2
Trabalhos Relacionados

2.1.
Introdução

Para a realização desta pesquisa foi feito um estudo extenso de diversos trabalhos e artigos acadêmicos relativos ao reconhecimento de objetos, preferencialmente em tempo real. Como resultado, esta dissertação recomenda uma série de trabalhos de escopo mais geral (Dockstader e Tekalp (2001), Han et al. (2004), Lannizzotto e Vita (2002), Kang et al. (2004), Molineros e Sharma (2001), Orwell et al. (2001), Kin e Kehtarnavaz (2005)) e outros específicos de vigilância (Amer (2003), Bramberger et al. (2004), Desurmont et al. (2004), Ziliani e Cavallaro (2001)), sem contudo entrar em detalhes. Por fim, esta dissertação termina por selecionar alguns poucos trabalhos considerados mais fortemente relacionados com a presente proposta e os apresenta neste capítulo. Estes últimos trabalhos também são agrupados em quatro classes, para se ter uma melhor visão crítica: *templates* de objetos; sistema de coordenadas; histogramas; e aprendizado de máquina e classificadores.

Neste capítulo, são apresentados os funcionamentos das principais técnicas dos trabalhos mencionados acima que motivaram o desenvolvimento da presente pesquisa. Ao longo de cada análise, são levantadas de uma maneira geral as vantagens e desvantagens de cada técnica apresentada, assim como as carências que ainda devem ser supridas.

2.2.
Template de Objetos

Alguns pesquisadores têm obtido bons resultados no reconhecimento e detecção de objetos em tempo real utilizando algoritmos baseados em gabaritos (*templates*) de objetos, ou seja, formas e padrões que os objetos possuem em comum.

Dentre tais resultados, destaca-se o trabalho realizado por Gavrila e Philomin (1999), onde é discutida a detecção de objetos em tempo real para carros “inteligentes”, capazes de detectar pedestres e sinalização de trânsito na

estrada. No trabalho desenvolvido por estes autores, é proposto um método baseado em *templates* de objetos usando-se Transformações de Distância (*Distance Transforms*), que consiste em envolver duas imagens binárias: um *template* segmentado *T* e uma imagem segmentada *I*, que são chamadas respectivamente de "*Template* de características" e de "Imagem de características".

Nestas imagens binárias, têm-se pixels "ligados" que denotam a existência de características, e pixels "desligados" que indicam a ausência das mesmas. Entretanto, para este método de combinação entre imagens, não importa o quê tais características significam, pois normalmente usam-se apenas pontos das bordas e dos cantos de cada *template*. A forma binária do *template* de características deve ser fornecida previamente para a aplicação, enquanto a imagem de características é gerada a partir do processamento da imagem de interesse, através da extração de tais características.

Para combinar *T* e *I* é necessário computar as Transformações de Distância (TD) da imagem de características *I*. O *template T* é então transformado (transladado, rotacionado ou escalado) e posicionado sobre a imagem TD resultante de *I*, onde a combinação da medida $D(T,I)$ é determinada pelos valores dos pixels da imagem TD que fica sob os pixels "ligados" do *template* transformado. Estes valores de pixels formam uma distribuição de distâncias do *template* de características para as características mais próximas da imagem. Quanto menores forem tais distâncias, melhor é a combinação entre o *template* e a imagem numa dada região de interesse. A figura 2 ilustra o processo de combinação de imagens.

