, p. 315:320, 2013.

GAZOLLI, K. e SALLES, E. GistCMCT: a combination of holistic descriptors for scene classification. Journal of Information Assurance and Security, v. 8, p. 129:136, 2013. (Ciência da Computação: Qualis B3; Engenharias IV: Qualis B5)

GAZOLLI, K. e SALLES, E. Using holistic features for scene classification by combining classifiers. Journal of Winter School of Computer Graphic, v. 21, n.1, p. 41:48, 2013. (Engenharias IV: Qualis B3)

GAZOLLI, K. e SALLES, E. Exploring neighborhood and spatial information for improving scene classification. Pattern Recogniton Letters, v. 46, p. 83:88, 2014. (Fator de impacto: 1,529; Ciência da Computação: Qualis A1; Engenharias IV: Qualis B1)

1.6 Organização do Trabalho

Este trabalho foi dividido em seis capítulos descritos a seguir.

No Capítulo 1, foi apresentado o problema, os trabalhos relacionados e o objetivo do trabalho.

No Capítulo 2, é apresentado o descritor Transformada Census Modificada Contextual - CMCT (*Contextual Modified Census Transform*), um descritor que associa informações sobre as estruturas locais com informação contextual.

O Capítulo 3 propõe a utilização do descritor CMCT com outras duas abordagens, apresentando assim, dois novos métodos para representação da imagem: o GistCMCT e o GistCMCT combinado com o *spatial* MCT.

O Capítulo 4 apresenta o descritor CMCT Estendido juntamente com a sua combinação com outras abordagens.

No Capítulo 5, são apresentados os experimentos em quatro bases de dados públicas e os resultados são discutidos.

Finalmente, no Capítulo 6, estão as conclusões e os trabalhos futuros.

2 Transformada Census Modificada Contextual (CMCT)

Neste capítulo, é proposto o descritor Transformada Census Modificada Contextual (CMCT - *Contextual Modified Census Transform*). O CMCT é um descritor holístico (ou seja, não divide a imagem em objetos ou regiões) inspirado no CENTRIST (WU; REHG, 2011), que leva em consideração o contexto na representação das estruturas presentes nas janelas utilizadas no cálculo da transformada não-paramétrica. Por contexto, entende-se as estruturas contidas em janelas vizinhas.

O CMCT captura as propriedades estruturais da imagem, assim como o CENTRIST. Entretanto, diferentemente do CENTRIST, o CMCT adiciona informação sobre a região em torno das estruturas locais, modelando as estruturas locais formadas por estruturas locais vizinhas, através de uma abordagem recursiva. Assim, a informação de contexto é introduzida a partir da comparação da estrutura contida na janela central com as estruturas contidas nas janelas que a cercam.

Primeiro, serão expostas as técnicas que foram utilizadas para a obtenção da Transformada Census Modificada Contextual e, depois o descritor proposto será apresentado.

2.1 A Transformada Census

A transformada census (CT - *Census Transform*) (ZABIH; WOODFILL, 1994) é uma transformada local não-paramétrica proposta originalmente para estabelecer correspondência entre *patches* de imagens. Transformadas não-paramétricas locais se baseiam na ordem relativa das intensidades dos pixeis e não nos valores da intensidade diretamente. Assim, a transformada census, $\mathcal{C}(x)$, compara o valor da intensidade de um pixel com a intensidade dos pixeis vizinhos. Se a intensidade do pixel central é maior ou igual a do seu vizinho um bit 1 é colocado no local correspondente. Caso contrário, é colocado um bit zero, como apresentado a seguir

$$\mathcal{C}(x) = \bigotimes_{y \in \mathcal{N}(x)} \zeta(I(x), I(y)), \quad \zeta(m,n) = \begin{cases} 1, & m \geq n \\ 0, & c.c. \end{cases} \tag{2.1}$$

onde $I(x)$ é o valor da intensidade do pixel central na posição x, $I(y)$ é o valor da intensidade do pixel na posição y, $\bigotimes$ indica a operação de concatenação e $\mathcal{N}(x)$ define a vizinhança local do pixel na posição x. Se uma vizinhança de tamanho 3 x 3 é utilizada, considerando-se que $x \notin \mathcal{N}(x)$, os 8 bits gerados pelas comparações de intensidade podem

ser colocados em qualquer ordem (neste trabalho, eles foram selecionados da esquerda para direita e de cima para baixo) e convertidos em um número na base decimal entre 0 e 255, chamado de valor CT.

Como o valor CT não é um nível de intensidade e todos os bits em $\mathcal{C}(x)$ têm o mesmo nível de significância, ele pode ser interpretado como um índice das estruturas definidas em $\mathcal{N}(x)$ com o centro igual a zero. O processo como um todo pode ser pensado como um filtro não-linear onde a imagem de entrada é associada ao índice da melhor estrutura para representar aquele pedaço da imagem (FROBA; ERNST, 2004). A Figura 8 apresenta seis possíveis estruturas de tamanho 3 x 3 pixeis. Como a comparação é feita com o pixel central, e em todos os exemplos ele tem valor 0 (zero), as seguintes sequências de bits são geradas: 11100000, 11100111, 11110100, 00101111, 01000110, 00101110, considerando as estruturas da esquerda para direita e de cima para baixo. Assim, cada estrutura recebe um valor CT diferente, ou seja, é associada a um índice diferente. Sendo, portanto, consideradas como estruturas diferentes.

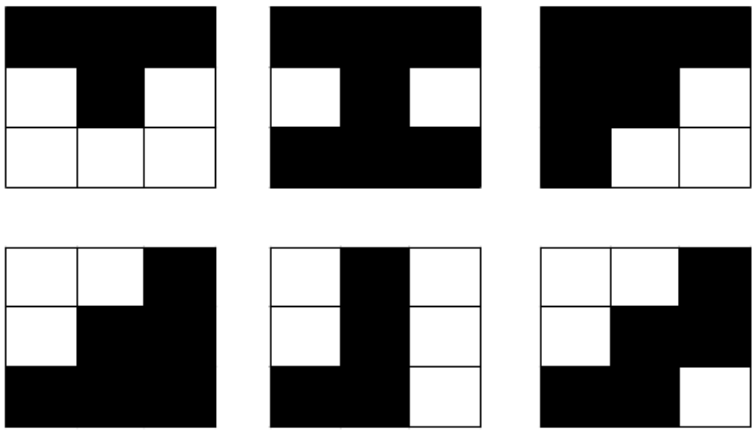


Figura 8 – Exemplos de padrões de pixeis que geram valores CT distintos. As seguintes sequências são geradas (de cima para baixo e da esquerda para direita): 11100000, 11100111, 11110100, 00101111, 01000110, 00101110

Aproximadamente ao mesmo tempo que a transformada census foi introduzida por Zabih e Woodfill (1994), o operador LBP (*Local Binary Pattern*) foi introduzido por Ojala, Pietikäinen e Harwood (1996), com o objetivo de descrever padrões binários locais em texturas. Esse operador também mapeia a vizinhança local de um pixel em uma sequência de bits, a única diferença em relação à transformada census é a ordem dos bits (MARCEL; RODRIGUEZ; HEUSCH, 2007).

2.2 CENTRIST

O *CENsus TRansfrom hISTogram* (CENTRIST) é um descritor visual proposto por Wu e Rehg (2011) para reconhecer locais e cenas. O CENTRIST, $H(ct)$, é uma

representação holística que captura propriedades estruturais através da modelagem de estruturas locais e pode ser definido da seguinte forma

$$H(\mathrm{ct}) = n_{\mathrm{ct}} \tag{2.2}$$

onde ct corresponde ao ct-ésimo valor CT no intervalo [0,255] e n_{ct} é o número de ocorrências do valor ct na imagem. Em outras palavras, o CENTRIST é um histograma de valores CT.

A motivação dos autores ao proporem o CENTRIST foi a criação de um descritor que capturasse as propriedades estruturais gerais da imagem, tais como formas retangulares, superfícies lisas, e suprimisse a informação sobre detalhes de textura, já que consideram que os detalhes da textura contribuem negativamente na etapa de classificação. Para exemplificar a importância das propriedades estruturais no reconhecimento de cenas, os autores aplicaram o filtro Sobel em uma imagem, pois esse filtro é utilizado na detecção de contornos e elimina detalhes de textura. A Figura 9 exibe a imagem pertencente à classe "Cozinha" e seus respectivos gradientes Sobel. Observando a imagem após a aplicação do filtro, é possível constatar que, apesar da ausência dos detalhes de textura e da acentuação das estruturas espaciais, um humano ainda é capaz de classificá-la como pertencente à classe "Cozinha".

(a) (b)

Figura 9 – (a) Imagem de uma cozinha. (b) A mesma imagem após filtragem. Imagens retiradas de Wu e Rehg (2011).

A transformada census utiliza o pixel central como limiar, o que faz com que diversas estruturas que apresentam combinações distintas de níveis de cinza, mas que não apresentam uma variação de níveis de cinza muito intensa entre si, sejam mapeadas em um mesmo padrão binário. Assim, a diferença entre essas estruturas são ignoradas e, como o cálculo do valor CT é feito localmente em janelas de tamanho 3 x 3, tem-se, como consequência, a supressão de informações sobre alguns detalhes de textura na representação produzida pelo CENTRIST. A Figura 10 ilustra essa situação. Nela, são apresentadas quatro estruturas de tamanho 3 x 3 pixeis e, apesar das diferentes combinações de níveis

de cinza, todas essas estruturas são mapeadas no padrão binário “00010110”, ou seja, a transformada census as considera como sendo iguais.

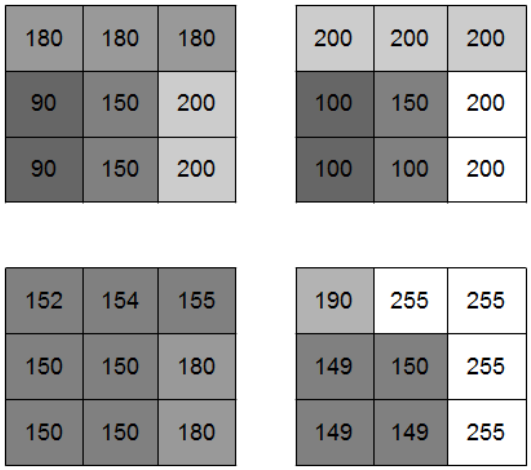


Figura 10 – Diferentes estruturas que são mapeadas em um mesmo padrão binário pela transformada census.

Uma vez que, na transformada census, o padrão binário é o resultado da comparação de cada pixel na janela com o pixel central e os bits são gerado de acordo com a posição de cada pixel na janela, essa operação é capaz de detectar superfícies lisas, pois nessa situação ocorre um grande número de janela onde não há variação na intensidade dos pixeis. Portanto, o CENTRIST herda da transformada census a capacidade de mapear estruturas similares, características locais da imagem, em um mesmo padrão binário, o que lhe confere certa robustez a iluminação. Isso se deve ao fato de que a transformada census se vale da ordem relativa das intensidades dos pixeis e não dos valores da intensidade. Já o CENTRIST interpreta o valor CT como uma variável aleatória uma vez que pequenos desvios podem ocorrer em seus valores mas, em essência, são eles que definem a estrutura local. Nesse sentido, seria possível dizer que as diversas realizações de uma mesma estrutura guardam uma certa relação entre si, o que poderia ser melhor capturado pela medição da frequência relativa de quantas vezes ocorreu aquele padrão, capturado pela resposta da transformada census.

De modo a ter um estimador simples e eficaz para representar a estrutura local, usa-se então o histograma para caracterizá-la. O histograma, como aproximador de uma função de densidade de probabilidade, mede amostralmente a ocorrência das estruturas locais detectadas pelo CT na imagem. Ou seja, o histograma funciona como um caracterizador de estruturas locais que constituem a cena em questão. Além disso, a comparação entre os pixeis permite detectar as variações entre claro e escuro, o que indica a existência de contornos e a posição dos bits no padrão binário dá pistas sobre a direção desses contornos. A Figura 11 exemplifica a detecção de contornos, exibindo duas estruturas distintas em janelas de tamanho 6 x 6 pixeis. Para cada uma dessas estruturas, é exibido o histograma de valores CT, calculados em janelas de tamanho 3 x 3 pixeis, onde se verifica a ocorrência

do valor CT 68 na representação da primeira estrutura (linha vertical) e 24, na da segunda (linha horizontal). Assim, apesar das duas figuras exibirem linhas, essas linhas apresentam direções diferentes e essa diferença se reflete no valor CT obtido para cada uma delas.

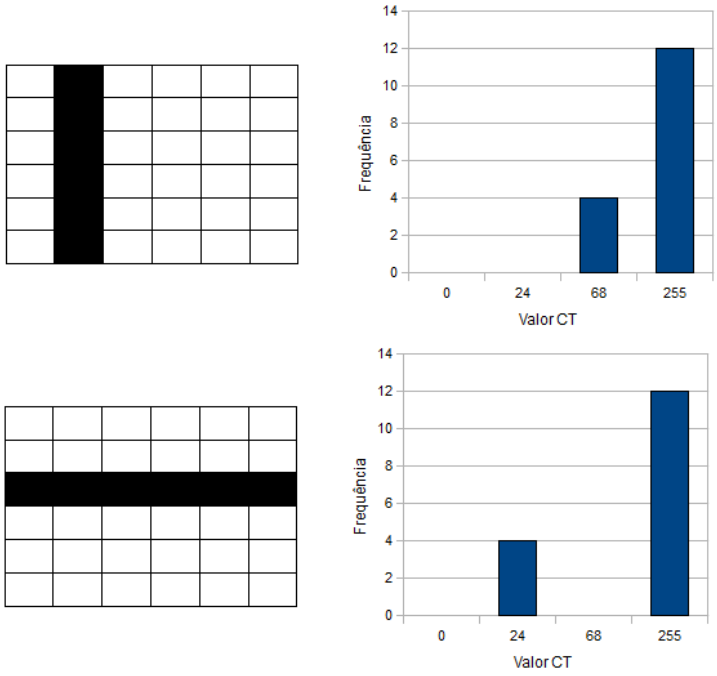


Figura 11 – Linhas vertical e horizontal e seus respectivos histogramas CT.

Como vantagens do CENTRIST, pode-se citar:

- como calcula um histograma de valores, sua obtenção é feita de forma rápida;
- poucos parâmetros a serem informados;
- é menos sensível à variações na iluminação, na correção gama e em ruídos introduzidos pela câmera.

Como limitações do CENTRIST, tem-se:

- não é invariante à rotação ou à mudança de escalas;
- não é um descritor preciso de formas;
- ignora a informação sobre cores.

2.3 Transformada Census Modificada

A transformada census tem se mostrado menos sensível à variações na iluminação, à correção de gama e a ruídos introduzidos pela câmera e é útil na extração das bordas

nas imagens (BHAVANI et al., 2007). No entanto, a transformada census não captura informações sobre algumas estruturas (FROBA; ERNST, 2004). Para evitar essa perda de informação, Froba e Ernst (2004) propuseram a Transformada Census Modificada (MCT - *Modified Census Transform*), $\Gamma(x)$. Para calculá-la, uma janela de pixeis é considerada e a média, $\overline{I}(x)$, desses pixeis é calculada. Todos os pixeis pertencentes à janela são comparados com $\overline{I}(x)$. Se o valor da intensidade do pixel é maior ou igual a $\overline{I}(x)$, um bit 1 é gerado. Caso contrário, um bit 0 é colocado no local correspondente, como apresentado a seguir

$$\Gamma(x) = \bigotimes_{y \in \mathcal{N}'(x)} \zeta(I(y), \overline{I}(x)), \quad \zeta(m,n) = \begin{cases} 1, & m \geq n \\ 0, & c.c. \end{cases} \tag{2.3}$$

onde $\overline{I}(x)$ é a média da intensidade dos pixeis na janela com o centro na posição x, $I(y)$ é a intensidade do pixel na posição y e $\mathcal{N}'(x)$ é a vizinhança espacial local do pixel na posição x, de forma que $\mathcal{N}'(x) = \mathcal{N}(x) \cup x$. Quando uma janela de tamanho 3 x 3 é utilizada, a técnica MCT gera 9 bits. Estes bits são, então, convertidos em um número na base decimal entre 0 e 511, chamado aqui de valor MCT. Neste trabalho, será utilizado o MCT, porém, não será feita a comparação da média com o pixel central. Dessa forma, o MCT aqui utilizado gerará apenas 8 bits e será convertido para um valor entre 0 e 255. Essa modificação tem como objetivo reduzir o valor do descritor proposto que será apresentado na Seção 2.4. Para diferenciar o MCT adotado, com 8 bits, do MCT original, o primeiro será referido como MCT8.

Na Figura 12, pode-se notar que o MCT8 é capaz de diferenciar algumas estruturas que são consideradas iguais pela transformação census, pois, apesar das janelas apresentarem estruturas diferentes, elas geram o mesmo valor CT, enquanto os valores MCT8 são distintos. Isso acontece porque o CT compara a intensidade dos seus pixeis em uma janela apenas com o pixel central, o que não é suficiente para detectar algumas diferenças entre os níveis de cinza. Por exemplo, quando o pixel central é o maior valor presente na janela, o valor CT obtido é sempre 255, não importa quais sejam os valores dos demais pixeis. O MCT8, por sua vez, compara as intensidades com a média das intensidades na janela e, portanto, é mais sensível às variações nos níveis de cinza. Esse fator permite que o MCT8 detecte mais estruturas do que a abordagem original. Por exemplo, a média assume o maior valor na janela, apenas quando todos os pixeis que a constituem forem iguais.

2.4 Transformada Census Modificada Contextual (CMCT)

Nesta seção, é proposta a Transformada Census Modificada Contextual (CMCT), um descritor de imagens inspirado no CENTRIST, que integra a informação sobre estruturas locais com informação contextual, com o intuito de diferenciar pedaços (ou janelas) de imagens que possuam estruturas similares, mas que apresentam diferenças significantes na

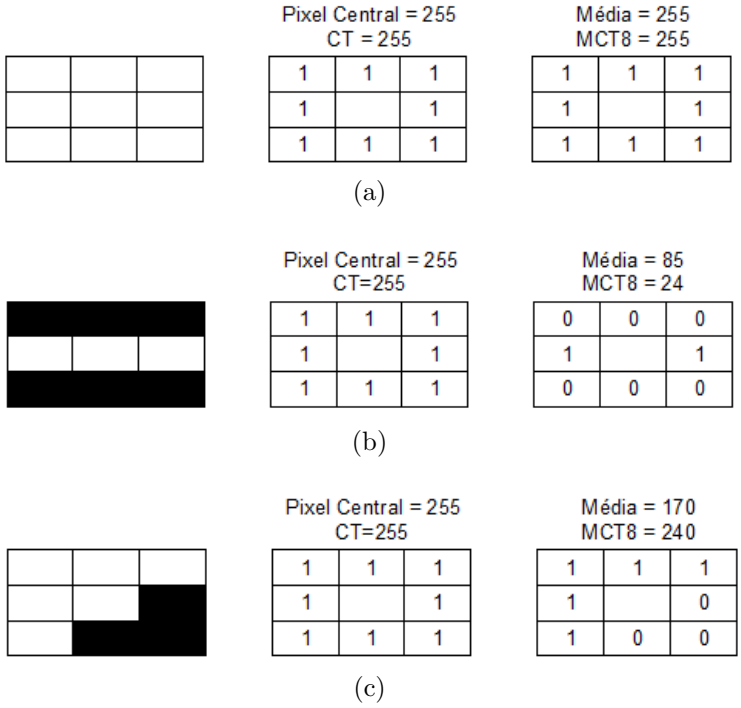


Figura 12 – Estruturas que são diferenciadas pelo MCT8, mas confundidas pelo CT. Em (a), os valores CT e MCT8 são os mesmos. Em (b), o valor MCT8 identifica os pixeis na cor preta, enquanto o CT, não. Em (c), a situação se repete.

vizinhança. Para isso, essa técnica 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. Essa informação adicional pode melhorar a representação da imagem e é chamada, aqui, de contexto.

Para entender melhor a relevância do contexto, pode-se fazer uma analogia com as palavras em uma frase. Há palavras que assumem diferentes significados de acordo com o contexto, o que é chamado de polissemia. É o caso da palavra MANGA nas frases abaixo:

- não posso sair com essa manga rasgada;
- não gosto de suco de manga.

Enquanto a primeira frase trata de uma parte de uma peça do vestuário, a segunda se refere a uma fruta. A diferença só é identificada através das demais palavras utilizadas em cada frase, ou seja, pelo contexto. O mesmo pode acontecer em uma imagem. Uma região branca em torno de uma região azul, provavelmente, pertence a uma nuvem, mas se essa mesma região estiver em uma região marrom, então, é provável que se trate de neve.

A Figura 13 esquematiza o processo para a obtenção da Transformada Census Modificada Contextual (CMCT). Primeiro, o valor MCT8 é calculado para cada pixel da imagem a ser representada. Depois, é gerado um histograma para toda a imagem de 256 posições com os valores MCT8 encontrados, o histograma de MCT8. Neste ponto,

portanto, é realizada a modelagem da distribuição das estruturas locais da imagem. Então, uma nova imagem é criada, a imagem MCT8, com os pixeis da imagem original sendo substituídos pelos seus respectivos valores MCT8, conforme mostra a Figura 14.

Na sequência, o valor MCT8 é calculado para cada pixel da nova imagem e é gerado, para toda a imagem MCT8, um histograma de 256 posições a partir desses novos valores MCT8 encontrados. Esse novo histograma é referenciado como histograma do contexto. O cálculo do MCT8 da imagem MCT8 é, portanto, uma abordagem recursiva e o histograma obtido neste ponto modela a distribuição das estruturas formadas por estruturas locais vizinhas. O contexto vem do fato dos valores obtidos nesta fase serem frutos da comparação da estrutura local central com as estruturas locais vizinhas. O histograma de MCT8 e o histograma do contexto são, então, concatenados, gerando, assim, um descritor para a imagem com 512 posições.

