Comparação entre HOG+SVM e Haar-like em cascata para a detecção de campos de futebol em imagens aéreas e orbitais

Juliano E. C. Cruz[1]
Elcio H. Shiguemori[2]
Lamartine N. F. Guimarães[2]

[1] Instituto Nacional de Pesquisas Espaciais – INPE
Caixa Postal 515 – 12245-970 – São José dos Campos - SP, Brasil
juliano.cruz@lac.inpe.br

[2] Instituto de Estudos Avançados – IEAv
Caixa Postal 6044 – 12231-970 – São José dos Campos - SP, Brasil
elcio,guimarae@ieav.cta.br

Abstract. Automatic object recognition in digital images is not an simple task due to diverse variations present within this process, consequently, different general purpose techniques have been proposed. In this paper, an approach combining HOG and SVM and other utilizing Haar-like cascade for automatic soccer field detection in airborne and satellite imagery is analyzed.

Keywords: image processing, pattern recognition, UAV processamento de imagens, reconhecimento de padrões, VANT

1. Introdução

Imagens aéreas ou orbitais com uma alta resolução espacial, permitem que a maioria dos objetos possam ser identificados por especialistas. Alguns objetos são de fácil reconhecimento visual, outros somente com experiência adquirida ao longo do tempo. Em certas aplicações é possível dividir todos os objetos em solo em dois grandes grupos: estáticos e móveis. Objetos em solo, quando estáticos, podem ser utilizados como marcos referenciais(RODRIGUES et al., 2009), como por exemplo, pista de pouso de aeroportos, campo de futebol, entre outros. Já os objetos móveis(REBOUçAS; SHIGUEMORI, 2012) são por exemplo pessoas, carros, embarcações, entre outros, e podem ser utilizados em aplicações de vigilância, fiscalização, resgate ou mesmo militar. Umas das possíveis utilizações da detecção de objetos seria em aplicações que utilizam plataformas VANT(veículo aéreo não tripulado), onde há atualmente uma crescente demanda de seu uso por forças policiais, armadas e em aplicações civis.

Existem hoje em dia muitas abordagens que propõe a detecção de objetos para fins gerais. Dentre as mais utilizadas estão as feições Haar-*like* em cascata(VIOLA; JONES, 2001) e a combinação de descritores HOG(DALAL; TRIGGS, 2005) com o classificador SVM(BOSER; GUYON; VAPNIK, 1992). Nesse trabalho compara-se, então, essas duas abordagens na detecção automática de campos de futebol em imagens aéreas e orbitais. Essas abordagens tem seu uso mais comum em outras áreas, como será visto na Seção 2. Além da possível utilização em sistemas autônomos, uma abordagem de detecção automática possui a vantagem de se eliminar total ou parcialmente o emprego de um operador humano em aplicações de fins gerais. Para um humano a detecção de objetos em imagens aéreas ou orbitais é uma tarefa cansativa e altamente suscetível à erro, pois além de ser uma tarefa entediante, há geralmente uma grande quantidade de informação a ser analisada, e ainda, em certos casos, há a necessidade de que o responsável pela tarefa tenha sido especialmente capacitado para o devido fim. As principais dificuldades encontradas na detecção automática de objetos nas abordagens utilizadas são que as imagens foram obtidas por diferentes tipos de sensores, os objetos podem estar em diferentes poses e

podem também ter sofrido transformações geométricas, entre outros.

2. Trabalhos relacionados

A combinação entre descritores HOG e classificador SVM é conhecido na literatura principalmente na identificação do corpo humano em imagens. A primeira proposta de utilização de HOG com SVM foi em (DALAL; TRIGGS, 2005), em imagens em solo obtidas por diferentes sensores e contendo pessoas diferentes, e conseguiu-se atingir um excelente desempenho. Em (BRECKON et al., 2010), propõe-se um sistema de reconhecimento humano em imagens aéreas de baixa qualidade, onde a combinação dos dois métodos é utilizada na fase de detecção de pessoas enquanto uma outra abordagem é empregada na fase de identificação. Com uma abordagem mais geral mas utilizando a associação HOG e SVM, (FELZENSZWALB et al., 2010) propõe um sistema de detecção em que partes de um determinado objeto são procuradas em uma certa configuração para detectá-lo.

Assim como a combinação HOG e SVM, o classificador Haar-*like* em cascata também possui grande utilização para a detecção de feições humanas. Um dos primeiros trabalhos a fazer uso dessa abordagem foi em (VIOLA; JONES, 2004), onde o detector de faces proposto conseguiu atingir uma excelente performance e robustez. Nessa mesma linha existem outros dois trabalhos: em (WILSON; FERNANDEZ, 2006) primeiramente detecta-se olhos, nariz e boca, e realiza-se, em uma segunda etapa, a análise das detecções para afirmar onde estão as regiões positivas, ou seja, os rostos; já em (PALIY, 2008) utiliza Haar-*like* em cascata para a detecção de rostos e através de uma rede neural realiza a confirmação. Também existem trabalhos que utilizam imagens aéreas obtidas por VANTs para detectar objetos. Em (GASZCZAKA; BRECKON; HANA, 2011) detecta-se carros no espectro visível enquanto em (BRECKON et al., 2009) utiliza-se tanto a faixa do visível quanto a do termal para detectar veículos e pessoas.