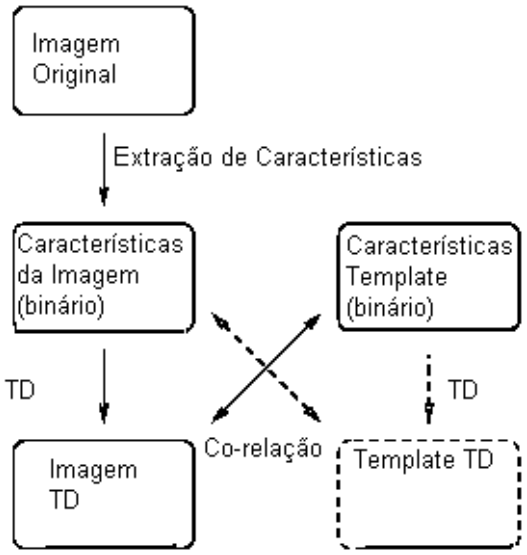
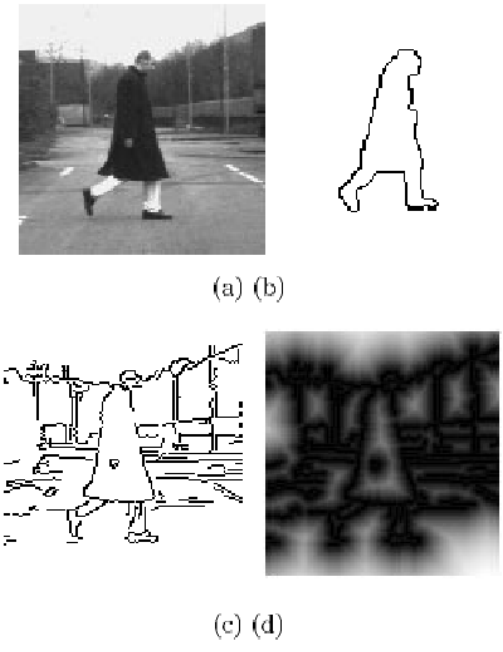


Figura 2. Combinação de Imagens usando Transformações de Distância

Na prática, um *template* é considerado como semelhante a regiões de uma determinada imagem se a medida de distância $D(T,I)$ for menor que um limiar θ fornecido pelo usuário. A figura 3 ilustra o esquema de comparação da figura 2 para o caso típico de características de bordas. A figura 3a-b mostra um exemplo de uma imagem e um *template* pré-definido. A figura 3c-d mostra a detecção de bordas e a transformação TD das bordas da imagem respectivamente. Pode-se notar que as distâncias na imagem TD estão codificadas pela intensidade de cores, onde as cores mais claras mostram valores de distância maiores e as mais escuras mostram distâncias menores.

Com a finalidade de se obter o reconhecimento de objetos em tempo real, Gavrila e Philomin (1999) propõem uma hierarquia de *templates* que permitam diminuir o número de testes feito para combinar regiões de interesse da imagem com o *template* em questão. A figura 4 mostra um exemplo de hierarquia de *template* de pedestres, onde se pode notar que quanto mais baixo o nível da hierarquia, mais semelhantes são as formas entre si.



(a) (b)

(c) (d)

Figura 3. Comparação usando características de bordas (a) original (b) *template* (c) bordas (d) imagem TD (extraído de Gavrila e Philomin (1999))

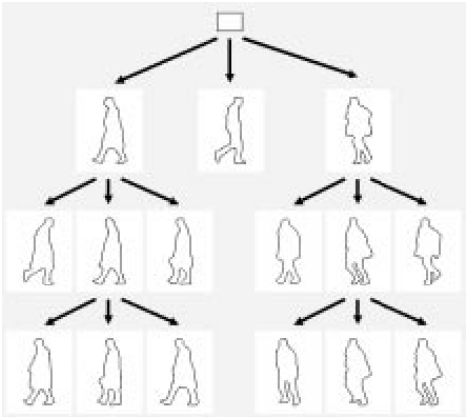


Figura 4. Hierarquia de *templates* de um pedestre (visão parcial) (extraído de Gavrila e Philomin (1999))

A técnica apresentada por Gavrila e Philomin (1999), apesar de ser capaz de detectar objetos em tempo real, apresenta um número elevado de falsos positivos, devido a um grande número de objetos similares a pedestres ou placas de trânsito. Além disto, a técnica também não é capaz de detectar objetos que estejam muito próximos ou distantes da câmera. O fato do algoritmo de comparação utilizar contornos e bordas das imagens impossibilita uma boa confiabilidade para a detecção de objetos mais complexos ou com muitas bordas, tais como árvores.