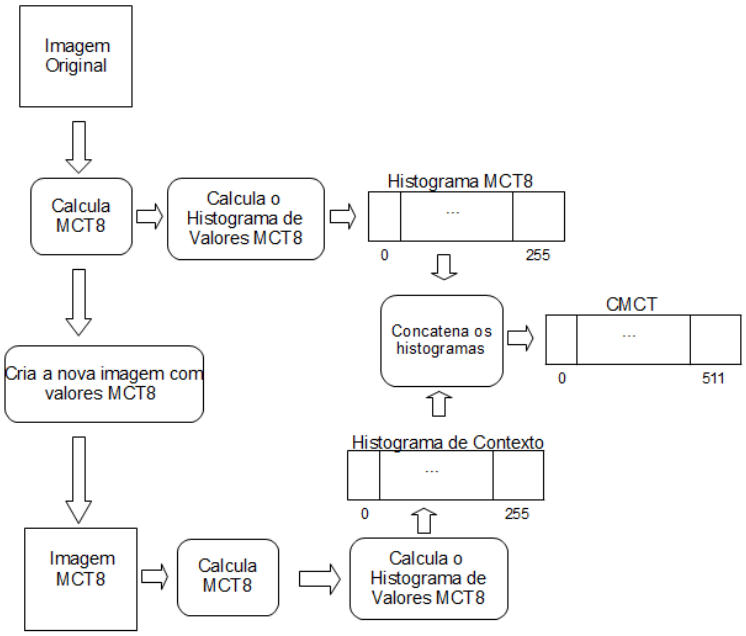


Figura 13 – Processo de extração do CMCT.

2.4.1 O Contexto

O MCT captura as estruturas locais da imagem e os valores por ele gerados agem como um índice para essas estruturas. Quando se criar uma nova imagem substituindo o valor do pixel pelo seu respectivo valor MCT8, como ilustrado na Figura 15, e se aplica o MCT8 novamente, modela-se as relações existente entre as estruturas locais. Isso significa que os valores obtidos através da aplicação do MCT8 na nova imagem são índices para as estruturas formadas pelas estruturas da imagem original. Assim, estruturas parecidas gerarão valores MCT8 iguais, mas poderão gerar valores MCT8 diferentes na nova imagem se estiverem inseridas em regiões com estruturas locais distintas. Essa situação é ilustrada

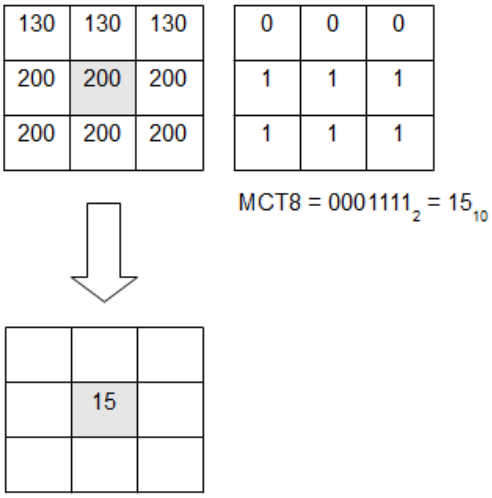
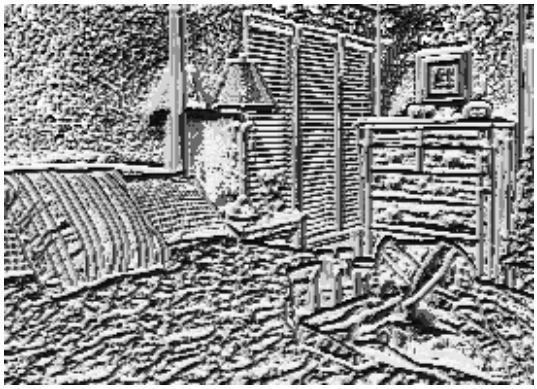


Figura 14 – O valor da intensidade do pixel é substituído pelo seu respectivo valor MCT8.

na Figura 16. Nela há uma janela de tamanho 9 x 9 pixeis contendo um padrão que se repete e, como pode-se verificar nos resultados da aplicação do MCT8 utilizando-se janelas de tamanho 3 x 3, os valores MCT8 para esse padrão também se repetem (células destacadas em cinza). Nos resultados da aplicação do MCT8 na imagem MCT8, no entanto, essa repetição desaparece, pois, apesar das estruturas locais serem iguais (isto é, as janelas de tamanho 3 x 3 pixeis são iguais em cada ocorrência do padrão), elas se encontram cercadas por conjuntos de pixeis diferentes, ou seja, apresentam vizinhanças diferentes, gerando assim valores de contexto diferentes.

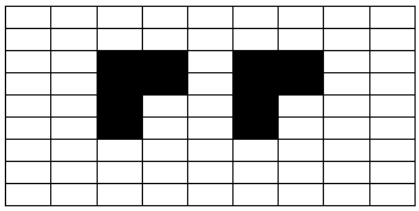
(a) Imagem original



(b) Imagem MCT8

Figura 15 – Um exemplo de uma imagem com os valores da intensidade do pixel substituídos pelos seus respectivos valores MCT8. Imagem retirada da base de dados de 15 categorias (LAZEBNIK; SCHMID; PONCE, 2006).

Na Figura 17, há a repetição do mesmo padrão em uma janela de tamanho 9 x 9 pixeis, porém com um deslocamento e, novamente, é possível verificar a repetição nos resultados da aplicação do MCT8 em janelas de tamanho 3 x 3 e, como no exemplo anterior, essa repetição não ocorre nos resultados da aplicação do MCT8 na imagem MCT8.



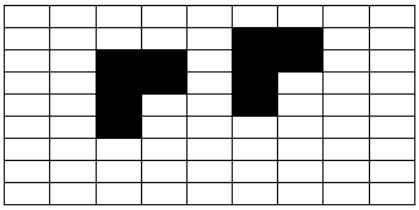
MCT8

254	252	249	250	252	249	251
246	244	233	226	244	233	235
214	149	43	70	149	43	111
214	157	43	86	157	43	127
215	191	111	215	191	111	255
223	191	127	223	191	127	255
255	255	255	255	255	255	255

MCT8 para a Imagem MCT8

252	248	248	248	248
244	244	233	226	228
212	149	43	70	149
150	157	43	86	29
151	47	71	151	47

Figura 16 – Figura com padrões que se repetem (esq.). Resultado da aplicação do MCT8 na figura original (meio). Resultado da aplicação do MCT8 na figura do meio (dir.).



MCT8

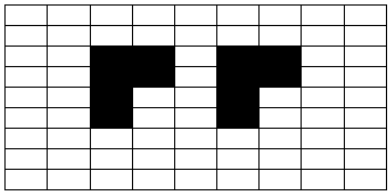
254	252	249	242	244	233	235
246	244	233	194	149	43	111
214	149	43	70	157	43	127
214	157	43	87	191	111	255
215	191	111	223	191	127	255
223	191	127	255	255	255	255
255	255	255	255	255	255	255

MCT8 para a Imagem MCT8

252	248	240	242	228
244	244	233	194	149
214	149	43	70	29
150	157	43	87	47
151	47	79	31	63

Figura 17 – Figura com padrões que se repetem com deslocamento (esq.). Resultado da aplicação do MCT8 na figura original (meio). Resultado da aplicação do MCT8 na figura do meio (dir.).

Ao se comparar os resultados da aplicação do MCT8 nas Figuras 16 e 17, encontra-se diversos valores em comum e o mesmo ocorre quando se compara os valores obtidos para a imagem MCT8, conforme ilustrado na Figura 18. Isso acontece porque as duas figuras apresentam várias estruturas locais semelhantes e essas estruturas estão em regiões que também são similares.

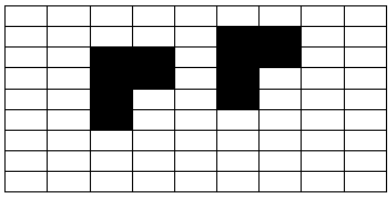


MCT8

254	252	249	250	252	249	251
246	244	233	226	244	233	235
214	149	43	70	149	43	111
214	157	43	86	157	43	127
215	191	111	215	191	111	255
223	191	127	223	191	127	255
255	255	255	255	255	255	255

MCT8 para a Imagem MCT8

252	248	248	248	248
244	244	233	226	228
212	149	43	70	149
150	157	43	86	29
151	47	71	151	47



MCT8

254	252	249	242	244	233	235
246	244	233	194	149	43	111
214	149	43	70	157	43	127
214	157	43	87	191	111	255
215	191	111	223	191	127	255
223	191	127	255	255	255	255
255	255	255	255	255	255	255

MCT8 para a Imagem MCT8

252	248	240	242	228
244	244	233	194	149
214	149	43	70	29
150	157	43	87	47
151	47	79	31	63

Figura 18 – Comparação dos resultados da aplicação do MCT8 e do MCT8 da imagem MCT8 nas figuras anteriores.

Assim, o histograma de valores MCT8 da imagem original indica a ocorrência das estruturas da imagem e contabiliza quantas são similares. Já a informação de contexto ressalta as diferenças entre as estruturas existentes. Como a representação final da imagem é um histograma, onde não há informação do arranjo espacial, a presença da informação contextual, advinda da aplicação do MCT8 na imagem MCT8, dá pistas relativas a esse

arranjo.

Na Figura 19, há duas imagens da classe "Costa" e uma da classe "Montanha". As duas primeiras imagens exibem uma faixa de areia, enquanto a imagem exibida em (c), uma faixa de neve. Areia e neve são conceitos diferentes, mas têm aparências bastante similares. Para verificar se as diferenças entre as imagens são ressaltadas através da inclusão do contexto, foi feito o cálculo da distância euclidiana entre os vetores que representam as imagens. A Tabela 1 exibe os resultados tanto para os histogramas MCT8 quanto para os histogramas CMCT. Quando se compara a distância entre as duas imagens da classe "Costa", a inclusão do contexto eleva a distância entre os vetores que as representam em, aproximadamente, 10%, enquanto, quando se compara os vetores de cada imagem da classe "Costa" com os vetores que representam a imagem da classe "Montanha", têm-se um aumento de 27%, com imagem exibida em (a), e 29%, com a imagem exibida em (b). Verificando-se, portanto, que a inclusão do contexto ajudou a ressaltar as diferenças entre elementos de classes distintas.

(a) (b) (c)

Figura 19 – (a) e (b) Cenas da categoria "Costa" (c) Cena da categoria "Montanha" . Imagens retiradas de Lazebnik, Schmid e Ponce (2006).

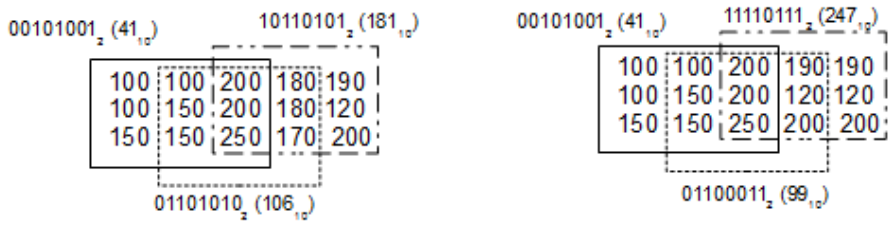
Tabela 1 – Distância euclidiana entre os descritores MCT8 e entre os descritores CMCT para as imagens exibidas na Figura 19.

Figuras	MCT8	CMCT
(a) e (b)	0,81	0,89
(a) e (c)	0,74	0,94
(b) e (c)	0,79	1,02

2.4.2 A Redundância

É importante notar que os valores MCT8 obtidos a partir da imagem original também contêm informações sobre a vizinhança, pois, dizer que duas estruturas locais semelhantes estão embutidas em vizinhanças diferentes, é o mesmo que dizer que as

estruturas locais dos vizinhos são diferentes, logo os valores MCT8 para esses vizinhos também serão distintos. A Figura 20 ilustra essa situação. Nela, é possível notar que, apesar dos pixeis da primeira janela (linha sólida) serem iguais nos dois exemplos, os valores nas janelas seguintes (linha pontilha e traço e ponto) são diferentes. Essa diferença é refletida nos valores MCT8 gerados.



(a) a (b)

Figura 20 – Estruturas locais semelhantes com vizinhos diferentes e os seus respectivos valores MCT8. Em (a), os valores MCT8 são 41, 106 e 181 e em (b), 41, 99 e 247

Como o MCT8 calculado sobre a imagem MCT8 apresenta novamente informações sobre a vizinhança, há uma redundância de informação. No entanto, neste segundo passo, as informações sobre a vizinhança são associadas a um contexto, onde é sinalizada a posição de cada estrutura dentro da vizinhança, fato que não ocorre no primeiro passo, já que a saída do mesmo é um histograma. Assim, os valores MCT8 obtidos a partir da imagem MCT8 são índices para um dos subconjuntos possíveis de arranjos de estruturas locais.

2.4.3 Resultados do CMCT na classificação

A aplicação do MCT8 na imagem MCT8 gera valores diferentes para estruturas locais similares que estão em regiões diferentes. Nesta seção, é verificado se essa informação contribui para o aprimoramento do descritor das imagens, melhorando, portanto, os resultados de suas classificações. Para isso, são analisadas as classificações realizadas através do classificador K-NN (*k-nearest neighbor*) .

O K-NN é um algoritmo de classificação supervisionada (DUDA; HART, 2001). Nele, as distâncias entre o elemento a ser classificado e as amostras do conjunto de treinamento são calculadas e os K vizinhos mais próximos são identificados. O elemento é, então, classificado na classe dos vizinhos mais próximos, o que é decidido através da maioria dos votos. O K-NN realiza a classificação medindo a similaridade entre o elemento que se deseja classificar e todos os elementos de treino. Assim, como elementos classificados em uma mesma classe apresentam uma representação similar, a representação dos dados que mais aproxima elementos de uma mesma classe e que distancia elementos de classes

distintas obtém uma taxa de classificação melhor. Em geral, o K-NN é um procedimento sub-ótimo uma vez que seu uso leva a uma taxa de erro maior que o mínimo possível, na hipótese de se ter um classificador bayesiano perfeitamente parametrizado. Entretanto, com um número ilimitado de amostras e K=1, como resultado assintótico, a probabilidade de erro do vizinho mais próximo é limitada por duas vezes a probabilidade de erro de Bayes (COVER; HART, 1967). Além disso, o classificador K-NN lida com a natureza multiclasse das cenas sem muito esforço (ZHANG et al., 2006).

A medida do desempenho de classificação é acurácia, a, estimada da seguinte forma

$$a = \frac{v_p}{f_p + v_p,} \tag{2.4}$$

onde v_p é o número de verdadeiros positivos, f_p é o número de falsos positivos.

Nos experimentos, cada categoria na base de dados é dividida de forma aleatória em um conjunto de dados de treino e outro de teste. Essa divisão é repetida 5 vezes, e, então, a média das acurácias, μ_a, e o desvio padrão, σ, são obtidos conforme adotado por Wu e Rehg (2011). O cálculo do desvio padrão é apresentado na Equação 2.5.

$$\sigma = \sqrt{\frac{1}{(N-1)} \sum_{i=1}^{N} (a_i - \mu_a)^2} \tag{2.5}$$

onde $\mu_a = \frac{\sum_{i=1}^{N} a_i}{N}$, N é o número de testes realizados e a_i é a acurácia obtida no i-ésimo teste.

Aqui, são feitas comparações entres os resultados de classificação utilizando-se os descritores MCT8, CMCT e MCT original na base de dados de 15 cenas (LAZEBNIK; SCHMID; PONCE, 2006), que contém 4.486 imagens em tons de cinza, divididas em 15 categorias de cenas: costa (360 imagens), floresta (328 imagens), montanha (374 imagens), campo aberto (410 imagens), autoestrada (260 imagens), dentro da cidade (308 imagens), edifício alto (356 imagens), rua (292 imagens), quarto (216 imagens), cozinha (210 imagens), sala de estar (289 imagens), escritório (215 imagens), subúrbio (241 imagens), indústria (311 imagens) e loja (315 imagens). A distância utilizada na comparação é a euclidiana. Todos os vetores foram normalizados antes da classificação e os valores apresentados são obtidos através da média dos resultados alcançados em 5 experimentos diferentes onde 100 imagens, selecionadas de forma aleatórias, de cada classe foram utilizadas para treino e as demais, para teste.

A Tabela 2 apresenta os resultados. Nela, é possível constatar que o melhor resultado de classificação obtido foi com o descritor CMCT e com o K igual a 5. Com esses resultados, pode-se notar que a inclusão da informação de contexto aprimora a capacidade de representação do vetor de características da imagem. A Tabela 3 apresenta a acurácia

de classificação por classe para os descritores MCT8 e para o CMCT em uma execução do experimento. Nela, é possível notar que o CMCT apresenta melhor desempenho em aproximadamente 73% das classes. É interessante ressaltar o aumento do desempenho da classificação, quando o descritor CMCT é utilizado, dos elementos das classes "Quarto" e "Sala", que geram bastante confusão.

Tabela 2 – Acurácia % na classificação utilizando-se o algoritmo K-NN e os descritores MCT8, CMCT e MCT. Em negrito, o melhor desempenho.

K	MCT8	CMCT	MCT
5	59,89 ± 0,80	**62,39 ± 1,18**	58,88 ±0,70
6	60,30 ± 0,74	62,29 ± 1,1	59,00 ± 0,90
7	59,86 ± 0,86	62,15 ± 0,88	58,77 ± 0,98

Tabela 3 – Acurácia % na classificação por classe utilizando-se o algoritmo K-NN e os descritores MCT8 e CMCT. Em negrito, os melhores desempenhos.

Classe	MCT8	CMCT
Quarto	33,62	**38,79**
Sala	39,68	**52,91**
Subúrbio	**97,87**	96,45
Indústria	38,39	**40,28**
Cozinha	55,45	**58,18**
Costa	**63,08**	62,31
Floresta	**91,23**	89,91
Rodovia	72,50	**75,62**
Dentro da cidade	65,38	**65,87**
Montanha	60,58	**63,87**
Campo	48,71	**50,65**
Rua	70,83	**75**
Prédio	35,94	**40,23**
Escritório	**92,17**	89,57
Loja	57,21	**59,07**

2.5 Considerações Finais

Neste capítulo, foi apresentado o CMCT, uma evolução do CENTRIST. O CENTRIST identifica as estruturas locais existentes na imagem e gera, como saída, um histograma, que é uma representação desordenada dessas estruturas. Apesar das estruturas estabelecerem relações entre pixeis vizinhos, essa informação é local e não dá uma visão da imagem como um todo. A introdução do contexto feita pelo CMCT ajuda a ampliar o escopo dessas relações, dando um indicativo maior da organização das estruturas encontradas na imagem, fornecendo, assim, mais subsídios para a classificação sem a necessidade de parâmetros adicionais.

No próximo capítulo, serão apresentadas técnicas que exploram as informações fornecidas pelo CMCT visando melhorar os resultados na classificação de cenas.

3 *Gist*CMCT

Neste capítulo, é proposta uma abordagem para melhorar a eficiência do CMCT na representação de imagens para a classificação de cenas. Assim, primeiro, é apresentado um modelo computacional para o reconhecimento de cenas do mundo real que ignora a segmentação e o processamento individual de objetos e regiões e tem como objetivo representar o *gist* da cena, ou seja, as informações perceptual e semântica que podem ser capturadas com apenas um olhar para uma cena complexa do mundo real independentemente da desordem ou da variedade de detalhes, e, depois, um novo descritor é proposto: o *Gist*CMCT.

3.1 O *Gist* da Cena

Considerando a hipótese de que não é necessário determinar os objetos que constituem uma cena para identificá-la, Oliva e Torralba (2001) propuseram uma abordagem holística cujo objetivo é satisfazer os requerimentos principais do *gist* - transmitir um sumário estrutural que é significativo o suficiente para permitir a identificação da imagem - o envelope espacial.

Essa abordagem leva em consideração o fato de que uma cena é um arranjo em um espaço tridimensional e a representa através de propriedades que funcionam como descritores do espaço que ela ocupa, as dimensões perceptuais (naturalidade, abertura, expansão, aspereza, irregularidade). Os autores observaram que as dimensões perceptuais capturam a maioria das estruturas tridimensionais das cenas do mundo real. Além disso, cenas que compartilham a mesma categoria semântica tendem a se aglomerar dentro da mesma região de um espaço multidimensional em que os eixos são as propriedades perceptuais, ou seja, possuem dimensões perceptuais similares. O conjunto de dimensões espaciais perceptuais forma o envelope espacial de uma cena.

O envelope espacial constrói uma representação significativa do *gist* da cena diretamente a partir de um conjunto de características de baixo nível, sem ligar os contornos para formar superfícies, e superfícies para formar objetos.

Cada propriedade do envelope espacial pode ser estimada a partir de uma coleção de *templates* de características globais medindo o quanto natural ou aberta (ou seja, se há a presença de uma linha no horizonte) é uma cena (OLIVA; TORRALBA, 2006). Assim, uma cena pode ser representada por uma combinação de características globais. Uma opção de representação é a combinação ponderada da saída de bancos de filtros multiescalares e multiorientados, como os filtros de Gabor, por exemplo. Devido à alta dimensão das

imagens, a combinação de todas as saídas de todos os filtros se torna inviável, sendo preciso, portanto, aplicar uma técnica de redução de dimensão. Uma abordagem possível é amostragem da saída de cada filtro em um tamanho N x N, sendo que todas as saídas são amostradas para o mesmo tamanho, independente da escala do filtro. Como resultado, cada imagem é representada por um vetor do tamanho N x N x K, onde K é o número de filtros.

A Figura 21 apresenta a média das magnitudes de saídas de um filtro multiescalar e multiorientado em um diagrama polar para imagens das classes "Quarto" e "Floresta". Nela, é possível verificar uma grande diferença entre as saídas dos descritores das imagens da classe "Quarto" e a da imagem da classe "Floresta". Para o cálculo dos descritores, foram utilizados 4 escalas e oito orientações distintas, além de uma amostragem de 4 x 4, gerando 16 saídas e um descritor de 512 dimensões.

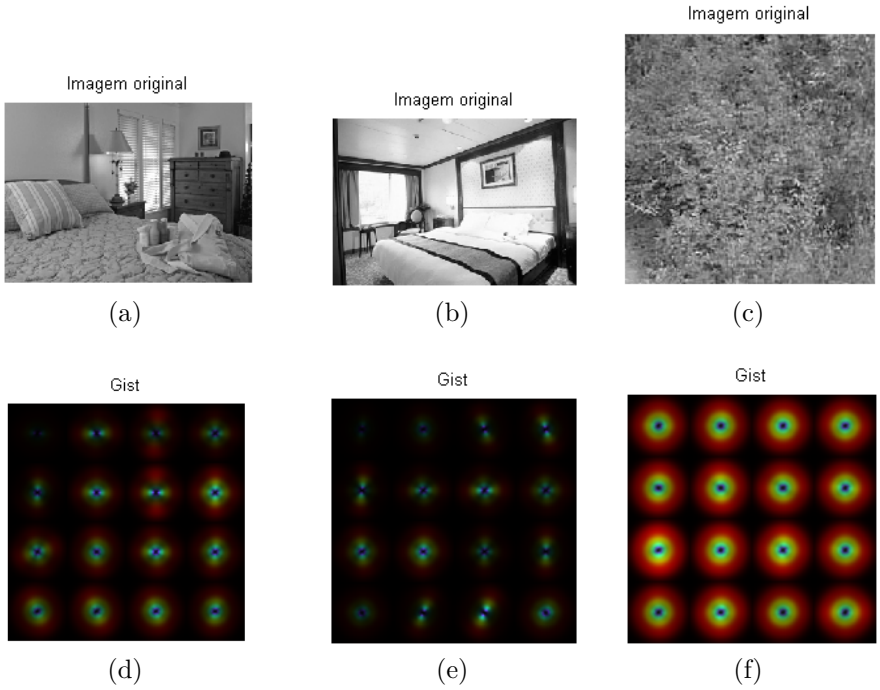


Figura 21 – Em (d), (e) e (f), observa-se o valor médio das magnitudes obtidas nas saídas dos filtros de Gabor para entradas a partir de (a), (b) e (c), respectivamente.

Uma vez que a abordagem do envelope espacial gera uma representação holística de baixa dimensionalidade da estrutura da cena e que não requer uma segmentação explícita da imagem e dos objetos, ela solicita baixos recursos computacionais (OLIVA; TORRALBA, 2006). Experimentos realizados em Wu e Rehg (2011) mostraram que o *gist* definido nesta abordagem tem bom desempenho, comparado com outras técnicas da literatura, quando há a classificação de cenas externas, mas apresenta um resultado um pouco pior quando cenas de ambientes internos são adicionadas.

3.2 *Gist*CMCT

Apesar do CMCT, apresentado no capítulo anterior, e da abordagem da obtenção do *gist* através do envelope espacial serem ambas holísticas, elas diferem entre si. Enquanto o envelope espacial representa a forma da cena através de estruturas espaciais estáveis dentro da imagem que refletem a funcionalidade de sua locação, CMCT sumariza a informação das formas locais.

A abordagem de obtenção do *gist* através do envelope espacial tem certas limitações em reconhecimento de cenas em ambientes fechados, mas é bastante eficiente no reconhecimento de cenas em ambientes abertos. O CMCT, como o CENTRIST, representa as propriedades através da distribuição de estruturas locais (por exemplo, percentual de estruturas que são linhas horizontais) (WU; REHG, 2011) o que ajuda na classificação de ambientes feitos pelo homem, incluindo ambientes internos. Assim, um vetor composto pelo descritores *gist*, obtido através do envelope espacial, e o CMCT irá reunir as qualidades de ambos e poderá obter um melhor desempenho na classificação de cenas.

Desse modo, é proposto o *Gist*CMCT uma abordagem holística que representa uma cena através de um vetor de características que é a composição do vetor gerado pelo CMCT com o vetor gerado pela abordagem do envelope espacial, o *gist*. A Figura 22 esquematiza o processo de obtenção do *Gist*CMCT. Nela é possível observar os seguintes passos na composição do descritor: obtenção do vetor de características *gist* através da aplicação de um banco de filtros de Gabor, obtenção do histograma dos valores MCT8 da imagem original, criação da imagem MCT8, obtenção do histograma de valores MCT8 da imagem MCT8 e, por fim, concatenação dos 3 vetores em um único vetor. Nenhum esquema de ponderação é utilizado na concatenação dos vetores.

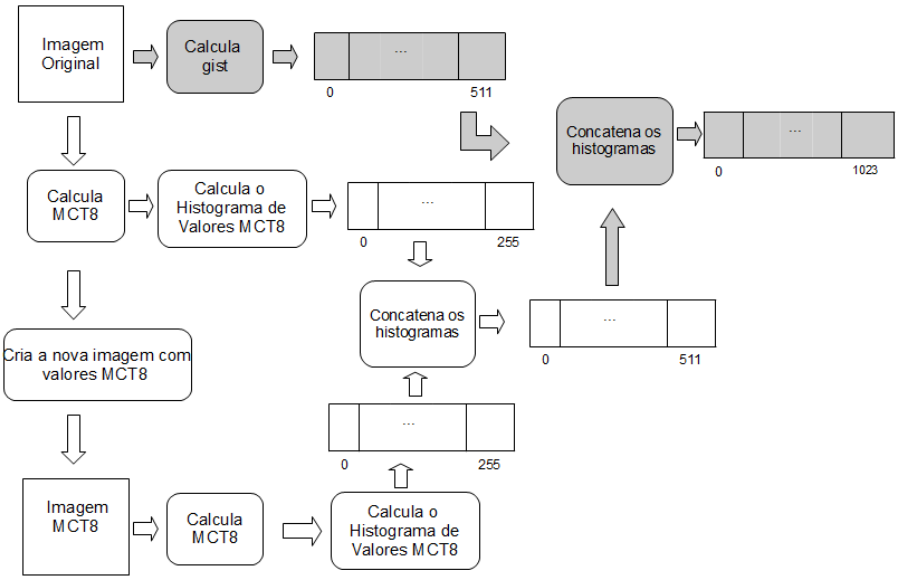


Figura 22 – Esquema para a obtenção do *Gist*CMCT.

Neste trabalho, os parâmetros empregados foram os mesmos apresentados na seção 3.1: filtros com 4 escalas e 8 orientações distintas e uma amostragem da imagem de 4 x 4, gerando um descritor *gist* de 512 dimensões. Como o CMCT também apresenta 512 dimensões, o descritor *Gist*CMCT apresenta 1024 dimensões.

3.2.1 Resultados do *Gist*CMCT na classificação

A Tabela 4 apresenta os resultados da classificação das imagens da base de dados de 15 categorias (LAZEBNIK; SCHMID; PONCE, 2006), utilizando-se o K-NN. A distância utilizada na comparação é a euclidiana. Todos os vetores foram normalizados antes da classificação e os valores apresentados são a média das acurácias (Equação 2.4) e o desvio padrão (Equação 2.5) obtidos em 5 experimentos diferentes onde 100 imagens, selecionadas de forma aleatórias, de cada classe foram utilizadas para treino e as demais, para teste.

É possível notar que a união dos diferentes descritores contribui para uma melhor separação das classes, já que o desempenho da classificação aumenta em aproximadamente 14%.

Tabela 4 – Acurácia % na classificação utilizando-se o algoritmo K-NN e os descritores CMCT, *gist* e *Gist*CMCT. Em negrito, o melhor desempenho.

K	CMCT	*gist*	*Gist*CMCT
5	62,39 ± 1,18	61,03 ±0,43	70,55 ±0,30
6	62,29 ± 1,1	62,39 ± 0,57	**71,06 ±0,44**
7	62,15 ± 0,88	62,45 ± 0,81	70,79 ±0,43

A Tabela 5 apresenta os resultados da classificação por classe para uma execução do K-NN na base de dados de 15 cenas para os descritores CMCT, *gist* e *Gist*CMCT. Há classes onde o desempenho do CMCT é superior ao do *gist*, como por exemplo na classe "Escritório", onde a taxa de acerto do CMCT é maior do que o dobro da taxa de acerto do *gist*, e classes onde o desempenho do *gist* é superior, como no caso da classe "Campo" onde a taxa de acerto do *gist* é aproximadamente 40% maior do que a taxa de acerto do CMCT. Quando se considera os resultados do *Gist*CMCT, percebe-se que em 80% dos casos o desempenho de classificação do *Gist*CMCT é superior aos resultados do *gist* e do CMCT, inclusive nos casos onde se pressupõe que a semelhança entre as imagens é grande, como no caso das classes "Sala" e "Quarto".

Desse modo, nota-se que a união desses dois descritores contribui para uma melhor descrição das imagens, permitindo uma classificação mais acurada das mesmas.

Tabela 5 – Acurácia % na classificação por classe utilizando-se o algoritmo K-NN e os descritores CMCT, *gist* e *Gist*CMCT. Em negrito, os melhores desempenhos.

Classe	CMCT	*gist*	*Gist*CMCT
Quarto	38,79	47,41	**53,45**
Sala	52,91	55,56	**65,08**
Subúrbio	96,45	91,49	**98,58**
Industria	40,28	29,38	**52,61**
Cozinha	58,18	36,36	**65,45**
Costa	62,31	**65,00**	62,31
Floresta	89,91	94,30	**94,74**
Rodovia	75,62	**79,37**	78,75
Dentro da cidade	65,87	56,25	**69,23**
Montanha	**63,87**	50,36	57,30
Campo	50,65	70,97	**76,77**
Rua	75	72,92	**77,60**
Prédio	40,23	57,03	**63,28**
Escritório	**89,57**	43,48	**89,57**
Loja	59,07	63,72	**68,84**

3.3 *Gist*CMCT e MCT Espacial Combinados

Nesta seção, é apresentado um novo método de classificação, o *Gist*CMCT e MCT Espacial combinados (*Gist*CMCT-SM), que emprega a combinação de classificadores. São utilizados dois classificadores baseados em características distintas: *Gist*CMCT e *spatial* MCT e os resultados gerados são utilizados na obtenção da classe final.

3.3.1 O MCT Espacial

Lazebnik, Schmid e Ponce (2006) mostraram que o arranjo espacial de características numa imagem fornece pistas poderosas para a tarefa de classificação de cenas.

Para explorar tais pistas, Wu e Rehg (2011) propuseram uma representação espacial para o CENTRIST que possibilita a captura das estruturas presentes em uma escala maior. Tal representação é baseada no esquema SPM (LAZEBNIK; SCHMID; PONCE, 2006). A representação espacial proposta consiste em dividir uma imagem em sub-regiões, aplicar o CENTRIST nestas regiões e reunir os resultados em um único vetor, representando, assim, a estrutura global aproximada da imagem, o que, normalmente, eleva os resultados de classificação.

A Figura 23 apresenta uma representação espacial em pirâmide com três níveis. Em cada nível, a imagem é dividida em $2^n \times 2^n$ blocos, onde n é o nível em questão. Ainda, é feito um deslocamento das divisões da imagem (retratado pela linha pontilhada), adicionando à representação $2^{n-1} \times 2^{n-1}$ blocos. Tal deslocamento é realizado para evitar artefatos criados pelas divisões sem interseções. Portanto, para cada nível $n > 0$, a imagem

é dividida em $2^{2n} + 2^{2(n-1)}$ blocos. Além disso, a imagem tem seu tamanho alterado em cada nível. Desse modo, todos os blocos contêm o mesmo número de pixeis. O CENTRIST de cada bloco é calculado e os histogramas resultantes são concatenados. Por fim, o PCA é aplicado para reduzir as dimensões do vetor final, que é chamado de *spatial Principal component Analysis of Census Transform* (*spatial* PACT ou sPACT).

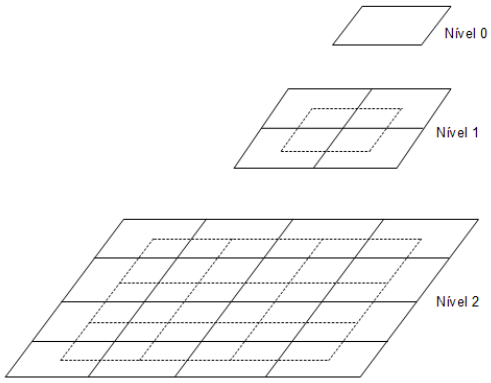


Figura 23 – Ilustração da divisão de uma imagem para a representação espacial.

A representação proposta em Wu e Rehg (2011) é adotada neste trabalho. Todavia, ao invés do descritor CENTRIST, é utilizado o MCT (o original com 9 bits) (FROBA; ERNST, 2004), já que este último diferencia estruturas que são consideradas iguais pelo CENTRIST. Os vetores MCT obtidos em todos os blocos são concatenados para formar um único vetor de características. Depois, é aplicada a abordagem análise de componentes principais (PCA) para reduzir a sua dimensionalidade, como foi feito em Wu e Rehg (2011). Três níveis de informação espacial são adotados, o que gera 31 blocos, e o vetor de cada bloco tem a sua dimensionalidade reduzida de 512 para 40. As informações extras (média e desvio padrão dos blocos de pixeis) também são utilizadas. A pirâmide espacial de MCT é chamada aqui de MCT Espacial.

3.3.2 Combinando Classificadores

Como visto na seção 3.2, o *Gist*CMCT combina o vetor de características do *gist* com o CMCT em um novo vetor com o objetivo de melhorar os resultados da classificação tanto para cenas internas quanto para cenas externas. Contudo, o *Gist*CMCT não considera a informação do *layout* espacial das características na imagem. O MCT Espacial, por sua vez, se beneficia da informação espacial, o que fornece informações adicionais que melhoram os resultados da classificação.

De acordo com Ho, Hull e Srihari (1994), a acurácia na classificação pode ser melhorada utilizando-se características e classificadores de diferentes tipos simultaneamente, através de classificadores múltiplos. Foi observado que classificadores e características de tipos diferentes se complementam nos resultados da classificação, ou seja, os padrões

classificados de forma incorreta por classificadores diferentes não são necessariamente os mesmos (KITTLER et al., 1998).

Assim, com o objetivo de tirar vantagem das qualidades do *Gist*CMCT e do MCT Espacial sem aumentar o tamanho final do vetor de características, a estratégia de combinação de classificadores é adotada. Dessa forma, cada uma dessas características treina um classificador diferente e os resultados por eles gerados são usados para a tomada de decisão através da combinação de suas opiniões individuais para se chegar a uma decisão consensual.

O classificador adotado para os dois descritores é o SVM, um padrão de classificação introduzido por Vapnik (1998). Apesar do emprego do mesmo tipo de classificador, as características fornecidas para cada um são distintas, uma vez que cada classificador utiliza a sua própria representação da imagem (*Gist*CMCT ou MCT Espacial). Dessa forma, os diferentes tipos de características são integrados fisicamente.

A implementação de sistemas de classificadores múltiplos implica na definição de uma regra de combinação para determinar a classe mais provável, baseando-se na classe atribuída por cada classificador individualmente (FOGGIA et al., 2007). Para a combinação das opiniões individuais de cada classificador, as seguintes regras de combinação foram adotadas: Máximo, Mediana e Produto, apresentadas em (KITTLER et al., 1998). Por simplicidade, foi assumido que os resultados de cada classificador são estatisticamente independentes. Para uma distribuição de classes equiprovável, a classe ω selecionada é a classe que satisfaz as seguintes equações:

Máximo

$$max_{i=1}^{R} P(\omega_j|x_i) = max_{k=1}^{m} max_{i=1}^{R} P(\omega_k|x_i), \tag{3.1}$$

Mediana

$$med_{i=1}^{R} P(\omega_j|x_i) = max_{k=1}^{m} med_{i=1}^{R} P(\omega_k|x_i), \tag{3.2}$$

Produto

$$prod_{i=1}^{R} P(\omega_j|x_i) = max_{k=1}^{m} prod_{i=1}^{R} P(\omega_k|x_i), \tag{3.3}$$

onde m é o número de classe4as ($\omega_1, ..., \omega_m$), x_i é o vetor de características utilizado pelo i-ésimo classificador, $P(a|b)$ é a probabilidade de b dado a e R é o número de classificadores, neste caso, $R = 2$.

3.4 Considerações Finais

Na classificação de cenas, é importante encontrar uma representação da imagem que seja capaz de aproximar exemplos que pertençam a uma mesma classe, enquanto mantém distantes exemplos de classes distintas. Com o intuito de encontrar uma forma

de representação da imagem mais eficiente, foi feita a combinação de dois descritores que representam diferentes tipos de informação da imagem, o *gist* e o CMCT. A intenção, nesse caso, é reunir as qualidades dos dois descritores e representar melhor tanto imagens de ambientes internos, como de ambientes externos.

Já com o objetivo de melhorar os resultados gerados pelos classificadores, uma segunda abordagem foi apresentada, onde optou-se pela combinação dos classificadores dos descritores *Gist*CMCT e MCT Espacial, pois o consenso entre as opiniões desses dois classificadores pode levar à classificação correta dos casos onde o erro de classificação é cometido por apenas um deles. Uma alternativa a essa abordagem seria a combinação desses dois descritores em um único vetor e a utilização de um único classificador. Porém, essa estratégia acarretaria em um aumento da dimensão do vetor final, podendo aumentar a complexidade da etapa de classificação, devido a problemas como a "maldição da dimensionalidade" (BAGGENSTOSS, 2004), que ocorre quando o aumento da dimensionalidade do espaço de características torna difícil a tarefa de encontrar os melhores limites entre as classes para um conjunto fixo de dados de treinamento.

No próximo capítulo, será apresentada uma estratégia que aumenta a quantidade de informação acerca do contexto na representação da imagem através da inclusão de informação no CMCT advinda de vizinhos não muito próximos das estruturas locais. Além disso, será proposta uma abordagem que introduz informação sobre o arranjo espacial nesse mesmo descritor.

4 CMCT Estendido

Neste capítulo, é proposto o descritor CMCT Estendido (ECMCT - *Extended* CMCT), uma evolução do CMCT, que melhora a sua capacidade de representação através da inclusão de dois tipos de informação: contextual, adicionando conhecimento acerca de pixeis próximos à janela de cálculo da transformada não-paramétrica, mas que não são considerados diretamente na obtenção do CMCT; e sobre o arranjo espacial, adicionando dados estatísticos obtidos em sub-regiões da imagem. Além disso, abordagens combinando o descritor ECMCT com o descritor *gist* e com o MCT Espacial também são propostas.

Primeiro, no entanto, são apresentadas duas novas abordagens: Transformada Census dos Vizinhos Distantes (CTDN - *Census Transform of Distant Neighbors*) e Vetor Estatístico, que são utilizadas na composição do descritor proposto.

4.1 Vizinhos Distantes

Conforme exposto no Capítulo 2, a inserção na descrição da imagem de dados sobre as estruturas vizinhas das estruturas locais, identificadas pelo MCT8, aprimora a representação da imagem e, consequentemente, leva a melhores resultados de classificação. Esse aperfeiçoamento é causado pela possibilidade de diferenciar estruturas similares que estão inseridas em regiões distintas. A partir deste conhecimento, são propostas modificações no descritor CMCT com a intenção de identificar, de forma mais apurada, as diferenças entre as regiões em tornos das estruturas locais identificadas. Assim, com o objetivo de ampliar a região de onde a informação local é extraída, é considerada a informação sobre os pixeis próximos, mas não limítrofes das estruturas locais, pixeis esses que não são considerados diretamente durante o cálculo do CMCT.

A abordagem proposta cria uma nova estrutura formada pelos pixeis presentes na janela 3 x 3 durante o cálculo do MCT8 e por pixeis que estão posicionados em regiões que estão próximas, mas que não fazem fronteira com a referida janela. Uma maneira direta de inserir esse tipo de informação é aumentar o tamanho da janela no cálculo do MCT8, mas essa abordagem também aumenta o tamanho do vetor final de forma sensível. Se, por exemplo, uma janela 5 x 5 pixeis é considerada, o tamanho do histograma de valores MCT8 será de 2^{24} posições. Dessa forma, com o intuito de evitar o aumento excessivo do vetor de características, a estratégia proposta considera apenas 8 pixeis, posicionados a um distância k a partir do pixel central da janela. A Figura 24 ilustra os pixeis considerados nesta abordagem, exibidos em cinza, para $k = 4$. Além dos oitos pixeis, é utilizada, também, a média dos pixeis da janela 3 x 3, $\overline{I}_w(x, y)$. Assim sendo, os oito pixeis são comparados com a média, $\overline{I}(x, y)$, calculada a partir dos valores do

pixeis e de $\overline{I}_w(x, y)$. O valor da Transformada Census dos Vizinhos Distantes (CTDN - *Census Transform of Distant Neighbors*) a uma distância k do pixel na posição (x, y), $CTDN_k(x, y)$ é calculado da seguinte forma:

$$CTDN_k(x,y) = \bigotimes_{p=0}^{p=7} \zeta(N_p, \overline{I}(x,y)), \quad \zeta(m,n) = \begin{cases} 1, & m \geq n \\ 0, & m < n \end{cases} \tag{4.1}$$

onde $\bigotimes$ representa a concatenação, N_p é o valor do pixel vizinho e assume um dos seguintes valores: $I(x-k, y-k)$, $I(x-k, y)$, $I(x-k, y+k)$, $I(x, y-k)$, $I(x, y+k)$, $I(x+k, y-k)$, $I(x+k, y)$, $I(x+k, y+k)$. Já, $\overline{I}(x, y) = (\sum_0^7 N_p + \overline{I}_w(x, y))/9$, sendo $\overline{I}_w(x, y)$ a média dos valores dos pixeis na janela 3 x 3 com pixel central na posição (x, y). Os oito pixeis gerados a partir das comparações são convertidos em um número na base decimal entre 0 e 255. Um histograma dos valores CTDN é obtido para representar a imagem.

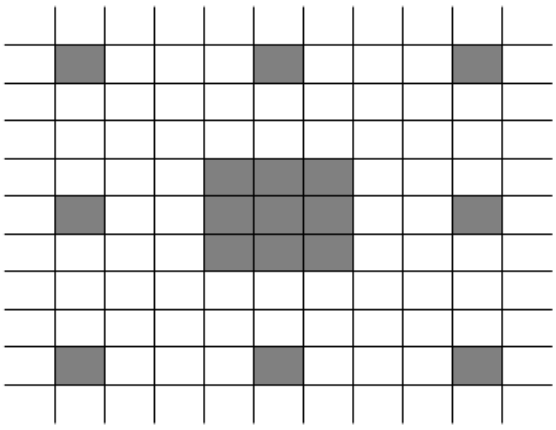


Figura 24 – Vizinhos distantes considerados pelo CMCT Estendido (GAZOLLI; SALLES, 2014).

O valor de k indica que vizinhança será considerada a partir do pixel central. Se k é muito pequeno, a nova informação poderá ser redundante, porque os pixeis muito próximos já são associados através do cálculo do MCT8. Se, ao contrário, o k for muito grande, será estabelecida uma relação entre pixeis muito distantes e a informação deixará de ser local. Assim, o melhor valor para k, para uma janela n x n no cálculo do MCT8, é n + 1. Os pixeis a essa distância são os que estão mais próximos do pixel central sem estar associados ao mesmo através do cálculo do MCT8, como ilustrado na Figura 25. Como, nesse trabalho, são utilizadas janelas 3 x 3 no cálculo do MCT8, o valor adotado para k é 4.