2.3. Sistema de Coordenadas

Ainda no problema específico de detecção de objetos relacionados com o trânsito, como pedestres, Guanglin et al. (2007) lançam um novo olhar sobre as técnicas existentes e apresentam uma solução capaz de detectar tais objetos em tempo real, com apenas uma única câmera monocromática instalada dentro de um carro. A solução proposta por estes autores é capaz de detectar tanto objetos parados quanto objetos em movimento. Para realizar tal tarefa, Guanglin et al. (2007) propõem três etapas: 1 - o algoritmo inicialmente detecta objetos que estão sobre o plano do chão realizando a simulação de um sistema estéreo “virtual”, através do uso do mapeamento inverso da perspectiva. 2 - Em seguida, um algoritmo de segmentação de imagem é executado a fim de encontrar pedestres em potencial, desenhando uma caixa envolvente para os objetos. 3 - É apresentado um algoritmo rápido para estabilização da imagem devido a ruídos causados por possíveis desníveis na estrada e movimentos do próprio

carro que podem fazer a câmera oscilar. Tais pontos da solução proposta são explicados mais detalhadamente a seguir.

O princípio do mapeamento inverso da perspectiva é o de remover o efeito da perspectiva quando os parâmetros da câmera (posição, orientação, distância focal, etc.) são conhecidos e quando uma boa suposição sobre a estrada pode ser feita (*e.g.* supor que a estrada é plana).

A idéia do algoritmo proposto por Guanglin et al. (2007) é ilustrada na figura 5, onde a visão lateral do plano da imagem é mostrada em dois instantes distintos t_0 e t_1. O ponto C_0 é o centro ótico da câmera no instante t_0 onde um frame da imagem é tirado, e C_1 é o ponto correspondente no *frame* tirado no instante t1. Dentro do espaço de tempo $\Delta t = t1 - t0$ a câmera move-se do ponto c_0 ao ponto c_1 com a velocidade *v* e taxa angular ω.

No exemplo da figura 5, através de uso de matrizes de transformação, é possível verificar, por exemplo, que no instante t_0 o ponto P_G^3 não é "visível" pela câmera, enquanto que no instante t_1 ele consegue ser visualizado, o que significa que o ponto P_O^1 é um obstáculo e pode ser entendido como um objeto em potencial.

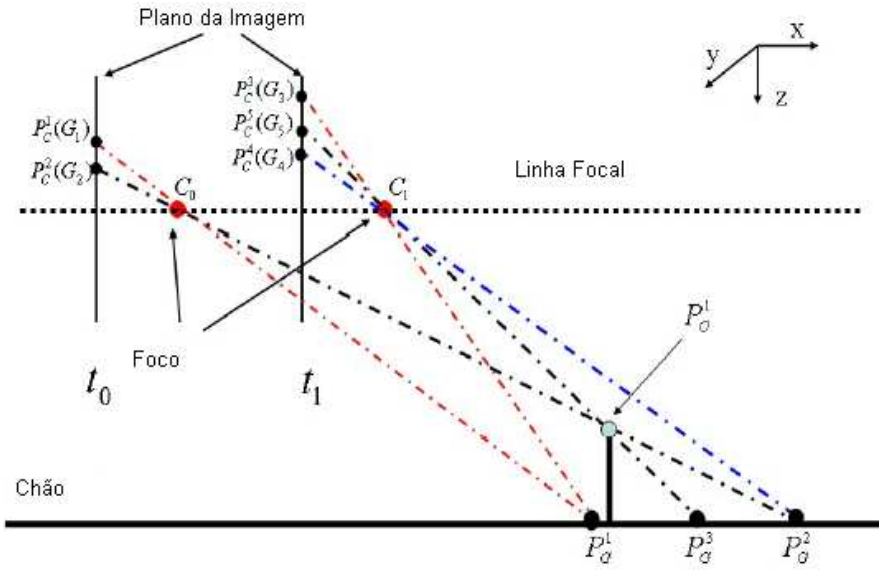


Figura 5. Ilustração do princípio de detecção por sistema de coordenadas (extraído de Guanglin et al. (2007))

Tendo-se como identificar objetos em potencial, Guanglin et al. (2007) propuseram uma heurística capaz de detectar com bastante eficiência se um objeto é um pedestre ou não, batizada de "Pedestrian Detection Strip" (PDS).

Através de observações realizadas durante os experimentos, estes autores concluem que pedestres, na grande maioria das vezes, estão próximos da linha do horizonte da estrada. Isto faz com que a área de procura da imagem possa ser substancialmente reduzida para uma região de trinta pixels de altura a partir do horizonte para baixo. A altura de trinta pixels é estabelecida por ser suficiente para encontrar um pedestre a uma distância de até cinqüenta metros de distância. Tal funcionamento está ilustrado na figura 6.

Entretanto, pequenas crianças próximas ao veículo não seriam detectadas apenas utilizando a heurística do PDS, pois somente a área próxima ao horizonte seria analisada. Para tanto, é necessário utilizar um segundo PDS, só que desta vez analisando apenas a área inferior da imagem, a fim de detectar tais crianças ou demais pedestres menores.

A técnica proposta por Guanglin et al. (2007) consegue detectar muito rapidamente pedestres na cena, a uma taxa de 64 *frames* por segundo. Porém, há um alto número de ocorrências de falsos positivos, principalmente de postes. Guanglin et al. (2007) alegam que não há problema, pois um algoritmo de classificação subseqüente poderia ser utilizado para reanálise das áreas selecionadas como pedestres. Entretanto, isto acarretaria numa perda de desempenho da detecção e faria com que a solução proposta deixasse de ser em tempo real. Além disto, a solução é focada exclusivamente para a detecção de pedestres. Se for necessário detectar outros tipos de objetos (e.g. carros), devem ser criadas novas heurísticas e funções de classificação.

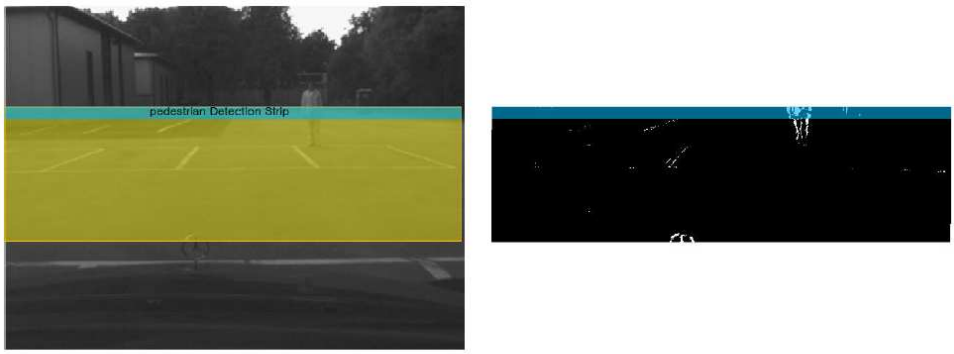


Figura 6. Detecção através do método “Pedestrian Detection Strip” (extraído de Guanglin et al. (2007))

2.4. Análise de Histogramas

Outra técnica utilizada para se detectar objetos ou ainda classes de cenários é através da análise de histogramas de regiões de uma imagem digital. Vários pesquisadores de visão computacional, entre eles Bosch et al. (2006), Fei-Fei e Perona (2005) e Torralba e Oliva (2001), têm estudado o reconhecimento de cenários principalmente para a área de robótica, onde é necessário saber com freqüência em que "tipo" de local o observador encontra-se (e.g. se está em um escritório ou em um quarto). Por muitas vezes, é necessário se saber inclusive em que cenário exato o observador encontra-se, ao invés de se saber apenas o "tipo" do cenário.

Na pesquisa realizada por Wu e Rehg (2008), é proposta uma nova técnica chamada de sPACT (Spatial Principal component Analysis of Census Transform histograms), que em comparação a técnicas utilizadas por outros pesquisadores, consegue um melhor desempenho no reconhecimento de cenários, possui menos parâmetros a serem balanceados, apresenta alta velocidade de computação (> 50 fps) e é fácil de implementar.