4.2 Informação Espacial

Um outro tipo de informação adicionada ao CMCT pelo novo descritor é o contraste. Visto que o MCT8 é um bom descritor de estruturas locais, mas não leva em conta

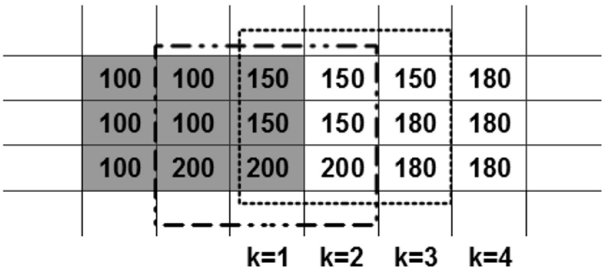


Figura 25 – Vizinhos, para k=4, não considerados diretamente pelo MCT8 com janelas de tamanho 3 x 3 (GAZOLLI; SALLES, 2014).

a informação de contraste dos pixeis dentro da janela 3 x 3. Essa informação indica a força da estrutura representada na janela e pode ajudar a diferenciar um grande número de estruturas. É possível medir o contraste, VAR_P, da seguinte maneira (OJALA; PIETIKAINEN; MAENPAA, 2002):

$$\mathrm{VAR}_P = \frac{1}{P}\sum_{p=0}^{P-1}(I_p - \mu)^2, \tag{4.2}$$

onde P é o número de pixeis na janela, $\mu = 1/P\sum_{p=0}^{P-1} I_p$ e I_p o nível de intensidade do p-ésimo pixel da imagem. Neste trabalho, $\mathrm{VAR_P}$ é calculado para todos os pixeis na imagem através de janelas de pixeis de tamanho 3 x 3.

Como mencionado no Capítulo 3, a composição espacial das características da imagem permite tirar vantagem das regularidades na composição da imagem e oferece pistas poderosas para a tarefa de classificação de cenas. Sendo assim, a informação de contraste é associada à informação espacial. Para isso, a imagem é dividida em sub-regiões cada vez menores e, para cada sub-região, a média dos contrastes é calculada. Além disso, a média e a variância dos valores MCT8 também são calculadas. Os valores obtidos em cada sub-região são, então, concatenados em um vetor chamado Vetor Estatístico. Apenas a média dos contrantes (isto é, o valor do contraste é desconsiderado no vetor) é utilizada para evitar o aumento excessivo do vetor final de características.

4.3 CMCT Estendido

Em seu trabalho, Wu e Rehg (2011) buscavam por uma representação que suprimisse uma informação detalhada sobre textura e, assim, propuseram o CENTRIST. Esse descritor considera apenas a informação das janelas de tamanho 3 x 3 durante o seu cálculo, uma porção muito pequena da imagem, o que dificulta a identificação de um padrão de textura mais grosseiro. Além disso, como mencionado no Capítulo 2, a própria transformada census ignora algumas variações de níveis de cinza, o que contribui para a não identificação de alguns detalhes de textura.

Excesso de detalhes de informação sobre textura pode atrapalhar o reconhecimento de cenas. Todavia, conforme apresentado no Capítulo 1, a representação de cenas sofre de problemas, tais como, alta variabilidade intraclasse e baixa variabilidade interclasse, e alguma informação sobre textura pode ajudar a definir de forma mais precisa os elementos pertencentes a uma determinada classe. Sobretudo quando se trabalha com cenas compostas por elementos não produzidos pelo ser humano.

O descritor CMCT se difere do CENTRIST por utilizar a informação de contexto, o que implica na consideração de informações de uma região maior do que a janela 3 x 3. Ademais, a adoção do MCT8 faz com que algumas variações de níveis de cinza, antes descartadas pelo CENTRIST, sejam consideradas. Assim, o CMCT considera mais informações de textura do que o CENTRIST.

Um outro tipo de informação descartada tanto pelo CENTRIST quanto pelo CMCT é a informação espacial. Esse tipo de informação ajuda na definição das classes, pois indica o local de ocorrência de uma determinada característica na imagem. Assim, dois conceitos distintos que apresentem características semelhantes podem ser diferenciados através das posições dessas características na imagem. Por exemplo, um conjunto de pixeis azuis no topo da imagem pode indicar o céu, já na parte inferior, mar.

Diante do exposto, é proposto o CMCT Estendido, descritor que amplia a informação de contexto fornecida pelo CMCT, através da representação de vizinhos distantes da janela 3 x 3, o que consequentemente aumenta a área analisada e agrega informação sobre textura. Além disso, esse descritor também provê alguma informação espacial através da divisão subsequente da imagem, sem, no entanto, aumentar em demasiado o tamanho do vetor final de características. Para atingir tal objetivo, essa abordagem realiza a concatenação do CMCT original com o histograma CTDN obtido a partir da imagem original, para ampliar a informação contextual, e com o Vetor Estatístico, para incluir a informação espacial, conforme ilustrado na Figura 26. É possível notar que o CMCT e o histograma CTDN são extraídos apenas globalmente, ou seja, da imagem como um todo, e que o Vetor Estatístico é extraído apenas das sub-regiões, isto é, para esse descritor, a imagem total é desconsiderada. Os resultados são, então, concatenados em um novo vetor.

O tamanho do vetor CMCT é 512 e o histograma CTDN tem 256 posições. Para obter o Vetor Estatístico, são utilizados apenas 2 níveis de divisão, resultando em um vetor de 60 posições. Assim, o tamanho final do vetor de características ECMCT apresenta 828 posições. Vale ressaltar que, do mesmo modo que o CMCT, o ECMCT mantém o atributo que possibilita a sua utilização em sistemas voltados para usuários leigos, ou seja, não há a necessidade de especificar parâmetros pelo usuário.

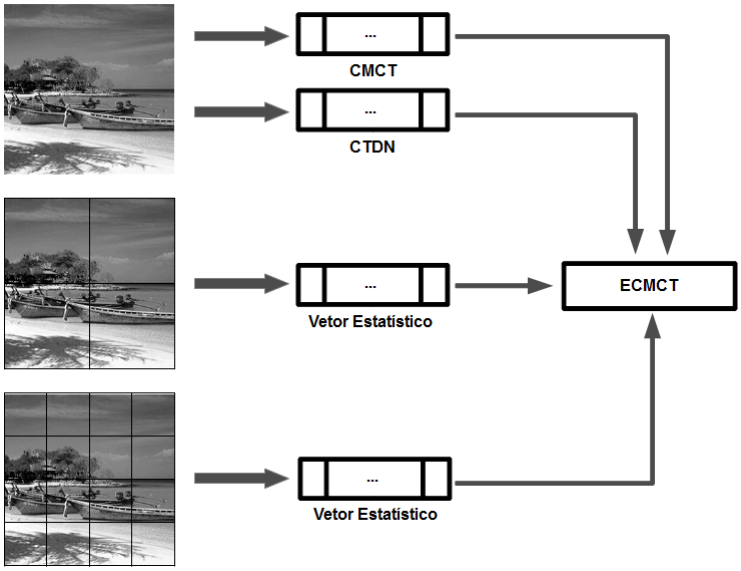


Figura 26 – Processo de obtenção do ECMCT (GAZOLLI; SALLES, 2014).

4.3.1 Resultados do ECMCT na Classificação

Nesta seção, é analisado o desempenho do descritor ECMCT na classificação, com o classificador K-NN, em duas bases de dados: 15 categorias (LAZEBNIK; SCHMID; PONCE, 2006) e base de dados de textura (LAZEBNIK; SCHMID; PONCE, 2005). A distância utilizada na comparação é a euclidiana. Todos os vetores foram normalizados antes da classificação e os valores apresentados são a média das acurácias (Equação 2.4) e o desvio padrão (Equação 2.5) obtidos em 5 experimentos diferentes. Na base de dados de 15 categorias, 100 imagens de cada classe foram selecionadas de forma aleatórias para treino e as demais foram utilizadas para teste. Já para a base de dados de textura, foram utilizadas, para cada classe, 20 amostras para treino e 20 para teste. Assim como na base de dados de 15 cenas, foram executados 5 experimentos, onde os exemplos de treino e teste foram selecionados de forma aleatória.

4.3.2 15 Categorias de Cenas

A Tabela 6 compara os resultados de classificação obtidos pelos descritores CMCT, CMCT concatenado com o histograma CTDN (CMCT + CTDN), CMCT concatenado com o Vetor Estatístico (CMCT + VE) e ECMCT na base de dados de 15 categorias (LAZEBNIK; SCHMID; PONCE, 2006). É possível ver que tanto o CTDN quanto o Vetor Estatístico fornecem informações que melhoram os resultados de classificação do CMCT. Já que o melhor desempenho do CMCT nesta base é 62,39% e, quando ele é associado ao CTDN, esse desempenho sobe para 64,95% e aumenta ainda mais quando associado ao vetor estatístico, chegando a 65,88%. O ECMCT, no entanto, tem o melhor resultado, indicando uma melhor representação da imagem quando há a concatenação dos três vetores.

Tabela 6 – Acurácia % na classificação utilizando-se o algoritmo K-NN na base de dados 15 categorias. Em negrito, o melhor desempenho de cada técnica. Em negrito, o melhor desempenho.

K	CMCT	CMCT + CTDN	CMCT + VE	ECMCT
5	62,39 ± 1,18	64,72 ± 0,73	65,56 ± 1,04	66,95 ±0,0,81
6	62,29 ± 1,1	64,95 ± 0,66	65,78 ± 0,95	**67,09 ± 0,76**
7	62,15 ± 0,88	64,65 ±0,82	65,88 ± 0,58	67,01 ± 0,81

A Tabela 7 apresenta os resultados da classificação por classe para uma execução do K-NN na base de dados de 15 categorias para os descritores CMCT e ECMCT. Nela é possível verificar que o ECMCT alcança melhores resultados de classificação em aproximadamente 73% das classes quando comparado ao CMCT. O ECMCT apresenta o seu pior desempenho na classe "Quarto", quando o resultado de classificação do ECMCT corresponde a 70% do resultado alcançado pelo CMCT. Já o melhor desempenho ocorre na classe "Campo", quando o resultado de classificação obtido pelo ECMCT é aproximadamente 39% maior do que o obtido pelo CMCT.

Tabela 7 – Acurácia % na classificação utilizando-se o algoritmo K-NN por classe na base de dados de 15 categorias. Em negrito, os melhores desempenhos.

Classe	CMCT	ECMCT
Quarto	**38,79**	27,59
Sala	52,91	**53,44**
Subúrbio	96,45	**97,87**
Industria	**40,28**	32,7
Cozinha	**58,18**	55,45
Costa	**62,31**	58,85
Floresta	89,91	**91,67**
Rodovia	75,62	**81,25**
Dentro da cidade	65,87	**73,56**
Montanha	63,87	**66,06**
Campo	50,65	**70,65**
Rua	75	**81,77**
Prédio	40,23	**49,22**
Escritório	89,57	**92,17**
Loja	59,07	**71,16**

4.3.3 Texturas

Comumente, superfícies naturais exibem padrões repetitivos ou variação de intensidade que são geralmente chamados de texturas (OJANSIVU; HEIKKILÄ, 2008). Conforme abordado na Seção 4.3, a inclusão da informação das regiões vizinhas nos descritores CMCT e ECMCT contribuem com informações sobre textura. Assim, nesta seção, é analisado o desempenho dos descritores propostos na classificação deste tipo de imagem, sendo

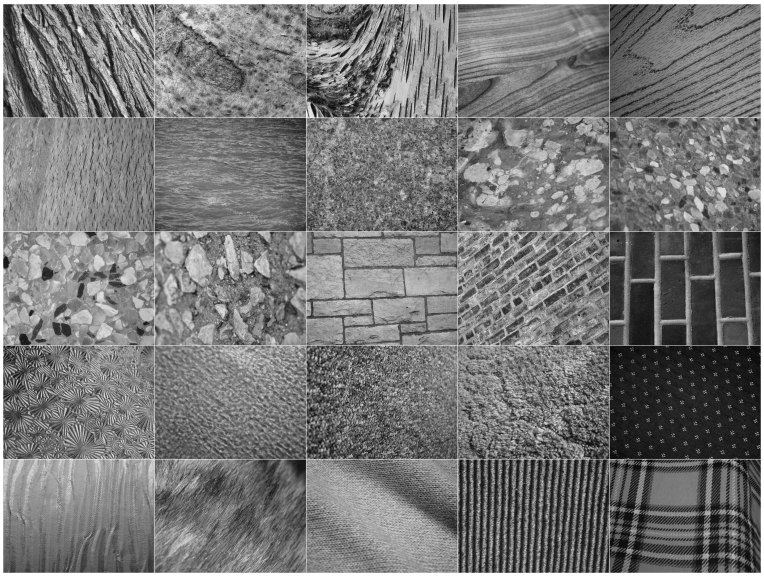


Figura 27 – Uma imagem de cada uma das 25 classe presentes na base de dados de textura.

feita uma comparação dos desempenhos de classificação de texturas representadas pelos descritores CENTRIST, CMCT, CMCT + CTDN, CMCT + VE e ECMCT na base de dados de textura. Essa base contém 25 classes de textura com 40 amostras cada. Todas as imagens estão em escala de cinza e têm tamanho 640 x 480 pixeis. A Figura 27 apresenta alguns exemplos de texturas presentes nesta base de dados.

Os resultados de classificação são apresentados na Tabela 8, onde é possível verificar que o ECMCT apresenta um desempenho aproximadamente 38% maior do que o CENTRIST e 11% maior do que o CMCT. Além disso, é possível constatar que a concataneção do CMCT com o CTDN representa melhor as texturas do que a concatenação do primeiro com o Vetor Estatístico. A concatenação das três técnicas, ou seja, o ECMCT tem um ganho pequeno na classificação quando comparado com o CMCT + CTDN. Assim, pode-se perceber que a inclusão das informações acerca dos vizinhos distantes permite uma melhor discriminação das texturas presentes na imagem, o que se deve a uma melhor representação das estruturas presentes em regiões de pixeis. Já que texturas não podem ser definidas através de um único pixel, mas através de um conjunto desses elementos.

Tabela 8 – Acurácia % na classificação utilizando o algoritmo K-NN na base de dados de textura. Em negrito, o melhor desempenho.

K	CENTRIST	CMCT	CMCT + VE	CMCT + CTDN	ECMCT
2	57,8 ± 2,2	72,3 ± 2,1	75,5 ± 1,9	79,3 ± 1,4	**80,1 ± 1,3**
3	55,6 ± 2,2	69,4 ± 1,7	70,9 ± 1,7	75,8 ± 1,5	77,6 ± 1,5
4	54,0 ± 1,0	68,1 ± 1,1	69,5 ± 2,0	74,5 ± 2,1	76,4 ± 1,3

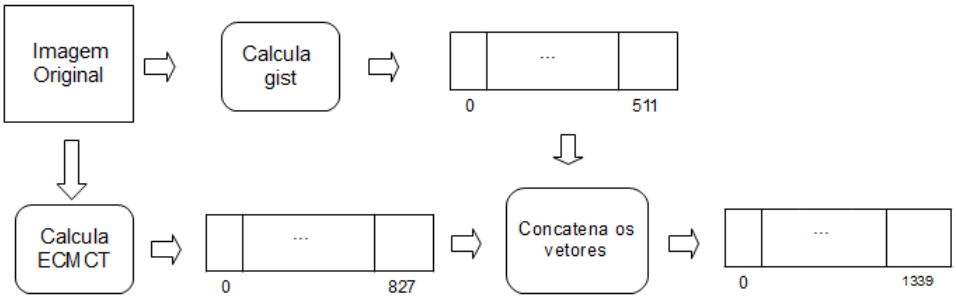


Figura 28 – Processo de obtenção do GECMCT.

4.4 Novos descritores baseados no ECMCT

No Capítulo 3, foram descritas melhorias no CMCT através da sua concatenação com o descritor *gist* e, também, através do uso da estratégia de combinação de classificadores. Nesta seção, os mesmos desenvolvimentos são feitos para o ECMCT. Assim, baseadas no ECMCT, três novas abordagem são expostas: *Gist*CMCTEstendido (GECMCT - *Gist Extended CMCT*), CMCT Estendido e MCT Espacial combinados (ECMCT-SM) e *Gist*CMCTEstendido e MCT Espacial combinados (GECMCT-SM).

4.4.1 *Gist*CMCTEstendido

Assim como o *Gist*CMCT apresentado no Capítulo 3, o *Gist*CMCTEstendido (GECMCT) é proposto com o objetivo de reunir as qualidades dos descritores *gist* e do CMCT Estendido, já que, assim como o CMCT, o ECMCT não captura as propriedades perceptuais da imagem. A Figura 28 ilustra a obtenção desse novo descritor. A imagem original é apresentada e os vetores dos descritores ECMCT e gist são extraídos. Por fim, os dois vetores são concatenados, gerando o descritor GECMCT.

Os parâmetros empregados para a obtenção do *gist* são os mesmos utilizados para o *Gist*CMCT na Seção 3.2, ou seja, filtros com 4 escalas e 8 orientações e uma amostragem da imagem de 4 x 4, gerando um descritor *gist* de 512 dimensões. Uma vez que o vetor gerado pelo ECMCT tem 828 posições, o descritor GECMCT é um vetor com 1.340 dimensões.

4.4.2 Descritores ECMCT-SM e GECMCT-SM

Na Seção 3.3.1, foi apresentado o descritor MCT Espacial, que tem como principal característica a extração de informações fornecidas pelo arranjo espacial na imagem. Apesar do ECMCT explorar alguma informação espacial, ela não é tão completa como a obtida pelo MCT Espacial, pois neste último o vetor de cada subdivisão da imagem é considerado, enquanto no ECMCT, utiliza-se apenas a informação estatística. Assim, com o propósito de complementar a informação espacial obtida pelo Vetor Estatístico no

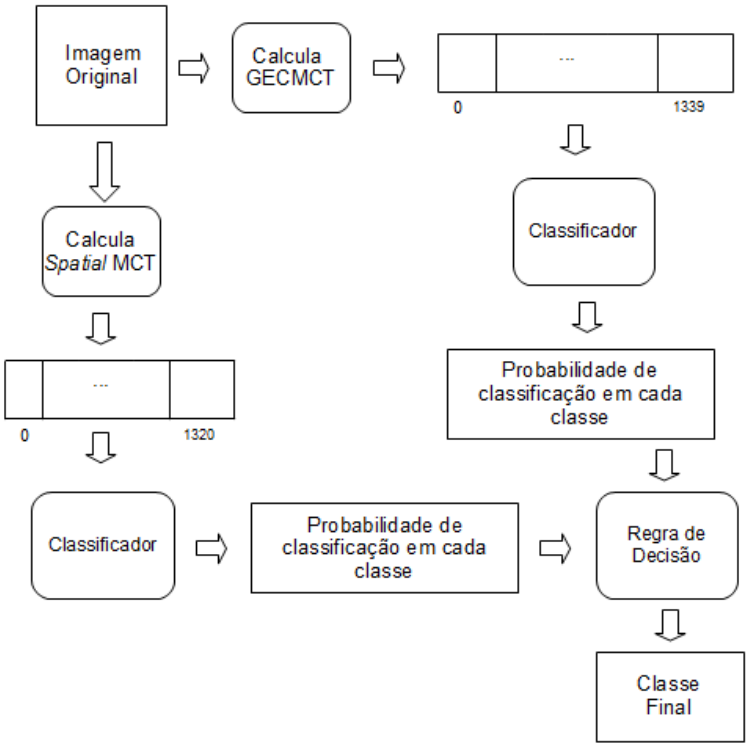


Figura 29 – Processo de obtenção do GECMCT-SM.

ECMCT, as características fornecidas pelo MCT Espacial são consideradas na classificação das imagens.

Para evitar o aumento excessivo do tamanho do vetor, ao invés de se adotar a concatenação dos vetores obtidos na extração dos dois descritores, é utilizada a estratégia de classificadores múltiplos. O classificador adotado é o SVM. Ademais, para a combinação dos resultados são adotadas as mesmas regras de combinação apresentadas na Seção 3.3.2: Produto, Máximo e Mediana. Essa nova estratégia de classificação é referenciada como CMCT Estendido e MCT Espacial combinados (ECMCT-SM).

Assim, como o ECMCT, o GECMCT também pode ter a informação espacial complementada através da inclusão das informações extraídas pelo MCT Espacial. Assim, adotou-se a mesma estratégia de classificadores múltiplos com o GECMCT e o MCT Espacial, referenciado como GistCMCTEstendido e MCT Espacial combinados (GECMCT-SM). As mesmas regras de combinação adotadas para o ECMCT-SM são utilizadas. A Figura 29 ilustra a obtenção do GECMCT-SM. Primeiro, os vetores dos descritores GECMCT e MCT Espacial são obtidos, depois, cada descritor é apresentado a um classificador individual. Cada classificador retorna, para cada classe, a probabilidade de que a imagem pertença a essa classe. Essas informações são, então, apresentadas à regra de combinação escolhida, quando a classe final é determinada.

4.5 Considerações Finais

A introdução da informação da vizinhança em torno das estruturas locais de uma imagem oferece um contexto para essas estruturas e faz com que a representação dessa imagem contenha informações sobre regiões mais amplas, compostas por essas estruturas e por seus vizinhos. Como visto no Capítulo 2, o CMCT inclui esse contexto. O ECMCT, por sua vez, tem como propósito ampliar ainda mais a vizinhança considerada, aumentando, desse modo, o tamanho das regiões identificadas. Além disso, o ECMCT inclui a informação de contraste, desconsiderada no CMCT, além de introduzir uma noção do arranjo espacial. As melhorias promovidas oferecem novas informações, tais como detalhes da textura, que permitem uma melhor definição dos elementos pertencentes a uma determinada classe. A nova abordagem mantém ainda a possibilidade de ser utilizada em sistemas voltados para usuários leigos, uma vez que não há necessidade de interação com o usuário.

As abordagens de combinação de descritores e combinação de classificadores também são explorados utilizando-se o descritor proposto, com a intenção de se obter uma representação mais acurada das imagens através do uso de técnicas distintas.

No próximo capítulo, serão apresentados os experimentos realizados para todos os descritores propostos neste trabalho e o desempenho de cada um deles será comparado com trabalhos similares existentes na literatura.

5 Experimentos Realizados

Este capítulo apresenta os resultados dos experimentos realizados utilizando-se as técnicas propostas, CMCT, *Gist*CMCT e *Gist*CMCT-SM, ECMCT, GECMCT, ECMCT-SM e GECMCT-SM em quatro bases de dados públicas. Antes, porém, são apresentadas informações sobre a implementação e sobre os métodos adotados.

5.1 O Classificador

Nos Capítulos 2 e 3, foram apresentados os resultados dos experimentos realizados utilizando-se o classificador K-NN, que apesar da sua simplicidade, exibiu resultados competitivos. Como exposto na Seção 2.4.3, de um ponto de vista teórico, para K=1 e um número ilimitado de amostras, a taxa de erro do K-NN nunca é pior do que duas vezes a taxa de Bayes. No entanto, na configuração prática, não é possível trabalhar com um número ilimitado de amostras. Além disso, o K-NN exige a determinação do número de vizinhos, e o valor ideal é variável de acordo com a base de dados.