Na técnica sPACT, calcula-se inicialmente a transformada de Censo (*CT – Census Transform*) da imagem, de acordo com a equação (1-1):

$$CT(i) = \underset{j \in N(i)}{\otimes} \xi(I(i), I(j)) \qquad (1\text{-}1)$$

que gera um *string* de bits em uma ordem qualquer, onde ⊗ é o operador de concatenação, $N(i)$ é a vizinhança espacial local do pixel i (*e.g.* os 8 pixels ao redor do pixel i) e $\xi(I(i),I(j))$ é uma função de comparação que é igual a 1 se $I(i) < I(j)$ e 0 caso contrário. Por exemplo:

$$\begin{array}{|c|c|c|} \hline 32 & 64 & 96 \\ \hline 32 & \mathbf{64} & 96 \\ \hline 32 & 32 & 96 \\ \hline \end{array} \Rightarrow \begin{array}{ccc} 1 & 1 & 0 \\ 1 & & 0 \\ 1 & 1 & 0 \end{array}$$

Figura 7. Ilustração da comparação do pixel central com seus vizinhos

onde *CT* = 11010110 ou *CT* = 214, que é o *string* na base decimal.

Assim como outras transformadas locais não-parametrizadas, a *CT* é robusta quanto a alterações de luminosidade, cor, variações de fator *gamma*, etc.

A partir do cálculo da *CT* de cada pixel, gera-se outra imagem (chamada de imagem *CT*) com o valor calculado para cada pixel (Figura 7). Nesta nova

imagem transformada, são então armazenadas informações globais (especialmente as descontinuidades da cena) e locais. Numa imagem *CT*, há várias regras implícitas nos valores de cada pixel que pode ser utilizado a fim de se obterem maiores informações. Por exemplo, para dois pixels vizinhos *(x,y)* e *(x,y+1)*, o bit 5 de *CT(x,y)* e o bit 4 de *CT(x,y+1)* são sempre complementares entre si (Figura 8).

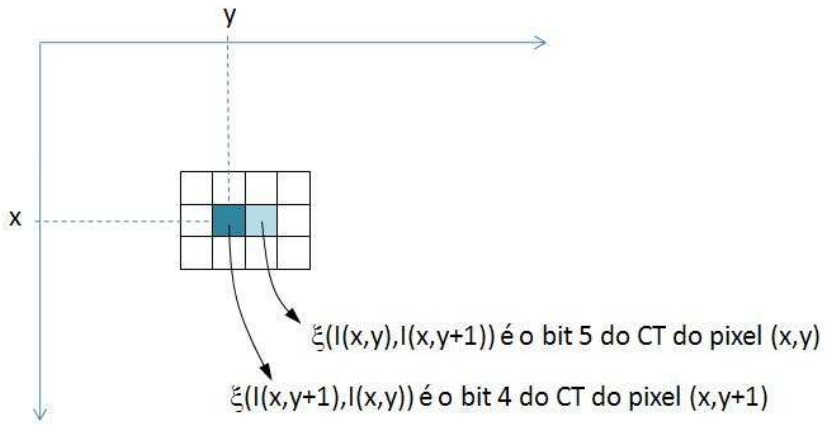


Figura 8. *CT*s para pixels vizinhos (bit 5 da *CT(x,y)* e bit 4 da *CT(x,y+1)* são complementares

Devido a tais informações implícitas na imagem, o seu histograma também contém informações importantes, podendo-se observar que formas de cenários e objetos semelhantes possuem também histogramas *CT* semelhantes.

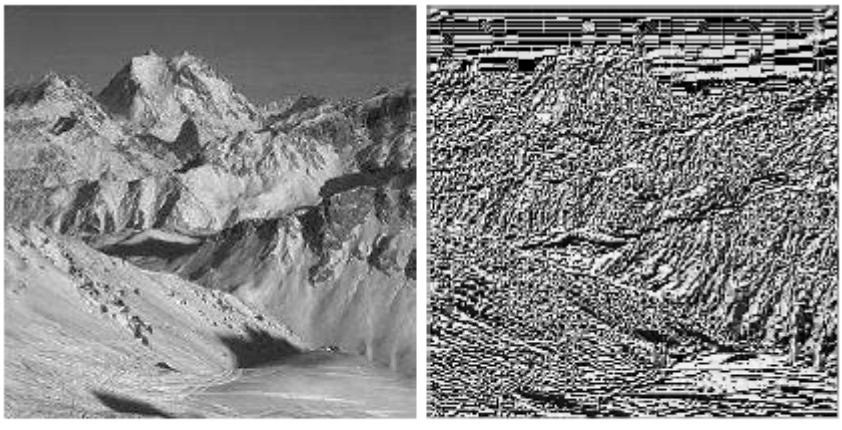


Figura 9. Exemplo de imagem CT (Census Transform) (extraído de Wu e Rehg (2008))