Assim, nos experimentos relatados neste capítulo, foi utilizado o classificador SVM (VAPNIK, 1998), que é a escolha padrão na literatura de classificação de cenas (LIU; XU; FENG, 2011). Isso se deve a seu alto desempenho de classificação mesmo quando a dimensão do espaço de entrada é muito alta (CHAPELLE; HAFFNER; VAPNIK, 1999), como é o caso das imagens.

Basicamente, o objetivo do SVM é encontrar um hiperplano que aja como um separador de espaço de decisão de forma que a margem de separação entre exemplos positivos e negativos seja maximizada, o que é conhecido como hiperplano ótimo. A Figura 30 ilustra o hiperplano e as margens de separação. Caso os dados não sejam linearmente separáveis, como ilustrado na Figura 31, pode ser necessário mapear os dados do vetor de treinamento em um novo espaço, geralmente, de dimensão mais alta, chamado espaço de características. Após o mapeamento, é preciso encontrar o hiperplano ótimo e implementar a classificação no espaço de características.

Para dados que já possuem uma alta dimensão, esse mapeamento pode ser um problema, uma vez que a dimensão do espaço de características pode explodir exponencialmente (BENNETT; CAMPBELL, 2000). No entanto, isso não é um empecilho para o SVM, já que ele trabalha com a maximização da margem, ou seja, ele maximiza a distância entre o hiperplano de separação e os dados de treinamento. Além disso, não é prático calcular a função que mapeia os dados em uma dimensão mais alta, porém o SVM contorna esse problema através da utilização de *kernels* (BENNETT; CAMPBELL, 2000).

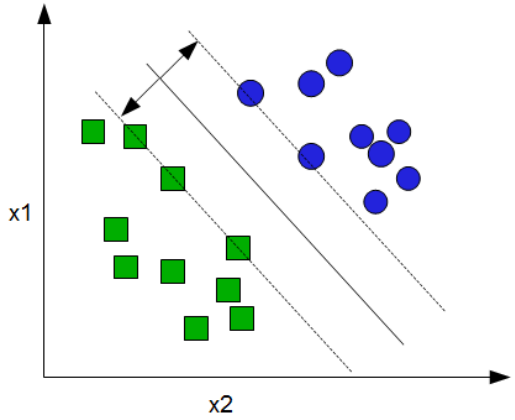


Figura 30 – Um hiperplano com suas margens de separação maximizada.

Uma função *kernel* k recebe dois pontos x_i e x_j e calcula o produto escalar $\Phi(x_i) \cdot \Phi(x_j)$ no espaço de característica, ou seja (BOSER; GUYON; VAPNIK, 1992),

$$k(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j) \tag{5.1}$$

Assim, é preciso apenas utilizar k no algoritmo e se torna desnecessário explicar ou conhecer a função Φ. As funções Φ mapeiam para o espaço de Hilbert de reprodução, um espaço onde é possível realizar o produto interno.

Neste trabalho, são relatados os resultados da classificação utilizando-se dois tipos de SVM: o linear, para os descritores CMCT e MCT8 e, para os demais descritores, o com *kernel* HIK (*Histogram Intersection kernel*)(BARLA; ODONE; VERRI, 2003), que é adequado quando se utiliza histogramas nos vetores de características, definido como a seguir:

$$k_{\mathrm{HI}}(x, y) = \sum_{j=1}^{D} \min(x_j, y_j) \tag{5.2}$$

onde x e y são histogramas com D dimensões.

5.2 Seleção de Características

Em alguns experimentos, foi adotada a técnica de seleção de características para a redução da dimensionalidade. A técnica empregada foi a BIF - (*Best Individual Features*) (JAIN; DUIN; MAO, 2000), que avalia as características de acordo com algum critério, as ordena e seleciona as k melhores. Esse método é rápido, eficiente e simples (NOVOVIČOVÄ; MALíK; PUDIL, 2004), o que o torna aplicável em problemas onde os vetores de características apresentam dimensões muito altas. O valor adotado para k foi 1.500. A

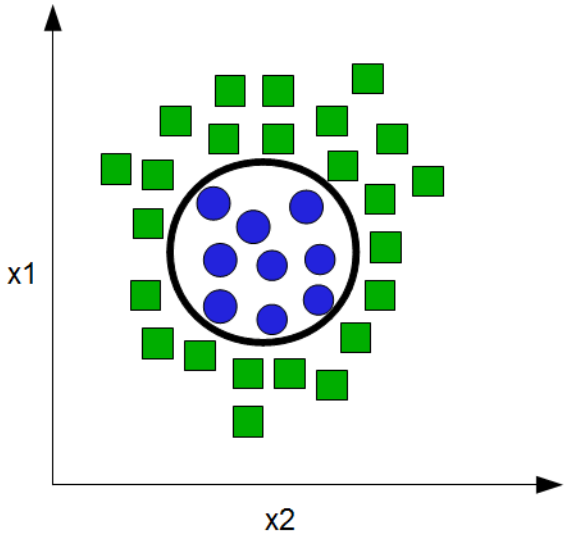


Figura 31 – Exemplo de dados que não podem ser separados linearmente.

execução deste algoritmo se deu através do PRTools4.1 (DUIN et al., 2007), desenvolvido em Matlab.

5.3 Algoritmos e Implementações

Com exceção do PCA que foi implementado em Matlab, todos os códigos foram implementados em C++ ou C. Além disso, para a manipulação de imagens, o pacote openCV (disponível em ⟨http://opencv.org/⟩) foi utilizado. Para a utilização do classificador SVM, foi adotado o pacote LibSVM (CHANG; LIN, 2011) modificado por Wu e Rehg (2009) e parâmetro de penalidade $C = 1$ (tal parâmetro controla o conflito entre os erros do SVM nos dados de treinamento e a maximização da margem). Para os descritores CMCT e MCT8, o *kernel* empregado no SVM foi o linear, já para os demais descritores, foi utilizado o *kernel* HIK (*Histogram Intersection kernel*)(BARLA; ODONE; VERRI, 2003). A escolha do *kernel* foi feita considerando-se os resultados obtidos através de diversos testes. A classificação com o descritor CMCT obteve melhores resultados com o *kernel* linear, enquanto a classificação com os demais descritores teve resultados melhores com o *kernel* HIK. É importante destacar, que o foco deste trabalho está em propor um descritor de imagens que alcance resultados competitivos na classificação de cenas e não em um classificador. Sendo assim, a escolha do classificador foi baseada nas escolhas da literatura e em testes exaustivos.

Para o *gist*, nos descritores *Gist*CMCT e GECMCT, foi utilizado o código de Lear (*Lear's Gist implementation*), uma reimplementação do código fornecido por Torralba, um dos autores do Envelope Espacial, e utilizado em Douze et al. (2009). No caso dos descritores que adotam a estratégia de combinação de classificadores, o MCT Espacial foi

obtido através de modificações do código C++ fornecido por Wu e Rehg (2011). Além disso, para encontrar a classe final, as regras de combinação necessitam da probabilidade que cada classificador atribui para uma imagem pertencer a uma classe, e o LibSVM oferece a probabilidade multiclasse estimada como opção de saída (WU; LIN; WENG, 2004).

Para realizar a classificação multiclasses, e sendo SVM um classificador binário, o LibSVM adota a estratégia um-contra-um, proposta em Knerr, Personnaz e Dreyfus (1990). Nesta estratégia, para k classes, são construídos $\frac{k(k-1)}{2}$ classificadores, cada um treina dados a partir de duas classes e a classificação final é encontrada através de votação (CHANG; LIN, 2011).

5.4 Base de Dados

As técnicas propostas foram testadas em 4 bases de dados públicas que incluem cenas internas e externas. As descrições dessas bases de dados são apresentadas abaixo e resumidas na Tabela 9:

- 8 categorias de cenas criada por Oliva e Torralba (2001). Essa base de dados contém 2.688 imagens coloridas, divididas em 8 categorias, com o número de imagens por classe variando entre 260 e 410. As categorias são as seguintes: costa (360 imagens), floresta (328 imagens), montanha (374 imagens), campo aberto (410 imagens), autoestrada (260 imagens), dentro da cidade (308 imagens), edifício alto (356 imagens) e rua (292 imagens). O tamanho de cada imagem é 256 x 256;

- 15 categorias de cenas (LAZEBNIK; SCHMID; PONCE, 2006). Uma extensão da base de dados de 8 categorias proposta por Oliva e Torralba (2001), através da adição das 7 classes seguintes: quarto (216 imagens), cozinha (210 imagens), sala de estar (289 imagens), escritório (215 imagens), subúrbio (241 imagens), indústria (311 imagens) e loja (315 imagens). No total, essa base de dados contém 4.486 imagens em tons de cinza. O tamanho de cada imagem é variado e em torno de 300 x 250. A Figura 32 apresenta alguns exemplos dessa base de dados. As categorias são (de cima para baixo e da esquerda para direita): costa, floresta, campo aberto, montanha, dentro da cidade, edifícios altos, autoestrada, quarto, rua, cozinha, sala de estar, escritório, loja, subúrbio e indústria.;

- 8 classes de eventos de esporte (LI; FEI-FEI, 2007). Essa base de dados contém 1.579 imagens de 8 esportes: *badminton* (214 imagens), bocha (142 imagens), críquete (245 imagens), polo (184 imagens), escalada em rocha (215 imagens), remo (254 imagens), navegação à vela (201 imagens) e *snowboard* (217 imagens). A Figura 33 apresenta alguns exemplos dessa base de dados. As categorias são (de cima para baixo e da

esquerda para direita): *badminton*, bocha, críquete, polo, escalada em rocha, remo, navegação à vela, e *snowboard*;

- 67 classes de cenas internas (QUATTONI; TORRALBA, 2009). Essa base de dados contém 15.620 imagens, divididas em 67 categorias, com o número de imagens por categoria variando entre 101 e 734. As cenas variam de corredor a padaria. Essa base de dados oferece um problema de classificação bastante desafiador (QUATTONI; TORRALBA, 2009).

Apesar de algumas bases apresentarem imagens coloridas, todas as imagens utilizadas nos experimentos foram convertidas para tons de cinza.

Tabela 9 – Resumo das informações sobre as bases de dados utilizadas neste trabalho.

Base de dados	**Total de imagens**	**Variação (entre)**
8 categorias de cenas	2.688	260 e 410
15 categorias de cenas	4.485	215 e 410
8 classes de eventos de esporte	1.579	142 e 254
67 classes de cenas internas	15.620	101 e 734

Figura 32 – Três imagens de cada uma das 15 categorias (LAZEBNIK; SCHMID; PONCE, 2006).

5.5 Procedimentos de Treino e Teste

A medida do desempenho do classificador é acurácia, estimada conforme a Equação 2.4 apresentada no Capítulo 2. Nos experimentos, cada categoria na base de dados é dividida de forma aleatória em um conjunto de dados de treino e outro de teste. Essa divisão é repetida 5 vezes, e, então, a média das acurácias e o desvio padrão são obtidos

Figura 33 – Duas imagens de cada uma das 8 classes de eventos de esporte (LI; FEI-FEI, 2007).

conforme adotado por Wu e Rehg (2011). O cálculo do desvio padrão é apresentado na Equação 2.5.

Seguindo a literatura, para as bases de dados de 8 e 15 categorias, foram utilizadas 100 imagens de treino em cada categoria e o restante foi utilizado para teste. Na categoria 8 eventos de esporte, foi utilizada a divisão adotada em Li e Fei-Fei (2007), na qual são utilizadas 70 imagens para treino e 60 para teste em cada categoria. Acompanhando Quattoni e Torralba (2009), na base de dados de 67 classes de cenas internas, foram utilizadas 80 imagens para treino e 20 para teste em cada categoria.

Para alguns dos resultados encontrados, é apresentada a matriz de confusão, uma ferramenta para a visualização de resultados de classificação, onde, em cada célula, é exibida a taxa de exemplos, pertencentes à classe exibida na linha, que foram classificadas na classe exibida na coluna. Assim, os valores exibidos na diagonal da matriz indicam a taxa de acertos para cada classe. Esse tipo de visualização dos resultados ajuda a identificar as "confusões" feitas pelo algoritmo de classificação entre classes. Para facilitar a leitura da matriz, somente as taxas de classificação maiores ou iguais a 0,1% foram exibidas. A taxa de classificação, t_{ci}, de elementos pertencentes à classe c classificados na classe i é calculada da seguinte forma:

$$t_{\mathrm{ci}} = \frac{n_i}{n_c} \tag{5.3}$$

onde n_i é o número de elementos pertencentes à classe c que foram classificados na classe

i e n_c é o total de elementos pertencentes à classe c.

5.6 Normalização dos Dados

A ponderação por log-frequência foi utilizada nos histogramas obtidos para o descritor CMCT. Essa é uma técnica utilizada na área de Recuperação de Informação (SALTON; MCGILL, 1983) que considera que a relevância de um termo não aumenta proporcionalmente com a sua frequência. A ponderação por log-frequência de um termo t em um documento d, $W_{t,d}$ é

$$W_{t,d} = \begin{cases} 1 + \log(f_{t,d}), & f_{t,d} > 0, \\ 0, & c.c. \end{cases} \tag{5.4}$$

onde $f_{t,d}$ é a frequência do termo t no documento d.

Um outro cálculo realizado é a normalização do vetor final de características de cada imagem. O valor final do termo t para a imagem i, $F_{t,i}$, é obtido da seguinte forma

$$F_{t,i} = \frac{V_{t,i} - V_{t\min}}{V_{t\max} - V_{t\min}}, \tag{5.5}$$

onde V_{tmax} é o maior valor encontrado para o termo t no conjunto de treino e V_{tmin}, o menor e $V_{t,i}$ é o valor inicial do termo t na imagem i. Assim, os valores no vetor de treinamento pertencem ao intervalo [0,1].

5.7 Resultados e Discussões

Nesta seção, são apresentados os resultados obtidos através do uso dos descritores CMCT, ECMCT, *Gist*CMCT e GECMT e é feita uma comparação com os trabalhos existentes na literatura.

5.7.1 Experimentos com o CMCT e o ECMCT

5.7.1.1 15 Categorias de Cenas

A primeira base de dados considerada é a de 15 categorias de cenas, onde a acurácia na classificação alcançada pelo CMCT é de 77,61 ± 0,42%. A Figura 34 apresenta a matriz de confusão para uma execução nessa base de dados. Nela é possível observar que o melhor valor de acurácia é de 97% e ocorre na classe "subúrbio". Essa classe exibe melhor desempenho, porque apresenta uma baixa variação intraclasse, o que acarreta em uma melhor representação dos seus elementos e uma boa taxa de acerto na classificação. As maiores confusões ocorrem entre as classes "sala de estar" e "quarto", que apresentam

objetos comuns, tais como: cadeiras, tv's e luminárias; "escritório" e "sala de estar", que também apresentam elementos comuns; "indústria" e "loja", que são classes que exibem uma grande variabilidade intraclasse e que, em, alguns casos, apresentam exemplos que possuem arranjos espaciais similares, como na imagem apresentada na Figura 35, que pertence à classe "loja", mas tem uma arranjo semelhante ao encontrado em alguns exemplos da classe "indústria"; "campo aberto" e "costa", que em algumas imagens apresentam arranjos espaciais semelhantes, como as exibidas na Figura 36 e "campo aberto" e "montanha", que também apresentam imagens com arranjos espaciais similares.

	Quarto	Sala de estar	Subúrbio	Indústria	Cozinha	Costa	Floresta	Autoestrada	Dentro da cidade	Montanha	Campo aberto	Rua	Edifício alto	Escritório	Loja
Quarto	0,6	0,22			0,11										
Sala de estar	0,26	0,57													
Subúrbio			0,97												
Indústria				0,66											0,16
Cozinha					0,67										0,11
Costa						0,82					0,11				
Floresta							0,9								
Autoestrada								0,84							
Dentro da cidade									0,86						
Montanha										0,81	0,11				
Campo aberto						0,1					0,71				
Rua												0,78			
Edifício alto													0,8		
Escritório		0,13												0,81	
Loja				0,1											0,77

Figura 34 – Matriz de confusão para uma execução do experimento de classificação na base de dados de 15 cenas utilizando o descritor CMCT. Somente as taxas (%) maiores ou iguais a 0,1% foram representadas.

Figura 35 – Imagem pertencente à classe "loja" com estrutura espacial similar a de elementos da classe "indústria". Imagem retirada de (LAZEBNIK; SCHMID; PONCE, 2006)

A Tabela 10 compara os resultados da classificação dos métodos propostos na base de dados de 15 cenas com os resultados existentes na literatura e com o MCT8 mencionado no

Figura 36 – Imagens pertencentes a classes distintas, mas com estruturas espaciais similares. A da esquerda pertencente à classe "costa" e a da direita, à classe "campo aberto". Ambas retiradas de Lazebnik, Schmid e Ponce (2006).

Tabela 10 – Resultados de classificação na base de dados de 15 cenas com os descritores CMCT, ECMCT e de trabalhos existentes na literatura. Em negrito, as abordagens propostas.

Método	**Acurácia(%)**
CENTRIST (WU; REHG, 2011)	73,29 ± 0,96
SPM (Caract. fracas) (LAZEBNIK; SCHMID; PONCE, 2006)	66,80 ± 0,6
SPM (200 centros)(LAZEBNIK; SCHMID; PONCE, 2006)	72,20 ± 0,6
SPM (400 centros) (LAZEBNIK; SCHMID; PONCE, 2006)	74,80 ± 0,3
gist (OLIVA; TORRALBA, 2001)	73,28 ± 0,67
RCVW (LIU; XU; FENG, 2011)	74,5
MCT8 (GAZOLLI; SALLES, 2012)	75,34 ± 0,76
CMCT (GAZOLLI; SALLES, 2012)	**77,61 ± 0,42**
ECMCT (GAZOLLI; SALLES, 2014)	**81,33 ± 0,49**

Capítulo 2. Todos os experimentos aqui referenciados utilizam o classificador SVM. Como o objetivo é comparar a eficiência dos descritores, somente os resultados de experimentos que não utilizaram a pirâmide espacial (ou número de níveis=0) foram considerados. Em Lazebnik, Schmid e Ponce (2006), dois tipos de características foram utilizados no experimentos com o método SPM, as características fracas, pontos onde a magnitude do gradiente em uma dada direção excede um limiar mínimo, e as fortes, descritores SIFT computados em pedaços de 16 x 16 da imagem. As características fortes apresentam maior acurácia em relação às fracas. O CMCT, por sua vez, superou tanto as características fracas como as fortes. O modelo RCVW *Region Contextual Visual Words* (LIU; XU; FENG, 2011), que introduz informação contextual na estratégia *bag-of-words* também foi superado pelo CMCT. O CMCT também superou o CENTRIST e, em relação ao MCT8, é possível constatar que a inclusão do contexto melhora o desempenho de classificação, uma vez que o MCT8 alcançou 75,34 ± 0,30% de acurácia.

Quando se observa os resultados obtidos pelo ECMCT, verifica-se que as informações adicionadas ao descritor CMCT aumentam a sua capacidade de representação para classi-

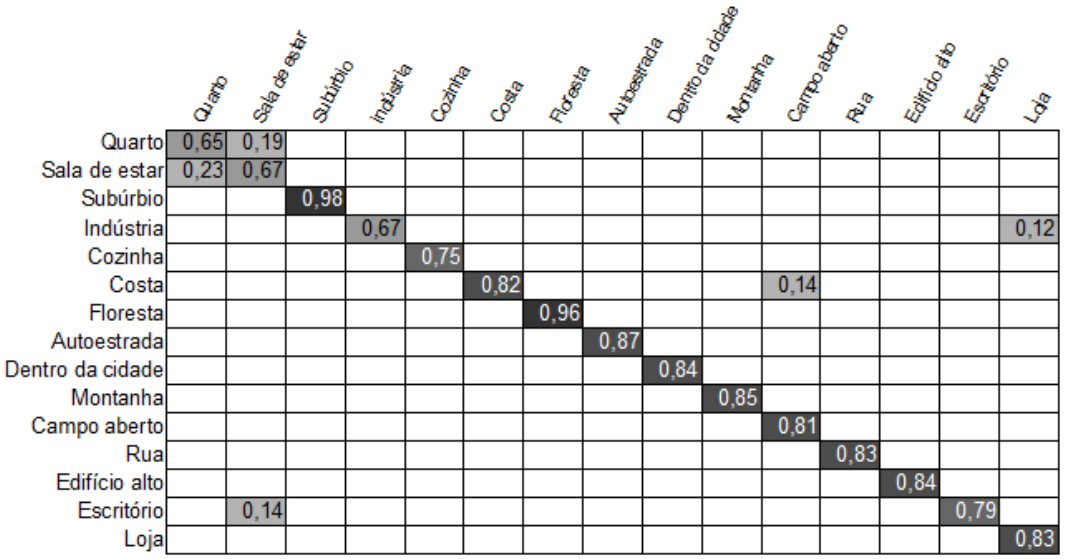


Figura 37 – Matriz de confusão para uma execução do experimento de classificação na base de dados de 15 cenas utilizando o descritor ECMCT. Somente as taxas (%) maiores ou iguais a 0,1% foram representadas.

ficação, pois o ECMCT obtém, na base de 15 categorias, um aumento de, aproximadamente, 5% na taxa de classificação em relação ao descritor original. A Figura 37 apresenta a matriz de confusão para o ECMCT, onde é possível perceber um aumento na taxa de classificação para todas as classes, exceto para a classe "costa", onde a taxa de acerto permanceu a mesma, e para as classes "escritório" e "dentro da cidade", onde houve uma redução. Houve também uma redução da confusão (taxas de classificação incorretas acima de 0,1%), exceto entre as classe "sala de estar" e "escritório".

5.7.1.2 8 Categorias de Cenas

A Figura 38 apresenta os resultados de classificação utilizando o CMCT para uma execução do experimento na base de dados de 8 categorias de cenas. É possível notar que as maiores confusões ocorrem entre as classes "campo aberto" e "montanha", com arranjos espaciais similares, e "rua" e "dentro da cidade", que possuem objetos em comuns, e o melhor resultado ocorreu para a classe "floresta", que apresenta pouca semelhança com as imagens pertencentes a outras classes.

A Tabela 11 apresenta os resultados experimentais para essa base de dados. Novamente, todos os experimentos aqui referenciados utilizam o classificador SVM. O descritor *gist* obteve uma acurácia na classificação de 82,60 ± 0,86%, que é maior do que a alcançada pelo CMCT. Entretanto, na base de dados de 15 cenas que inclui muitas cenas de ambientes internos, a acurácia obtida pelo descritor *gist* cai para 73,28 ± 0,67%, que é menor do que a alcançada pelo CMCT. Como na base de dados de 15 cenas, CMCT apresenta melhores resultados do que o CENTRIST e o MCT8.

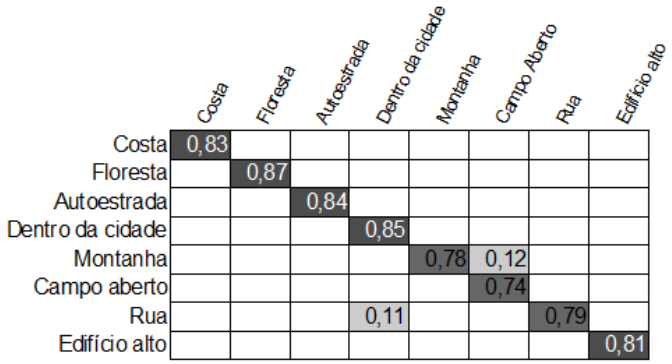


Figura 38 – Matriz de confusão para uma execução do experimento de classificação na base de dados de 8 categorias de cenas utilizando o descritor CMCT. Somente as taxas (%) maiores ou iguais a 0,1% foram representadas.

O CMCT Estendido novamente obteve resultados melhores do que a sua versão original, elevando em, aproximadamente, 5% a taxa de acerto na classificação. A Figura 39, apresenta a matriz de confusão para esse descritor, onde é possível perceber um aumento na taxa de classficação para todas as classes. A confusão foi reduzida, mas entre as classes "rua" e "dentro da cidade", que apresentam elementos comuns, ainda atingiu 0,1%. Ao contrário do CMCT, o ECMCT obteve um desempenho superior ao obtido pelo *gist*. Assim, a informação adicional, introduzida no CMCT, suprimiu a deficiência na representação das cenas externas presentes nesta base de dados.

Tabela 11 – Resultados alcançados nos experimentos realizados na base de dados de 8 categorias de cenas com os descritores CMCT e ECMCT e com outros métodos da literatura. Em negrito, as abordagens propostas.

Tipo de Característica	Acurácia (%)
gist	82,60 ± 0,86
CENTRIST	76,49 ± 0,84
MCT8 (GAZOLLI; SALLES, 2012)	78,97 ± 1,26
CMCT (GAZOLLI; SALLES, 2012)	**81,23 ± 0,73**
ECMCT (GAZOLLI; SALLES, 2014)	**85,31 ± 0,74**

5.7.1.3 8 Eventos de Esporte

A Tabela 12 apresenta os resultados dos experimentos realizados na base de dados 8 classes de eventos de esporte. O descritor CMCT alcançou nessa base 70,79 ± 1,86%, enquanto o CENTRIST alcançou 63,91 ± 2,44%. Em Li e Fei-Fei (2007), que utiliza um modelo generativo, e onde a classificação dos eventos é o resultado da combinação da classificação do ambiente com a dos objetos, a acurácia é de 73,4% e, portanto, maior do que o CMCT. Entretanto, nessa abordagem há a segmentação manual e os rótulos dos objetos são usados como entradas adicionais, um procedimento que não é adotado no CMCT. O ECMCT, novamente, atingiu taxas de acertos melhores do que o CMCT,

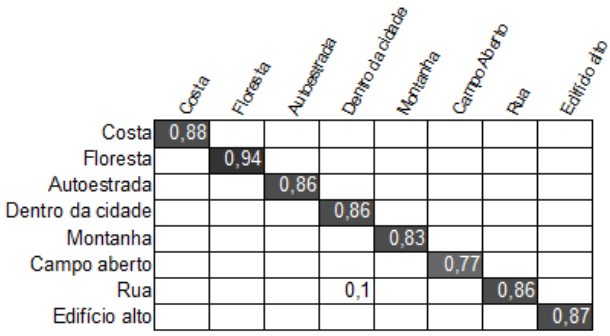


Figura 39 – Matriz de confusão para uma execução do experimento de classificação na base de dados de 8 categorias de cenas utilizando o descritor ECMCT. Somente as taxas (%) maiores ou iguais a 0,1% foram representadas.

chegando a 75,04 ± 2,35%, superando, inclusive, os resultados alcançados por Li e Fei-Fei (2007).

Tabela 12 – Resultados alcançados nos experimentos realizados na base de dados de 8 categorias de eventos de esporte com os descritores CMCT, ECMCT e com outros métodos da literatura. Em negrito, as técnicas propostas.

Método	Acurácia (%)
Ambiente + objeto	73,40
CENTRIST	63,91 ± 2,44
MCT8	69,33 ± 2,11
CMCT (GAZOLLI; SALLES, 2012)	70,79 ± 1,86
ECMCT (GAZOLLI; SALLES, 2014)	**75,04 ± 2,35**

5.7.1.4 67 Classes de Cenas Internas

Na base de dados 67 classes de cenas internas, os experimentos executados por Quattoni e Torralba (2009) com o *gist* alcançaram 21% de taxa de reconhecimento. Com o uso de informação global e local (*gist* + histograma de palavras visuais) para a representação de cenas, a acurácia foi aumentada para 25%. Usando o CMCT, alcançou-se 25,82 ± 0,72%. Já os experimentos executados com o CENTRIST alcançaram 22,46 ± 0,84%. Assim, nesta base de dados desafiadora, CMCT apresentou melhores resultados do que os métodos apresentados. Já, o ECMCT obteve um aumento na taxa de classificação de, aproximadamente, 22% em relação ao CMCT. A Tabela 13 apresenta os resultados para os experimentos nessa base.

5.7.2 Experimentos com o *Gist*CMCT e GECMCT

5.7.2.1 15 Categorias de Cenas

Na base de dados de 15 cenas, os experimentos realizados com o *Gist*CMCT levaram a 82,72 ± 0,55% de acurácia, o que significa um aumento de, aproximadamente, 7% em

Tabela 13 – Resultados alcançados nos experimentos realizados na base de 67 cenas de ambientes internos com os descritores CMCT, ECMCT e com outros métodos da literatura. Em negrito, as abordagens propostas.

Método	Acurácia (%)
gist	21
CENTRIST	22,46 ± 0,84
local + global	25
MCT8	24,32 ± 1,09
CMCT (GAZOLLI; SALLES, 2012)	**27,67 ± 0,47**
ECMCT (GAZOLLI; SALLES, 2014)	**33,64 ± 0,75**

relação ao descritor CMCT, conforme visto na Tabela 14. Os resultados obtidos pelo *Gist*CMCT também superaram os resultados do método *gist* com ARP (LIU; KIRANYAZ; GABBOUJ, 2012), uma técnica que melhora o *gist* através da modificação das divisões feitas na imagem durante o cálculo e que obteve 75,25 ± 0,67% de acurácia utilizando 4 ângulos e um vetor de 2048 dimensões.

Tabela 14 – Resumo dos resultados alcançados (acurácia %) pelos descritores CMCT, ECMCT, *Gist*CMCT e GECMCT.

Base de Dados	CMCT	ECMCT	*Gist*CMCT	GECMCT
15 categorias	77,61 ± 0,42	81,33 ± 0,49	82,72 ± 0,55	84,06 ± 0,19
8 categorias	81,23 ± 0,73	85,31 ± 0,74	85,72 ± 0,99	86,95 ± 0,81
8 eventos de esporte	70,99 ± 1,86	75,04 ± 2,35	74,37 ± 1,34	77 ± 1,63
67 categorias	27,67 ± 0,47	33,64 ± 0,75	33,60 ± 1,30	36,57 ± 1,23

Ao se comparar as matrizes de confusão das Figura 34 e Figura 40, nota-se que, a não ser para classe “montanha”, as taxas de acerto em todas as classes, tanto de ambientes internos como externos, aumentaram. Assim, é possível verificar que as informações da imagem presentes no *gist* ajudam a superar as deficiências do CMCT no que tange o reconhecimento de ambientes externos, já que o *gist* superou o CMCT nos experimentos realizados na base de dados de 8 cenas, a qual só apresenta cenas de ambientes externos. Um outro ponto a se notar é que nem todas as confusões que ocorrem no CMCT ocorrem no *Gist*CMCT, mas as confusões entre as classes “sala de estar” e “quarto”, “campo aberto” e “costa” e ‘indústria” e “loja” ainda permanecem acima de 0,1%.

Já, o GECMCT obteve melhores resultados que o ECMCT nesta base de dados, alcançando uma taxa de acerto de 84,06 ± 0,19%, ou seja, aproximadamente, 3% maior do que a taxa obtida pelo ECMCT e 2% maior do que a obtida pelo GistCMCT. A Figura 41 apresenta a matriz de confusão para o GECMCT, onde é possível verificar um aumento em quase todas as classes (exceto para as classes “floresta”, “autoestrada” e “loja”), quando comparado com o ECMCT (Figura 34). Merece destaque a classe “escritório”, para a qual obteve-se um aumento de 11% na taxa de classificação. A confusão entre as classes “quarto”

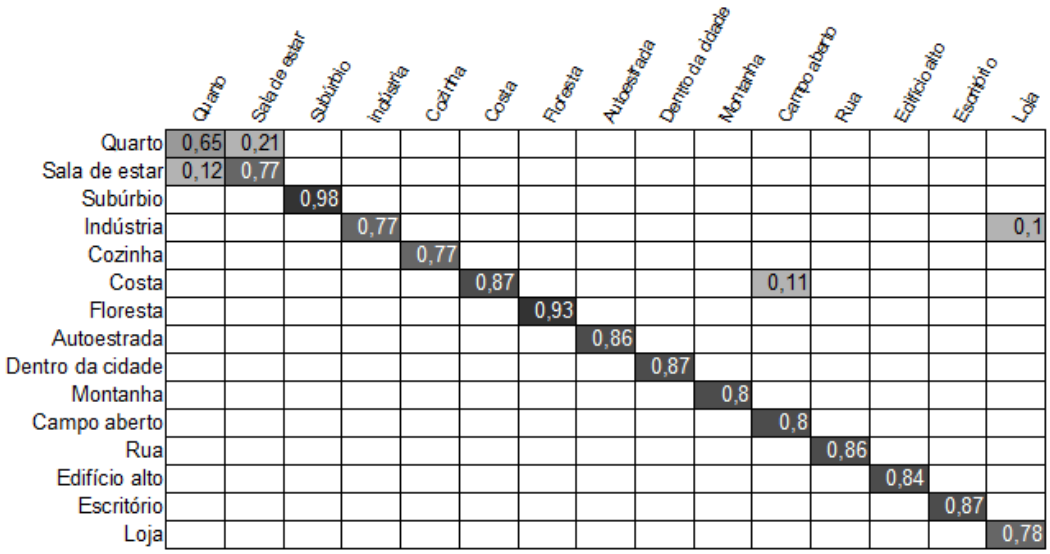


Figura 40 – Matriz de confusão para uma execução do experimento de classificação com o descritor *Gist*CMCT na base de dados 15 cenas. Somente as taxas (%) maiores ou iguais a 0,1% foram representadas.

e “sala” e “loja” e “indústria” permanece.

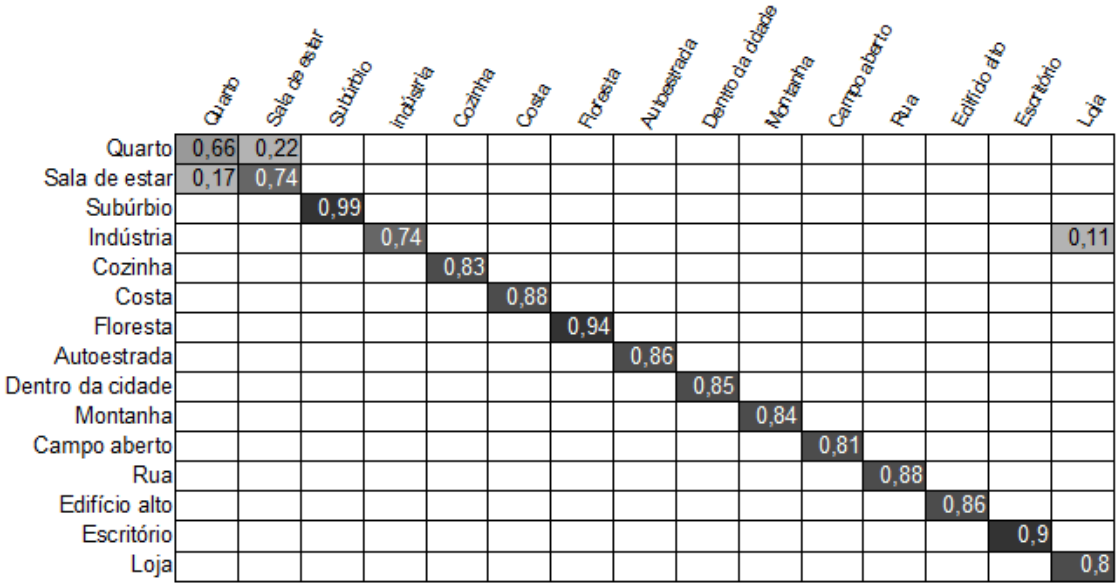


Figura 41 – Matriz de confusão para uma execução do experimento de classificação com o descritor GECMCT na base de dados 15 cenas. Somente as taxas (%) maiores ou iguais a 0,1% foram representadas.

5.7.2.2 8 Categorias de Cenas

O *Gist*CMCT também obteve melhores resultados que o CMCT na base de dados de 8 categorias de cenas (Tabela 14), levando a um aumento de 6% no desempenho de classificação, visto que o *Gist*CMCT alcançou uma acurácia de 85,72 ± 0,99% nessa base

de dados. O método *gist* com ARP (LIU; KIRANYAZ; GABBOUJ, 2012) que alcançou uma acurácia de 84,77 ± 0,71%, com 3 ângulos e um vetor de 1.536 dimensões, também foi superado. Ao contrário de CMCT, o *Gist*CMCT supera o *gist* nessa base de dados. A Figura 42 apresenta um gráfico com as taxas de classificação para o *gist*, CMCT e *Gist*CMCT onde é possível verificar que o *Gist*CMCT melhora o desempenho para todas as classes quando comparado com o *gist* ou com o CMCT isoladamente.

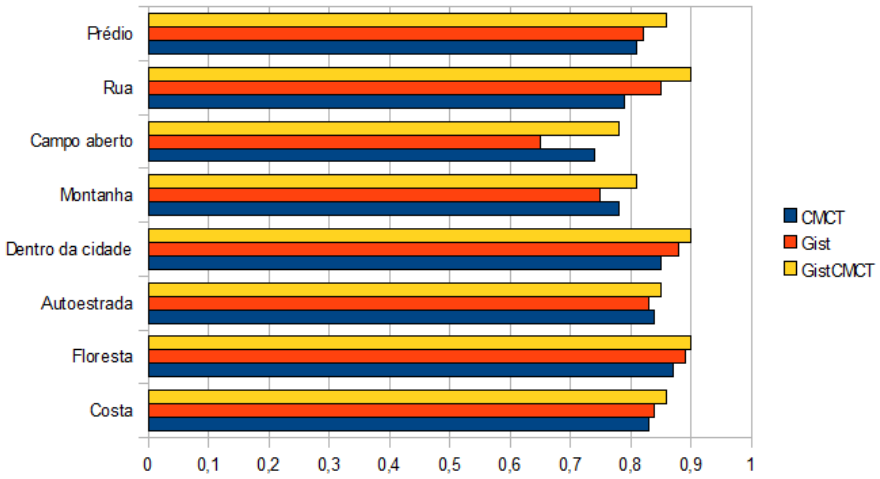


Figura 42 – Taxas de classificação por classe para os descritores *gist*, CMCT e *Gist*CMCT para uma execução na base de dados de 8 cenas.

O GECMCT alcançou, nesta base de dados, uma taxa de acerto na classificação de 86,95 ± 0,81, obtendo, portanto, melhor desempenho que o ECMCT e o *Gist*CMCT.

5.7.2.3 8 Cenas de Eventos de Esporte

Na base de dados de 8 cenas de eventos de esporte, o *Gist*CMCT alcançou uma acurácia de 74,37 ± 1,34% e, portanto, superou o CMCT (Tabela 14) e os resultados apresentados em Li e Fei-Fei (2007). O GECMCT, por sua vez, alcançou uma taxa de acerto de 77 ± 1,63% e, novamente, obteve um desempenho superior ao obtido pelo ECMCT e pelo *Gist*CMCT.

5.7.2.4 67 Classes de Cenas Internas

Na base de dados de 67 classes de cenas internas, o *Gist*CMCT alcançou uma acurácia de 33,60 ± 1,30% exibindo um resultado 30% superior ao CMCT, conforme visto na Tabela 14. É interessante notar que mesmo com a ausência de classes de ambiente externos, a concatenação do *gist* com o CMCT melhora os desempenhos de classificação nessa base de dados. A abordagem GECMCT também eleva os resultados nesta base de dados, atingindo um desempenho de classificação de 36,57 ± 1,23% e, superando, mais uma vez, os resultados do *Gist*CMCT e do ECMCT.

5.8 Experimentos com Estratégias de Combinação dos Classificadores

Nesta seção são apresentados os resultados obtidos através do uso das estratégias combinadas, *Gist*CMCT-SM, ECMCT-SM e GECMCT-SM, que são comparados com resultados existentes na literatura.

5.8.1 15 Categorias de Cenas

Na base de dados de 15 cenas, a abordagem *Gist*CMCT-SM alcançou 86,25 ± 0,51%. A Figura 43 apresenta a matriz de confusão para uma execução na base de dados de 15 cenas usando a regra do produto. Nela é possível observar que a maior taxa de reconhecimento ocorre na classe "subúrbio". A maior confusão ocorre entre as classes "sala de estar" e "quarto". Como dito anteriormente, essas classes se confundem por apresentarem elementos comuns. A Tabela 15 apresenta o desempenho de classificação das abordagens combinadas propostas para a base de dados de 15 cenas comparados com métodos existentes na literatura. Todas as abordagem exibidas nesta seção utilizaram o classificador SVM.

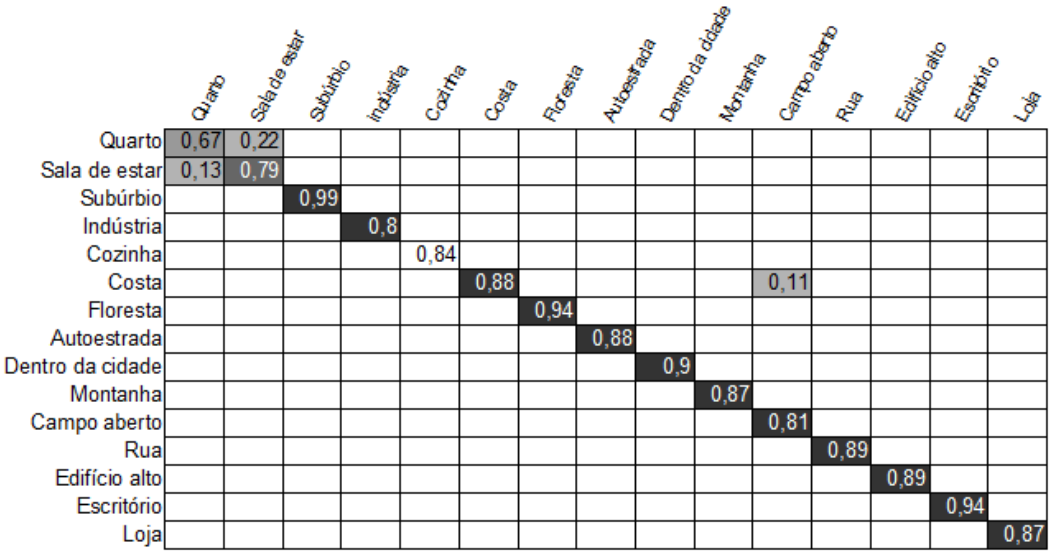


Figura 43 – Matriz de confusão para uma execução do experimento de classificação utilizando a estratégia *gist*CMCT-SM na base de dados de 15 categorias de cenas.

Em Lazebnik, Schmid e Ponce (2006), o melhor desempenho ocorre com o tamanho do vocabulário igual a 400 e com uma pirâmide espacial de 3 níveis (o que leva a um vetor final de 34.000 dimensões). No método proposto por Ergul e Arica (2010), que combina duas abordagens populares na literatura: SPM e PLSA, o PLSA em cascata, os experimentos referenciados utilizam um vetor de 5.900 posições. Em Wu e Rehg (2011), o *spatial* PACT (CENTRIST com níveis) com 3 níveis e 40 autovetores e 1.302 dimensões

Tabela 15 – Resultados dos experimentos realizados na base de dados de 15 cenas com as estratégias combinadas e com métodos existentes na literatura. Em negrito, as abordagens propostas.

Método	Acurácia(%)
SPM (LAZEBNIK; SCHMID; PONCE, 2006)	81,4 ± 0,5
PLSA em cascata (ERGUL; ARICA, 2010)	83,31
spatial Pact (3 níveis) (WU; REHG, 2011)	83,88 ± 0,76
LDBP (MENG; WANG; WU, 2012)	84,10 ± 0,96
CBoW _SL (LI et al., 2011)	85,1
LBP + Semantic (LI; DEWEN, 2011)	85,1
*Wave*LBP (SONG; LI, 2013)	81,5
GBPWHGO (ZHOU; ZHOU; HU, 2013)	85,2 ± 0,7
MCT Espacial (GAZOLLI; SALLES, 2013)	84,47 ± 0,63
*Gist*CMCT-SM (BIF)	84,34 ± 0,36
***Gist*CMCT-SM (Máximo)** (GAZOLLI; SALLES, 2013)	**86,02 ± 0,46**
***Gist*CMCT-SM (Mediana)** (GAZOLLI; SALLES, 2013)	**86,30 ± 0,62**
***Gist*CMCT-SM (Produto)** (GAZOLLI; SALLES, 2013)	**86,25 ± 0,51**
ECMCT-SM (Produto) (GAZOLLI; SALLES, 2014)	**85,82 ± 0,62**
ECMCT-SM (BIF)	85,08 ± 0,48
GECMCT-SM (Produto) (GAZOLLI; SALLES, 2014)	**86,75 ± 0,41**

é o que alcança melhores resultados. Em Meng, Wang e Wu (2012), as imagens são representadas por histogramas de transformações locais, uma extensão do histograma de transformada census, através da Diferença Local de Padrões Binários, LDBP (*Local Difference Binary Pattern*). Além disso, essa abordagem também utiliza a representação de pirâmide espacial e o descritor final tem 840 dimensões. Em Liu, Xu e Feng (2011), onde uma nova representação contextual de *bag-of-words*, CBoW (*Contextual Bag-of-Words*), foi proposta para modelar duas relações contextuais típicas: relação semântica conceitual e relação espacial de vizinhança, o melhor desempenho alcançado utiliza um vetor de 2.250 dimensões. Li e Dewen (2011) que combinam descritores de baixo nível, utilizando o LBP, com estratégias de modelagem semântica, extração de características locais e geração de *codebook*, não informaram o tamanho do *codebook* utilizado. O LBP também é utilizado na abordagem proposta por Song e Li (2013), que o combina com a transformada *wavelet* em uma estrutura hierárquica, gerando um vetor de 5.360 dimensões. Já em Zhou, Zhou e Hu (2013), é feita a combinação do GBP (*Gradiente Binary Pattern*) com o WHGO (*Weighted Histogram of Gradient Orientation*). Como é possível verificar, a estratégia *Gist*CMCT-SM proposta oferece melhores resultados do que os métodos mencionados acima, incluindo o *Gist*CMCT e o MCT Espacial isoladamente. Com respeito às regras de combinação aqui utilizadas, os resultados foram próximos, mas a regra do máximo obteve o pior resultado.