Por fim é utilizado o conceito de “pirâmide espacial”, onde se divide a imagem em sub-regiões menores (blocos) e em 3 níveis de resolução diferentes,

conforme ilustra a figura 10. Esta suddivisão é uma adaptação da idéia de casamento de pirâmide espacial (Spatial Pyramid Matching) de Lazebnik et al. (2006), que mostraram ser esta pirâmide uma estratégia eficiente para o aumento da capacidade de reconhecimento de objetos. Em seguida, para cada bloco, calculam-se histogramas de *CT* e removem-se efeitos de correlação com PCA (Principal Component Analysis) – o que nomeia a representação proposta por Wu e Rehg (2008) de sPACT (Spatial Principal component Analysis of Census Transform histograms). PACT em todos os blocos são então concatenados para formar um único vetor geral de características. Por exemplo, com 40 autovetores usados na análise, a pirâmide até o nível 2 da figura 10 gera um vetor com $40 \times (25 + 5 + 1) = 1240$ dimensões. Depois que os vetores de características PACT são extraídos das imagens, escolhem-se diferentes classificadores (e.g. Nearest Neighbor classifier, SVM classifier) para reconhecimento de objetos, dependendo da situação.

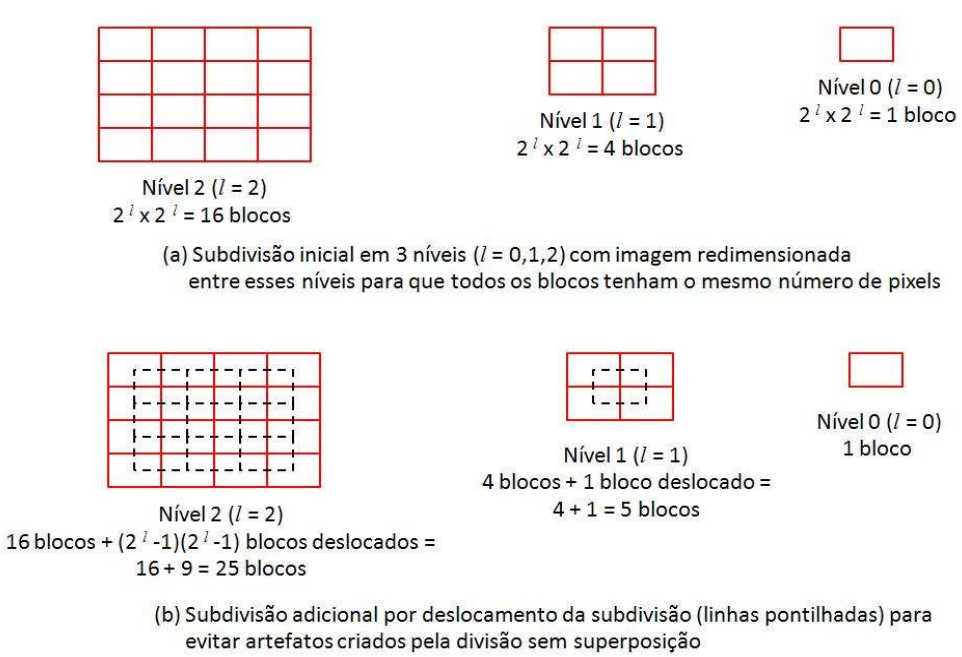


Figura 10. Pirâmide espacial de uma imagem com 3 níveis

Apesar de a técnica sPACT possuir baixo tempo de computação, a comparação com grandes números de histogramas a torna inviável. Portanto, o conjunto de histogramas para comparação e análise deve possuir um tamanho pequeno a fim de não prejudicar o desempenho do algoritmo. Entretanto, neste caso, a quantidade de cenários e objetos que poderão ser reconhecidos é substancialmente reduzida. Além disto, o algoritmo não é invariante a rotações

do cenário, o que faz com que cenas ou objetos, mesmo que levemente rotacionados ou transformados, não sejam mais detectados.