Em relação às combinações feitas utilizando-se o ECMCT, verifica-se que o *Gist*CMCT-SM apresenta melhor resultado do que o ECMCT-SM e um resultado muito próximo ao obtido pelo GECMCT-SM. Assim, para essa base de dados as informações

adicionais introduzidas pelo ECMCT não melhoram o desempenho da classificação. Quando se compara o ECMCT com o ECMCT-SM e o GECMCT com o GECMCT-SM, no entanto, é possível verificar que a estratégia de combinação aumenta o desempenho na classificação em relação aos descritores originais.

Já, a técnica de seleção de caraterísticas utilizando o BIF apresentou resultados inferiores aos alcançados pela estratégia de combinação de características tanto para *Gist*CMCT-SM como para o ECMCT-SM.

5.8.2 8 Categorias de Cenas

Na base de dados de 8 cenas, a estratégia *Gist*CMCT-SM alcançou 88,95 ± 0,49% com a regra do produto. A Tabela 16 apresenta os resultados dos experimentos realizados nessa base de dados. Como é possível observar, as abordagens propostas superam todos os métodos apresentados. O método *Novel Gist* (MENG; WANG, 2010) é uma extensão da transformada census e também utiliza informação espacial. Nessa técnica, os histogramas de padrão superior e inferior são calculados e concatenados. Os experimentos aqui relatados com o *Novel Gist* utilizam o classificador SVM e um vetor de 1.610 dimensões. Com relação aos descritores *Gist*CMCT e MCT Espacial, mais uma vez, o método *Gist*CMCT-SM alcançou melhores resultados do que as estratégias isoladas. Com relação à regra de combinação, os resultados, novamente, foram próximos, mas, como na base de dados de 15 cenas, a regra do Máximo alcançou os piores resultados.

Tabela 16 – Resultados dos experimentos realizados na base de dados de 8 cenas com as estratégias combinadas e com métodos existentes na literatura. Em negrito, as abordagens propostas.

Método	Acurácia (%)
gist (OLIVA; TORRALBA, 2001)	82,60 ± 0,86
Novel Gist (MENG; WANG, 2010)	86,60 ± 0,53
GBPWHGO (ZHOU; ZHOU; HU, 2013)	87,1 ± 0,6
spatial PACT (3 níveis) (WU; REHG, 2011)	86,2 ± 1,02
MCT Espacial (GAZOLLI; SALLES, 2013)	87,65 ± 0,24
*Gist*CMCT-SM (BIF)	87,56 ± 1,03
***Gist*CMCT-SM (Máximo)** (GAZOLLI; SALLES, 2013)	**88,51 ± 0,33**
***Gist*CMCT-SM (Mediana)** (GAZOLLI; SALLES, 2013)	**88,83 ± 0,46**
***Gist*CMCT-SM (Produto)** (GAZOLLI; SALLES, 2013)	**88,95 ± 0,49**
ECMCT-SM (BIF)	88,11 ± 0,44
ECMCT-SM (Produto) (GAZOLLI; SALLES, 2014)	**88,60 ± 0,69**
GECMCT-SM (Produto) (GAZOLLI; SALLES, 2014)	**89,06 ± 0,47**

A estratégia ECMCT-SM teve desempenho muito próximo ao *Gist*CMCT-SM e, também, superou os resultados da abordagem isolada. Já a abordagem GECMCT-SM superou tanto o *Gist*CMCT-SM, quanto as abordagens isoladas.

A técnica de seleção de caraterísticas para o *Gist*CMCT-SM apresentou resultados inferiores aos alcançados pela estratégia de combinação de classificadores, mas obteve resultados próximos aos dessa abordagem, quando aplicada ao ECMCT-SM. A adoção da seleção de características tem como vantagem a simplificação do processo de classificação, já que o classificador trabalha com um número menor de características, mas necessita da definição de parâmetros, critério e número de características, e, no caso da classificação de cenas, nem sempre alcança resultados próximos aos da estratégia de combinação de classificadores.

5.8.3 8 Classes de Eventos de Esportes

A Tabela 17 apresenta os resultados alcançados na base de dados de 8 classes de eventos de esporte. O método *Gist*CMCT-SM superou os resultados alcançados pelo *spatial* PACT (com três níveis) e pelo GBPWHGO, mas obteve resultados próximos ao do MCT Espacial com uma diferença menor do que 1%.

A Figura 44 apresenta a comparação entre *Gist*CMCT, MCT Espacial e o método combinado para uma execução nessa base de dados. Nela é possível observar que a taxa de reconhecimento na abordagem combinada para as classes "navegação à vela" e "bocha " são piores na abordagem combinada do que no MCT Espacial. Assim, para essas classes, a informação introduzida pelo *Gist*CMCT levou à redução da acurácia da classificação. Com relação às regras de combinação, os resultados foram muito próximos, mas a regra Máximo foi a que obteve pior desempenho.

Quando se compara os resultados obtidos pelo *Gist*CMCT-SM, ECMCT-SM e GECMCT-SM, percebe-se resultados muito próximos, mas, no caso do ECMCT-SM, a diferença em relação ao MCT Espacial representa uma aumento de apenas 1,5%.

Tabela 17 – Comparação dos resultados dos experimentos realizados na base de dados de 8 classes de eventos de esporte com as estratégias combinadas e com métodos existentes na literatura. Em negrito, as abordagens propostas.

Método	Acurácia (%)
spatial PACT (3 níveis) (WU; REHG, 2011)	78,25 ± 1,27
GBPWHGO (ZHOU; ZHOU; HU, 2013)	79,3 ± 1,5
MCT Espacial (GAZOLLI; SALLES, 2013)	80,63 ± 1,17
***Gist*CMCT-SM (Máximo)** (GAZOLLI; SALLES, 2013)	**80,96 ± 0,86**
***Gist*CMCT-SM (Mediana)** (GAZOLLI; SALLES, 2013)	**81,33 ± 1,19**
***Gist*CMCT-SM (Produto)** (GAZOLLI; SALLES, 2013)	**81,54 ± 0,62**
ECMCT-SM (Produto) (GAZOLLI; SALLES, 2014)	**81,87 ± 1,81**
GECMCT-SM (Produto) (GAZOLLI; SALLES, 2014)	**81,46 ± 1,88**

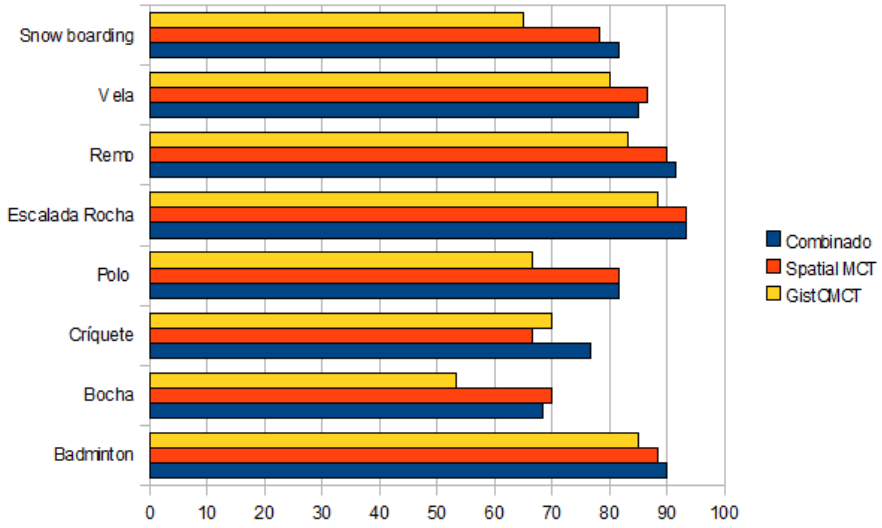


Figura 44 – Taxa de acerto na classificação por classe para os descritores *Gist*CMCT, MCT Espacial e a estratégia combinada para uma execução de um experimento na base de dados de 8 classes de eventos de esporte.

5.8.4 67 Classes de Cenas Internas

Com relação à base de dados 67 classes de cenas internas, GBPWHGO (NIU; ZHOU; SHI, 2010), uma técnica que combina informação global com estruturas locais em um vetor de 34.692 dimensões, e *spatial* PACT (com 3 níveis) obtiveram resultados de classificação inferiores aos obtidos pelo *Gist*CMCT-SM. Os resultados podem ser comparados na Tabela 18. O aumento do desempenho da abordagem *Gist*CMCT-SM em relação ao MCT Espacial foi de aproximadamente 9%. Com relação às regras de combinação, a regra do máximo atingiu o pior resultado novamente.

A abordagem ECMCT-SM foi superada pelo *Gist*CMCT-SM, porém a abordagem GECMCT-SM superou essa última nesta base de dados, que contém apenas cenas internas.

Tabela 18 – Comparação dos resultados dos experimentos realizados na base de dados de 67 classes de cenas internas com as estratégias combinadas e com métodos existentes na literatura. Em negrito, as abordagens propostas.

Método	Acurácia (%)
global + local (NIU; ZHOU; SHI, 2010)	40,19
GBPWHGO (ZHOU; ZHOU; HU, 2013)	40,4 ± 0,7
spatial PACT (3 níveis) (WU; REHG, 2011)	36,88 ± 1,10
MCT Espacial (GAZOLLI; SALLES, 2013)	38,58 ± 1,44
***Gist*CMCT-SM (Máximo)** (GAZOLLI; SALLES, 2013)	**40,45 ± 1,41**
***Gist*CMCT-SM (Mediana)** (GAZOLLI; SALLES, 2013)	**41,83 ± 1,22**
***Gist*CMCT-SM (Produto)** (GAZOLLI; SALLES, 2013)	**42,42 ± 1,32**
ECMCT-SM (Produto) (GAZOLLI; SALLES, 2014)	**41,98 ± 1,54**
GECMCT-SM (Produto) (GAZOLLI; SALLES, 2014)	**43,27 ± 1,38**

5.9 Considerações Finais

No Capítulo 2, levantou-se a hipótese de que a inclusão da informação de contexto no descritor MCT8 melhoraria a qualidade da representação da imagem. Os resultados obtidos nos experimento realizados com o descritor CMCT sugerem a validade dessa ideia, uma vez que, em todos as bases de dados, a classificação utilizando o CMCT superou os resultados obtidos utilizando o descritor MCT8. O mesmo aconteceu com o descritor ECMCT, que, associando novas informações ao CMCT, obteve melhores resultados que o descritor original.

Outra proposta que se mostrou promissora foi a combinação dos descritores *gist* e CMCT. Os resultados dos experimentos realizados em todas as bases foram melhores do que os resultados obtidos por cada descritor individualmente. Pode-se argumentar que a melhora nos resultados é causada pelo o aumento do tamanho do vetor de características, mas, como pode-se observar nos resultados relatados, a abordagem *gist* com ARP (LIU; KIRANYAZ; GABBOUJ, 2012) utiliza um vetor com dimensões maiores e nem por isso alcança os melhores resultados. O GECMCT também obteve melhores resultados que cada descritor individualmente, *gist* e o ECMCT, e superou, inclusive o *Gist*CMCT, indicando que as novas informações atribuídas ao CMCT ajudam a diferenciar características que não são contempladas nem pelo *gist* nem pelo ECMCT.

A combinação de classificadores implementada no *Gist*CMCT-SM também encontrou bons resultados e, mesmo havendo redundância nas informações trazidas pelos dois descritores separadamente, a combinação dos mesmos obteve melhor desempenho. É interessante notar que, nos experimentos realizados, o descritor MCT Espacial obteve melhores resultados de classificação do que o *spatial* PACT. Esse fato não surpreende, uma vez que o *spatial* PACT utiliza a transformada census, que, como foi exposto na Seção 2.3, não é capaz de capturar informações sobre determinadas estruturas, enquanto o MCT o é. Tal fato pode ser percebido no seguinte cenário: considerando apenas 2 níveis de cinza (0 e o 255, por exemplo), em uma janela 3 x 3, há 512 arranjos possíveis. O MCT mapeia 2 desses arranjos no valor 511, quando todos os pixeis são iguais à media, isto é, quando todos os bits são iguais a zero ou todos os bits são iguais a 255. A transformada census, por sua vez, mapeia 257 desses arranjos no valor 255, pois, quando o pixel central é igual a 255, os 8 bits restantes são menores ou iguais a ele e recebem valor 1 na transformada census. O mesmo acontece quando todos os pixeis são iguais a zero.

Os descritores ECMCT-SM e GECMCT-SM apresentam resultados melhores do que cada descritor que os compõem individualmente, porém, quando comparados com o descritor *Gist*CMCT-SM os resultados, em alguns casos, são inferiores ou muito próximos aos obtidos por esse último. Isso indica que, apesar do GECMCT ter um desempenho melhor do que o *Gist*CMCT, os acertos obtidos pelo primeiro coincidem com os acertos obtidos pelo MCT Espacial, fazendo com que a combinação desses dois descritores não

eleve os resultados da classificação em relação ao *Gist*CMCT-SM. Quanto ao ECMCT-SM, os resultados inferiores são esperados, uma vez que o desempenho do ECMCT não é melhor do que o do *Gist*CMCT. No entanto, aquele descritor apresenta a vantagem de não utilizar os filtros de Gabor, adotados no *gist*, apresentando, desse modo, uma complexidade computacional menor.

Vale ressaltar que, comparados com as demais técnicas, os resultados não foram muito maiores, exibindo, na maioria dos casos, um aumento inferior a 5%. Essa é uma característica dos métodos propostos nesta área. Como a classificação de cenas é um problema que tem sido explorado há bastante tempo e que apresenta grandes dificuldades, as melhorias propostas, normalmente, não provocam grandes saltos nos resultados das classificações. Contudo, apesar da pequena diferença nos resultados, o tamanho dos vetores de características empregados em algumas das técnicas existentes na literatura, tais como SPM (LAZEBNIK; SCHMID; PONCE, 2006), PLSA em cascata (ERGUL; ARICA, 2010) e CBoW (LI et al., 2011), é muito maior do que o tamanho de cada um dos vetores das técnicas combinadas.

Com relação aos classificadores utilizados nos trabalhos existentes na literatura, todos os resultados reportados, com exceção de Quattoni e Torralba (2009) e Li e Fei-Fei (2007), utilizam o SVM.

Em Li e Fei-Fei (2007), a novidade está exatamente em inferir o tipo de evento presente na imagem utilizando um modelo de grafo generativo, ou seja, um modelo que explica um conjunto de dados observados através de parâmetros não-observáveis. Eles utilizam essa estratégia porque, além de associar um evento a uma imagem, eles pretendem associar rótulos semânticos aos objetos e cenas que compõem os eventos. Já em (QUATTONI; TORRALBA, 2009), os autores utilizam protótipos de imagens, obtidos através da segmentação manual de objetos, para definir um mapeamento entre a imagem e um rótulo de cena, considerando a hipótese de que imagens com objetos similares devem ter rótulos similares e que alguns objetos na cena são mais importantes do que outros.

Assim, nesses dois casos, o foco não está na proposta de uma nova técnica de representação de características, mas no uso de modelos generativos na classificação de cenas e, ainda assim, algumas das técnicas propostas superam os resultados desses trabalhos, mesmo utilizando modelos discriminativos.

Nos outros casos, onde o classificador utilizado é o SVM, como o mesmo classificador é utilizado para todos, as diferenças são provenientes da forma como as imagens são representadas, garantindo, assim, a comparação do desempenho dos métodos de representação de características.

No Anexo A são exibidos os resultados de outras experimentos realizados com os descritores propostos.

6 Conclusões e Trabalhos Futuros

Neste trabalho, foi proposto o *Contextual Modified Census Transform* (CMCT) um descritor que modela a distribuição de estruturas locais e adiciona informação de contexto, através da inclusão de informações sobre as estruturas locais vizinhas. Esse descritor facilita a utilização dos sistemas de classificação de cenas por usuários comuns, visto que não necessita que nenhum parâmetro seja informado na construção da representação da imagem, como ocorre nos trabalhos que utilizam a técnica de *bag-of-features*. Além disso, o CMCT faz uso de cálculos simples e não exige nenhum método de agrupamento e nem realiza subdivisões sequenciais da imagem como muitos dos métodos existentes na literatura, tais como, Lazebnik, Schmid e Ponce (2006), Liu, Xu e Feng (2011), Quattoni e Torralba (2009). No entanto, apesar do CMCT superar outros métodos computacionalmente mais intensos, como exibido nas Tabelas 10, 11, 12, e 13, esse descritor apresenta resultados aquém dos métodos mais elaborados, como os presentes nas Tabelas 15 e 16. O CMCT também mantém as desvantagens apontadas pelo CENTRIST, as mais relevantes: sensibilidade à rotação e à mudança de escalas.

Para reduzir a desvantagem no desempenho da classificação em relação aos outros métodos, foi proposto o *Gist*CMCT, que explora as qualidades do descritor *gist* e obtém melhor desempenho na classificação de cenas internas e externas quando comparada com cada descritor individualmente. Nela a vantagem de não informar parâmetro algum é mantida, já que é utilizado um banco de filtros com pesos previamente estabelecidos. Porém, o desempenho computacional fica dependente do custo associado aos filtros de Gabor.

Assim, com o intuito de se obter uma representação mais eficiente de imagens, melhorando o desempenho da classificação e eliminando, ao mesmo tempo, os custos associados ao filtro de Gabor, foi proposta uma melhoria do CMCT, o CMCT Estendido (ECMCT). Este novo descritor explora informações de pixeis próximos à janela de cálculo da transformada não-paramétrica, além de fornecer dicas sobre o arranjo espacial. Ao se comparar os resultados obtidos pelo ECMCT com os obtidos pelo *Gist*CMCT, nota-se, que os desempenhos desses dois descritores são muito próximos, o que torna o ECMCT um descritor mais vantajoso do que o *Gist*CMCT, uma vez, que ele também não necessita de parâmetros e faz uso de cálculos mais simples, apresentando, portanto, uma complexidade computacional inferior ao *Gist*CMCT.

Quando Wu e Rehg (2011) propuseram o CENTRIST, eles almejavam um descritor que suprimisse as informações sobre os detalhes de textura. Os descritores CMCT e ECMCT consideram a vizinhança na representação da imagem, o que incorpora informações sobre

textura que não são contemplados pelo CENTRIST. Essa situação pode ser percebida nos experimentos realizados com a base de dados de texturas (LAZEBNIK; SCHMID; PONCE, 2005) na Seção 4.3.3. A Tabela 8 mostra que o desempenho no reconhecimento de texturas pelo ECMCT, que considera uma vizinhança maior, é superior ao do CMCT, que, por sua vez, supera o CENTRIST. Assim, pode-se concluir que, já que o CMCT e o ECMCT têm melhor desempenho na classificação do que o CENTRIST, um pouco mais de informação sobre os detalhes de textura torna a representação da imagem mais eficiente para a classificação de cenas.

Uma outra abordagem proposta neste trabalho é a utilização de classificadores múltiplos. Foi proposta a combinação do descritor MCT Espacial, que tem como ponto forte a representação da informação sobre o arranjo espacial, com os descritores ECMCT, *Gist*CMCT e GECMCT. Conforme visto no Capítulo 5, as três estratégias obtiveram resultados muito próximos, o que é um ponto favorável ao ECMCT-SM, uma vez que esse descritor apresenta uma complexidade computacional inferior em relação aos outros dois, pois não utiliza filtros de Gabor, sem que isso se reflita no seu desempenho. Vale dizer que a adoção dos classificadores múltiplos, utilizando os descritores propostos, apresentou resultados competitivos quando comparada com outros trabalhos na área.

Através dos descritores CMCT e ECMCT , pode-se verificar que as hipóteses levantadas neste trabalho foram verificadas: i) é possível extrair a informação contextual da imagem sem que para isso tenha que se apelar para o uso de técnicas que adotem representações intermediárias ou que exijam entrada de parâmetros e ii) a adição da informação contextual melhora a representação da imagem de forma que resultados competitivos de classificação sejam alcançados, mesmo quando mais de duas classes estão envolvidas. Sendo, também, atingido o objetivo deste trabalho: utilizar o contexto sem elevar demasiadamente a dimensão do vetor de características e sem recorrer a técnica de representação intermediárias *bag-of-features*.

No entanto, alguns problemas não são tratados por esses descritores como, por exemplo, a invariância à rotação. A informação sobre cor, nas imagens coloridas, também não é considerada. Assim, como trabalho futuro, propõe-se a criação de um descritor que mantenha as características do CMCT e do ECMCT, mas que considere as cores na representação da imagem. Além disso, propõe-se o uso de uma transformada não-paramétrica que seja invariante à rotação. Há também a possibilidade de se investigar a aplicação sucessiva do MCT8 de forma recursiva. No entanto, é preciso encontrar uma forma de quantizar os resultados para que não haja uma explosão do tamanho do vetor de características.

Um outro ponto a ser explorado é a combinação de classificadores, que pode ser melhorada através da inclusão de outros classificadores que utilizem representações que não sejam baseadas em transformadas não-paramétricas.

Referências

BAGGENSTOSS, P. Class-specific classifier: avoiding the curse of dimensionality. *IEEE Aerospace and Electronic Systems Magazine*, IEEE, v. 19, n. 1, p. 37–52, 2004. Citado 2 vezes nas páginas 53 e 96.

BARLA, A.; ODONE, F.; VERRI, A. Histogram intersection kernel for image classification. In: *Proceedings of 2003 International Conference on Image Processing (ICIP 2003)*. Barcelona, Espanha: [s.n.], 2003. v. 3, p. 513–516. Citado 2 vezes nas páginas 65 e 66.