2.5. Aprendizado de Máquina e Classificadores

A utilização de técnicas de aprendizado de máquina tem apresentado bons resultados para a seleção de características de um objeto e para a consequente construção de classificadores que permitem detectá-lo, como mostram os trabalhos de Freund e Schapire (1995), Rowley et al. (1998) e Schneiderman e Kanade (2000). Papageorgiou e Poggio (2000) foram um dos primeiros a proporem o uso de uma nova forma de representação de imagem, invariável a diferenças moderadas de luminosidade e tamanho, que é fácil de ser computada quando algoritmos de aprendizado de máquina são utilizados. Tal técnica é baseada em uma função matemática conhecida como Haar Wavelets (Mallat, 1989), definida na equação (2-1).

$$\psi(t) = \begin{cases} 1 & 0 \leq t < 1/2, \\ -1 & 1/2 \leq t < 1, \\ 0 & \text{Caso contrário} \end{cases} \tag{2-1}$$

Nesta nova representação, a imagem é inteiramente convertida para o formato de Haar Wavelets, onde são identificadas as principais características retangulares que serão utilizadas durante a fase de treinamento. Tais características são independentes das demais informações da cena, como luminosidade, cor, etc. Para que se possa ter uma idéia do que seria uma imagem transformada para Haar Wavelets, ela é mostrada na figura 11, onde é feita a média das imagens de transformadas em três sentidos (horizontal, vertical e diagonal) e em duas resoluções distintas.

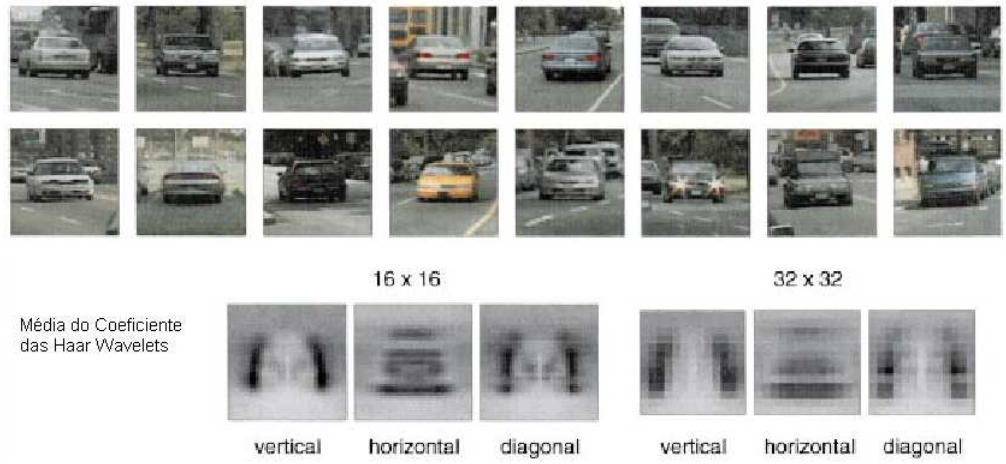


Figura 11. Exemplo de imagens no formato de Haar Wavelets (extraído de Papageorgiou e Poggio (2000))

Para que o treinamento possa ocorrer, são necessárias duas entradas: um conjunto de imagens do objeto de interesse com o mesmo tamanho e posições parecidas do objeto; e outro conjunto de imagens que não possuam o objeto em questão. Tais conjuntos de imagens para treinamento são chamados respectivamente de ***conjunto de imagens positivas*** e ***conjunto de imagens negativas***. Através do processamento destes dois conjuntos, é possível gerar um padrão de classificação capaz de responder se uma determinada região da imagem contém ou não o objeto desejado.