BENNETT, K.; CAMPBELL, C. Support vector machines: hype or hallelujah? *SIGKDD Explorations Newsl.*, ACM, v. 2, n. 2, p. 1–13, 2000. Citado na página 64.

BHAVANI, S. et al. Illumination invariant face recognition for frontal faces using modified census transform. In: *IEEE TENCON 2007*. Taipei, Taiwan: [s.n.], 2007. p. 1–4. Citado na página 36.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. *Journal of Machine Learning Research*, JMLR. org, v. 3, n. 4-5, p. 993–1022, 2003. Citado na página 24.

BOLOVINOU, A.; PRATIKAKIS, I.; PERANTONIS, S. Bag of spatio-visual words for context inference in scene classification. *Pattern Recognition*, v. 46, n. 3, p. 1039 – 1053, 2013. Citado 2 vezes nas páginas 26 e 28.

BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. A training algorithm for optimal margin classifiers. In: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. Pittsburgh, Pennsylvania, USA: ACM, 1992. p. 144–152. Citado na página 65.

CHANG, C.-C.; LIN, C.-J. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, ACM, v. 2, n. 3, p. 1–27, maio 2011. Citado 2 vezes nas páginas 66 e 67.

CHANG, E. et al. Cbsa: content-based soft annotation for multimodal image retrieval using bayes point machines. *IEEE Transactions on Circuits and Systems for Video Technology*, IEEE, v. 13, n. 1, p. 26–38, 2003. Citado na página 21.

CHAPELLE, O.; HAFFNER, P.; VAPNIK, V. Support vector machines for histogram-based image classification. *IEEE Transactions on Neural Networks*, IEEE, v. 10, n. 5, p. 1055–1064, 1999. Citado na página 64.

COVER, T.; HART, P. Nearest neighbor pattern classification. IEEE, p. 21–27, 1967. Citado na página 43.

DOUZE, D. et al. Evaluation of gist descriptors for web-scale image search. In: *International Conference on Image and Video Retrieval*. [S.l.: s.n.], 2009. Citado na página 66.

DUDA, R. O.; HART, D. G. *Pattern Classification*. [S.l.]: Wiley-Interscience, 2001. Citado na página 42.

DUIN, R. P. W. et al. *PRTools4.1, A Matlab Toolbox for Pattern Recognition*. [S.l.], 2007. Citado na página 66.

ERGUL, E.; ARICA, N. Scene classification using spatial pyramid of latent topics. In: *Proceedings of the 2010 20th International Conference on Pattern Recognition*. Istambul, Turquia: [s.n.], 2010. (ICPR '10), p. 3603–3606. Citado 4 vezes nas páginas 25, 79, 80 e 85.

FEI-FEI, L.; PERONA, P. A bayesian hierarchical model for learning natural scene categories. In: *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. San Diego, EUA: [s.n.], 2005. p. 524–531. Citado na página 24.

FOGGIA, P. et al. Evaluating classification reliability for combining classifiers. In: *14th International Conference on Image Analysis and Processing*. [S.l.: s.n.], 2007. p. 711–716. Citado na página 52.

FREDEMBACH, C.; SCHRODER, M.; SUSSTRUNK, S. Eigenregions for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 26, n. 12, p. 1645–1649, 2004. Citado na página 22.

FROBA, B.; ERNST, A. Face detection with the modified census transform. In: *Proceedings of the Sixth IEEE international conference on Automatic face and gesture recognition*. Seoul, Coreia: [s.n.], 2004. p. 91–96. Citado 3 vezes nas páginas 32, 36 e 51.

GAZOLLI, K.; SALLES, E. A contextual image descriptor for scene classification. In: *Online Proceedings on Trends in Innovative Computing*. [S.l.: s.n.], 2012. p. 66–71. Citado 4 vezes nas páginas 72, 74, 75 e 76.

GAZOLLI, K.; SALLES, E. Using holistic features for scene classification by combining classifiers. *Journal of WSCG*, v. 21, n. 1, p. 41–48, 2013. Citado 4 vezes nas páginas 80, 81, 82 e 83.

GAZOLLI, K.; SALLES, E. Exploring neighborhood and spatial information for improving scene classification. *Pattern Recogniton Letters*, v. 46, p. 83–88, 2014. Citado 12 vezes nas páginas 8, 55, 56, 58, 72, 74, 75, 76, 80, 81, 82 e 83.

GORKANI, M.; PICARD, R. Texture orientation for sorting photos "at a glance". In: *Pattern Recognition, 1994. Vol. 1 - Conference A: Computer Vision amp; Image Processing., Proceedings of the 12th IAPR International Conference on*. [S.l.: s.n.], 1994. v. 1, p. 459–464 vol.1. Citado na página 21.

GRAUMAN, K.; DARRELL, T. The pyramid match kernel: Efficient learning with sets of features. *J. Mach. Learn. Res.*, JMLR.org, v. 8, p. 725–760, 2007. Citado na página 25.

HO, T. K.; HULL, J. J.; SRIHARI, S. N. Decision combination in multiple classifier systems. *IEEE Trans. Pattern Anal. Mach. Intell.*, IEEE, v. 16, p. 66–75, 1994. Citado na página 51.

HOFFMAN, T. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, Springer, v. 42, p. 177–196, 2001. Citado na página 25.

JAIN, A.; DUIN, R. P. W.; MAO, J. Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 22, n. 1, p. 4–37, 2000. Citado na página 65.

KITTLER, J. et al. On combining classifiers. *IEEE Trans. Pattern Anal. Mach. Intell.*, IEEEy, v. 20, n. 3, p. 226–239, 1998. Citado na página 52.

KNERR, S.; PERSONNAZ, L.; DREYFUS, G. Single-layer learning revisited: a stepwise procedure for building and training a neural network. In: *Neurocomputing*. [S.l.]: Springer Berlin Heidelberg, 1990. v. 68, p. 41–50. Citado na página 67.

KULLBACK, S.; LEIBLER, R. A. On information and sufficiency. *Annals of Mathematical Statistics*, Institute of Mathematical Statistics, v. 22, 1951. Citado na página 26.

LAZEBNIK, S.; SCHMID, C.; PONCE, J. A sparse texture representation using local affine regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 27, n. 8, p. 1265–1278, 2005. Citado 3 vezes nas páginas 58, 87 e 98.

LAZEBNIK, S.; SCHMID, C.; PONCE, J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Nova York, EUA: [s.n.], 2006. (CVPR '06), p. 2169–2178. Citado 20 vezes nas páginas 7, 8, 9, 20, 21, 25, 39, 41, 43, 49, 50, 58, 67, 68, 71, 72, 79, 80, 85 e 86.

LI, L.-J.; FEI-FEI, L. What, where and who? classifying events by scene and object recognition. In: *IEEE 11th International Conference on Computer Vision*. [S.l.: s.n.], 2007. p. 1–8. Citado 7 vezes nas páginas 8, 67, 69, 74, 75, 78 e 85.

LI, T. et al. Contextual bag-of-words for visual categorization. *IEEE Transactions on Circuits and Systems for Video Technology,*, v. 21, n. 4, p. 381 –392, 2011. Citado 4 vezes nas páginas 26, 28, 80 e 85.

LI, Z.; DEWEN, H. Scene classification combining low-level and semantic modeling strategies. In: *Second International Conference oN Digital Manufacturing and Automation*. [S.l.: s.n.], 2011. p. 1071 –1075. Citado na página 80.

LIU, S.; XU, D.; FENG, S. Region contextual visual words for scene categorization. *Expert Systems with Applications*, Pergamon-Elsevier Science LTD, v. 38, n. 9, p. 11591–11597, 2011. Citado 6 vezes nas páginas 26, 28, 64, 72, 80 e 86.

LIU, W.; KIRANYAZ, S.; GABBOUJ, M. Robust scene classification by gist with angular radial partitioning. In: *5th International Symposium onCommunications Control and Signal Processing (ISCCSP)*. [S.l.: s.n.], 2012. p. 2–4. Citado 4 vezes nas páginas 24, 76, 78 e 84.

LOWE, D. G. Object recognition from local scale-invariant features. In: *Proceedings of the International Conference on Computer Vision*. [S.l.: s.n.], 1999. (ICCV '99), p. 1150–1157. Citado na página 24.

LUO, J.; SAVAKIS, A. E.; SINGHAL, A. A bayesian network-based framework for semantic image understanding. *Pattern Recognition*, Elsevier Science, v. 38, n. 6, p. 919–934, 2005. Citado 2 vezes nas páginas 7 e 23.

MARCEL, S.; RODRIGUEZ, Y.; HEUSCH, G. On the recent use of local binary patterns for face authentication. *International Journal on Image and Video Processing Special Issue on Facial Image Processing*, IDIAP, 2007. Citado na página 32.

MENG, X.; WANG, Z. Rapid scene categorization using novel gist model. In: *2nd International Conference on Information Engineering and Computer Science (ICIECS)*. [S.l.: s.n.], 2010. p. 1–4. Citado na página 81.

MENG, X.; WANG, Z.; WU, L. Building global image features for scene recognition. *Pattern Recogn.*, Elsevier Science Inc., v. 45, p. 373–380, 2012. Citado 2 vezes nas páginas 27 e 80.

NIU, Z.; ZHOU, Y.; SHI, K. A hybrid image representation for indoor scene classification. In: *25th International Conference of Image and Vision Computing New Zealand*. [S.l.: s.n.], 2010. p. 1 –7. Citado na página 83.

NOVOVICOVÄ, J.; MALíK, A.; PUDIL, P. Feature selection using improved mutual information for text classification. In: *Structural, Syntactic, and Statistical Pattern Recognition*. [S.l.]: Springer Berlin Heidelberg, 2004. v. 3138, p. 1010–1017. Citado na página 65.

OHTA, Y.; KANADE, T.; SAKAI, T. Color information for region segmentation. *Computer Graphics and Image Processing*, Elsevier Science Inc., v. 13, n. 3, p. 222 – 241, 1980. Citado na página 22.

OJALA, T.; PIETIKÄINEN, M.; HARWOOD, D. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, Elsevier Science Inc., v. 29, n. 1, p. 51–59, 1996. Citado na página 32.

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 (7), p. 971–987, 2002. Citado na página 56.

OJANSIVU, V.; HEIKKILÄ, J. Blur insensitive texture classification using local phase quantization. Springer Berlin Heidelberg, v. 5099, p. 236–243, 2008. Citado na página 59.

OLIVA, A.; TORRALBA, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vision*, Kluwer Academic Publishers, v. 42, n. 3, p. 145–175, 2001. Citado 5 vezes nas páginas 23, 46, 67, 72 e 81.

OLIVA, A.; TORRALBA, A. Building the gist of a scene: The role of global image features in recognition. In: MARTINEZ-CONDE, S. et al. (Ed.). *Visual Perception Fundamentals of Awareness: Multi-Sensory Integration and High-Order Perception*. [S.l.]: Elsevier Science Inc., 2006, (Progress in Brain Research, v. 155). p. 23–36. Citado 3 vezes nas páginas 23, 46 e 47.

POTTER, M. C. Meaning in visual search. *Science*, Elsevier Science Inc., v. 187, p. 965–966, 1975. Citado na página 23.

QIN, J.; YUNG, N. Scene categorization via contextual visual words. *Pattern Recognition*, Elsevier Science Inc., v. 43, n. 5, p. 1874–1888, 2010. Citado 4 vezes nas páginas 7, 25, 26 e 28.

QUATTONI, A.; TORRALBA, A. Recognizing indoor scenes. In: *Proceedings IEEE CS Conf. Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2009. p. 413–420. Citado 5 vezes nas páginas 68, 69, 75, 85 e 86.

QUELHAS, P. et al. Modeling scenes with local descriptors and latent aspects. In: *Proceedings of the Tenth IEEE International Conference on Computer Vision*. Beijing, China: [s.n.], 2005. (ICCV '05), p. 883–890. Citado na página 25.

QUELHAS, P. et al. A thousand words in a scene. *IEEE Trans. Pattern Anal. Mach. Intell.*, IEEE, v. 29, n. 9, p. 1575–1589, 2007. Citado na página 25.

SALTON, G.; MCGILL, M. J. *Introduction to modern information retrieval*. [S.l.]: New York: McGraw-Hill, 1983. Citado na página 70.

SHEN, J.; SHEPHERD, J.; NGU, A. H. H. Semantic-sensitive classification for large image libraries. In: *Proceedings of the 11th International Multimedia Modelling Conference*. [S.l.: s.n.], 2005. p. 340–345. Citado na página 21.

SONG, T.; LI, H. Wavelbp based hierarchical features for image classification. *Pattern Recogn. Lett.*, Elsevier Science Inc., v. 34, n. 12, p. 1323–1328, 2013. Citado na página 80.

SZUMMER, M.; PICARD, R. Indoor-outdoor image classification. In: *IEEE International Workshop on Content-Based Access of Image and Video Database*. [S.l.: s.n.], 1998. p. 42–51. Citado 2 vezes nas páginas 18 e 22.

VAILAYA, A. et al. Content-based hierarchical classification of vacation images. In: *IEEE International Conference on Multimedia Computing and Systems*. [S.l.: s.n.], 1999. v. 1, p. 518–523. Citado na página 21.

VAPNIK, V. The support vector method of function estimation. In: SUYKENS, J.; VANDEWALLE, J. (Ed.). *Nonlinear Modeling*. [S.l.]: Springer US, 1998. p. 55–85. Citado 2 vezes nas páginas 52 e 64.

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 22.

WEBB, A. R. *Statistical Pattern Recognition, 2nd Edition*. [S.l.]: John Wiley & Sons, 2002. Citado na página 28.

WU, J.; REHG, J. M. Where am i: Place instance and category recognition using spatial pact. In: *IEEE Conference on Computer Vision and Pattern Recognition (2008)*. Alaska, USA: [s.n.], 2008. p. 1–8. Citado na página 18.

WU, J.; REHG, J. M. Beyond the euclidean distance: Creating effective visual codebooks using the histogram intersection kernel. In: *IEEE 12th International Conference on Computer Vision*. Kyoto, Japão: [s.n.], 2009. p. 630–637. Citado na página 66.

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 20 vezes nas páginas 7, 27, 31, 32, 33, 43, 47, 48, 50, 51, 56, 67, 69, 72, 79, 80, 81, 82, 83 e 86.

WU, T.-F.; LIN, C.-J.; WENG, R. C. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, MIT Press, v. 5, p. 975–1005, 2004. Citado na página 67.

ZABIH, R.; WOODFILL, J. Non-parametric local transforms for computing visual correspondence. In: *Computer Vision ECCV '94*. [S.l.]: Springer Berlin Heidelberg, 1994. v. 801, p. 151–158. Citado 3 vezes nas páginas 27, 31 e 32.

ZHANG, H. et al. Svm-knn: Discriminative nearest neighbor classification for visual category recognition. In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2006. v. 2, p. 2126–2136. Citado na página 43.

ZHOU, L.; ZHOU, Z.; HU, D. Scene classification using a multi-resolution bag-of-features model. *Pattern Recogn.*, Elsevier Science Inc., v. 46, n. 1, p. 424–433, 2013. Citado 4 vezes nas páginas 80, 81, 82 e 83.

Anexos

ANEXO A – Outros Resultados

A.1 GistCMCT e Espacial MCT com PCA

Na Seção 3.3, optou-se por utilizar a estratégia de classificadores múltiplos com o intuito de não aumentar as dimensões finais do vetor de classificação. Uma outra abordagem possível seria a concatenação do vetor GistCMCT com o vetores MCT para cada subdivisão da pirâmide e a aplicação do PCA para a redução da dimensionalidade. Isto posto, nesta seção são apresentados os resultados obtidos para essa estratégia alternativa nas bases de dados de 8 e 15 categorias de cenas e 8 classes de eventos de esportes. O número de componentes adotado para cada base de dados foi variado e selecionados através do nível de energia dos autovetores.

Tabela 19 – Resultados da classificação utilizando-se os descritores GistCMCT e Espacial MCT com PCA.

Base de Dados	**Acurácia(%)**
8 Categorias de Cenas	88,50 ± 0,58
15 Categorias de Cenas	85,63 ± 0,62
8 Classes de Eventos de esporte	80,67 ± 1,24

Ao se comparar os resultados com os obtidos através do uso da estratégia de classificadores múltiplos (Tabelas 15, 16 e 17), nota-se que os valores foram alcançados foram próximos, porém um pouco menores do que os da estratégia proposta.

A.2 Estratégias Combinadas

Nesta seção, são apresentados os resultados de classificação obtidos utilizando-se os descritores ECMCT-SM e GECMCT-SM com as regras de combinação Máximo e Mediana.

Tabela 20 – Resultados utilizando-se os descritores ECMCT-SM e GECMCT-SM com as regras de combinação Máximo e Mediana para base de dados 15 categorias de cenas.

Método	**Acurácia(%)**
ECMCT-SM (Máximo)	85,35 ± 0,70
ECMCT-SM (Mediana)	85,70 ± 0,70
GECMCT-SM (Máximo)	86,53 ± 0,48
GECMCT-SM (Mediana)	86,68 ± 0,56

Comparando-se os resultados apresentados com os das Tabelas 15, 16, 17 e 18, verifica-se que os valores para as três técnica de combinação utilizadas são muito próximos,

Tabela 21 – Resultados utilizando-se os descritores ECMCT-SM e GECMCT-SM com as regras de combinação Máximo e Mediana para base de dados 8 categorias de cenas.

Método	**Acurácia(%)**
ECMCT-SM (Máximo)	88,33 ± 0,54
ECMCT-SM (Mediana)	88,45 ± 0,78
GECMCT-SM (Máximo)	88,92 ± 0,31
GECMCT-SM (Mediana)	88,96 ± 0,46

Tabela 22 – Resultados utilizando-se os descritores ECMCT-SM e GECMCT-SM com as regras de combinação Máximo e Mediana para base de dados 8 eventos de esporte.

Método	**Acurácia(%)**
ECMCT-SM (Máximo)	81,63 ± 1,47
ECMCT-SM (Mediana)	82,21 ± 1,99
GECMCT-SM (Máximo)	80,96 ± 1,29
GECMCT-SM (Mediana)	81,54 ± 1,73

Tabela 23 – Resultados utilizando-se os descritores ECMCT-SM e GECMCT-SM com as regras de combinação Máximo e Mediana para base de dados 67 classes de cenas internas.

Método	**Acurácia(%)**
ECMCT-SM (Máximo)	39,76 ± 1,84
ECMCT-SM (Mediana)	41,61 ± 1,54
GECMCT-SM (Máximo)	40,28 ± 1,42
GECMCT-SM (Mediana)	42,85 ± 1,42

porém um pouco melhores para a regra de combinação Produto, exceto na base de dados 8 eventos de esportes, onde a regra de combinação Mediana se saiu um pouco melhor.

A.3 Validação Cruzada

Nesta seção, são apresentados os resultados utilizando-se, para a partição dos dados, a técnica de validação cruzada k-*fold*, com k = 5, para todos os descritores propostos. Assim, as amostras foram divididas em 5 subconjuntos de tamanhos iguais, sendo que um subconjunto foi utilizado para teste e os outros 4, para treino. Esse procedimento foi realizado 5 vezes, com cada subconjunto sendo utilizado exatamente uma vez como teste.

Nota-se que os resultados obtidos através do uso da validação cruzada são superiores aos apresentados no Capítulo 5. Isso ocorre devido ao tamanho dos vetores de características, pois de acordo com a "maldição da dimensionalidade" (BAGGENSTOSS, 2004), aumentar o tamanho do vetor de características, mantendo o mesmo número de amostras, pode

Tabela 24 – Resultados da validação cruzada na base de dados 15 categorias de cenas.

Método	**Acurácia(%)**
CMCT	80,33 ± 0,58
ECMCT	85 ± 1,65
*Gist*CMCT	86,79 ± 0,75
GECMCT	87,47 ± 0,77
*Gist*CMCT-SM (Produto)	89,14 ± 0,45
ECMCT-SM (Produto)	88,56 ± 0,53
GECMCT-SM (Produto)	89,34 ± 0,61

Tabela 25 – Resultados da validação cruzada na base de dados 8 categorias de cenas.

Método	**Acurácia(%)**
CMCT	84,62 ± 1,34
ECMCT	88,34 ± 1,38
*Gist*CMCT	89,58 ± 1,08
GECMCT	89,83 ± 1,50
*Gist*CMCT-SM (Produto)	90,91 ± 1,35
ECMCT-SM (Produto)	90,54 ± 1,49
GECMCT-SM (Produto)	91,40 ± 1,54

Tabela 26 – Resultados da validação cruzada na base de dados 8 eventos de esporte.

Método	**Acurácia(%)**
CMCT	70,85 ± 1,68
ECMCT	82,06 ± 1,95
*Gist*CMCT	80,94 ± 1,95
GECMCT	83,09 ± 2,24
*Gist*CMCT-SM (Produto)	86,18 ± 1,51
ECMCT-SM (Produto)	87,06 ± 2,31
GECMCT-SM (Produto)	86,75 ± 2,52

levar a uma piora nos resultados de classificação. Isso acontece porque o número maior de combinação de variáveis explode exponencialmente, o que causa um aumento na esparsidade dos dados, quando poucos exemplos são apresentados. Nesse caso, existe um risco de *over-fitting* dos dados de treinamento, o que implica em uma generalização fraca para classificar novos exemplos. Assim, como na validação cruzada um número maior de exemplos de treino são apresentados, os resultados obtidos são melhores. No entanto, a alta dimensionalidade está presente na maioria dos métodos de classificação, e, mesmo com essa desvantagem, os métodos propostos neste trabalho alcançaram resultados competitivos.

Vale ressaltar que o procedimento de divisão de dados adotado no Capítulo 5 é o utilizado em todos os métodos referenciados e, portanto, haveria uma comparação desigual, caso se aplicasse a validação cruzada para testar a eficiência dos descritores propostos.

A.4 Base de Dados de Texturas

Nesta seção, são apresentados os resultados de classificação obtidos na base de dados de textura (LAZEBNIK; SCHMID; PONCE, 2005), utilizando-se o classificador SMV e os descritores CENTRIST, MCT8, CMCT, ECMCT.

Tabela 27 – Resultados de classificação utilizando-se os descritores MCT8, CMCT e ECMCT na base de dados de textura.

Método	**Acurácia(%)**
CENTRIST	70,6 ± 0,71
MCT8	83,24 ± 1,03
CMCT	87,96 ± 1,02
ECMCT	90,92 ± 0,97

Nota-se que, quando comparado ao CENTRIST, o ECMCT tem uma taxa de acerto aproximadamente 29% superior na classificação das texturas presentes nesta base.