Baseado na técnica de utilização de Haar Wavelets, estudada por Papageorgiou e Poggio(200), Viola e Jones (2001) propõem então um novo conjunto de algoritmos a fim de aprimorar a detecção de objetos, tornando-a ainda mais robusta e extremamente rápida.

O primeiro ponto proposto por Viola e Jones (2001) é a utilização de uma nova representação de imagem, chamada ***Imagem Integral***. Neste tipo de representação, é armazenado, para cada pixel da imagem, o valor do somatório de pixels da área que é formada desde a origem da imagem até o ponto (*x,y*). Com estes valores armazenados, é possível calcular qualquer área retangular da imagem em tempo constante, bastando realizar uma simples conta de subtrações de áreas, conforme demonstrado com mais detalhes no capítulo 3. O fato de serem utilizadas áreas retangulares é proveniente da idéia de Haar Wavelets, pois o gráfico que representa sua função gera também áreas retangulares.

O segundo ponto da proposta de Viola e Jones (2001) é a utilização de algum método de aprendizado de máquina, neste caso o método escolhido é o

AdaBoost (Freund e Schapire, 1995). A terceira contribuição original destes autores é um método para combinar classificadores em uma “cascata” que permite regiões de fundo da imagem serem rapidamente descartadas, enquanto se gasta mais tempo de computação em regiões potencialmente representativas de objetos. Durante o treinamento, o algoritmo de aprendizado é capaz de selecionar as características mais marcantes do objeto a fim de se criar uma cascata de classificadores capaz de reconhecer o objeto numa área qualquer da imagem digital.

Os primeiros níveis da cascata devem possuir um alto índice de rejeição, descartando a maioria das áreas da imagem que não possuem o objeto procurado, enquanto os demais níveis da cascata vão refinando ainda mais as comparações realizadas. Se uma determinada região conseguir passar por todos os níveis da cascata, é porque ela contém o objeto de interesse.

O conjunto de algoritmos propostos por Viola e Jones (2001) mostra-se extremamente eficiente, provendo um ambiente de detecção de objetos extremamente rápido e com baixos índices de falsos positivos. Entretanto, a detecção começa a tornar-se lenta ao ser utilizada em vídeos com resoluções maiores ou quando se procura por vários objetos simultaneamente. Além disto, o tempo que se leva para treinar uma nova cascata de classificadores é demasiadamente longo, podendo levar dias para se obter uma cascata com índice de detecção elevado e baixa taxa de falsos positivos.

Apesar destes pontos negativos, a técnica desenvolvida por Viola e Jones mostra-se tão promissora que vários outros pesquisadores procuram estendê-la no sentido de aprimorar ainda mais os resultados obtidos (Lienhart e Maydt (2002), Messom e Barczak (2006), Wu et al. (2008)). No capítulo seguinte, a técnica de Viola e Jones (2001) é apresentada de forma mais minuciosa.

2.6. Conclusão

Neste capítulo são analisadas as técnicas mais estreitamente relacionadas com o trabalho da presente dissertação. De uma maneira geral, estas técnicas possuem um bom resultado para o problema de detecção de objetos em tempo real. Porém, tais técnicas apresentam apenas soluções para problemas muito bem delimitados, como a detecção de pedestres, placas de trânsito, um conjunto de cenários pré-definidos; ou, então, permitem procurar qualquer tipo de objeto,

mas o seu uso deve ser feito em vídeos de baixa resolução ou com apenas um objeto de cada vez.

O trabalho desta dissertação visa suprir algumas das carências dos trabalhos citados acima, propondo uma técnica inspirada na pesquisa de Viola e Jones (2001). Estes autores apresentam uma das técnicas de detecção de objetos mais abrangentes da literatura, sendo capaz de detectar quaisquer objetos previamente treinados e em baixo tempo de processamento comparado às demais técnicas.

O próximo capítulo apresenta, de forma detalhada, a fundamentação da técnica de Viola e Jones (2001) para o treinamento de novos objetos e para geração/uso da cascata de classificadores, com vistas ao reconhecimento de objetos em imagens e vídeos digitais.