3. Métodos

3.1. HOG+SVM

A associação de HOG com SVM, funciona da seguinte maneira: o primeiro é um descritor de alvos que utiliza histogramas de orientação dos vetores gradientes, o segundo é um classificador com alto poder de generalização que utiliza os dados do primeiro método para executar o treinamento e classificação.

O Histograma de Gradientes Orientados (HOG, do inglês *Histogram of Oriented Gradients*) foi primeiramente descrito em 2005 em(DALAL; TRIGGS, 2005). A ideia principal deste descritor é que a aparência e forma de objetos em uma imagem podem ser descritos através da distribuição dos gradientes de intensidade dos pixels ou pelas direções das bordas(GRITTI et al., 2008). O processo para gerar o descritor pode ser dividido em quatro etapas: cálculo do gradiente em cada pixel, agrupamento dos pixels em células, agrupamento das células em blocos e obtenção do descritor.

Primeiramente utiliza-se máscaras unidimensionais (Equação 1) de derivada discreta pontual tanto no eixo vertical como horizontal para o cálculo do gradiente de cada pixel, o resultado é visto na Figura 1(b). O passo seguinte é responsável por agrupar os pixels de uma determinada região, criando-se o que se chama de célula, como pode-se ver na Figura 1(c) e 2(a). Após a segunda etapa, os blocos são criados através do agrupamento de células de uma certa região, como pode-se ver na Figura 2(b). Na etapa final, cria-se o descritor. O descritor nada mais é do que uma lista dos histogramas de todas as células de todos os blocos. A atenuação do problema das variações locais de iluminação ou de contraste entre o primeiro plano e o plano de fundo, se dá através da normalização de cada histograma de acordo com seus próprios valores(DALAL; TRIGGS, 2005). No método canônico, a normalização do

vetor utilizada é o *L2-hys*(LOWE, 2004). *L2-hys* consiste em em aplicar *L2-norm*, descrito na Equação 2, e limitar os resultados em um teto padrão, em seguida calcula-se *L2-norm* novamente.

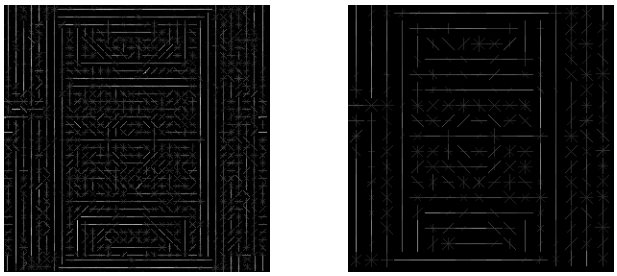
$$[-1, 0, 1] \text{ e } [-1, 0, 1]^T \tag{1}$$

$$\textit{L2-norm: } \frac{v}{\sqrt{\|v\|_2^2 + e^2}} \tag{2}$$

onde v é o vetor descritor, $\|v\|_k$ a sua *k*-norma para $k = 1, 2$ e e uma constante muito pequena(DALAL; TRIGGS, 2005).

(a) (b) (c)

Figura 1: Imagem de entrada (a), magnitude do vetor gradiente dos pixels (b) e das células (c)



(a) (b)

Figura 2: Histograma de orientação das células (a) e dos blocos (b)

Maquina de Vetores de Suporte (SVM, do inglês *Support Vector Machine*) foi descrita em 1992 em (BOSER; GUYON; VAPNIK, 1992) e é um método de aprendizado supervisionado que analisa dados e reconhece padrões usado para classificação e análise de regressão. O SVM é um classificador linear binário, mas existem abordagens que o tornam capaz de classificar um conjunto de dados com classes não-linearmente separáveis ou mesmo com mais de uma classe(THEODORIDIS; KOUTROUMBAS, 2006).

Seja $\mathbf{X}$ um conjunto de treinamento, onde $\mathbf{x}_i$, $i = 1, 2, ..., N$, são vetores de atributos. Estes vetores pertencem a somente a duas classes ω_1 ou ω_2 e assumi-se que são linearmente separáveis. O objetivo é então encontrar um hiperplano(Equação 3) que classifique corretamente os vetores de treinamento.

$$g(\mathbf{x}) = \mathbf{w^T x} + w_0 \tag{3}$$

onde $\mathbf{w}$ é uma matriz unidimensional contendo os vetores de suporte, w_0 é o *bias* e $\mathbf{x}$ o vetor de atributos.

Portanto, tal hiperplano não é único, como pode-se ver na Figura 3(a). Mas há um conceito que deve-se sempre levar em consideração: o poder de generalização do classificador, ou seja, a capacidade do classificador de operar satisfatoriamente com dados de fora do conjunto de treinamento sendo somente projetado com os dados de treinamento. O que o SVM faz a respeito dessa questão, é durante o treinamento escolher o hiperplano que possui maior margem entre as classes, tendo como exemplo a Figura 3(b) (THEODORIDIS; KOUTROUMBAS, 2006).

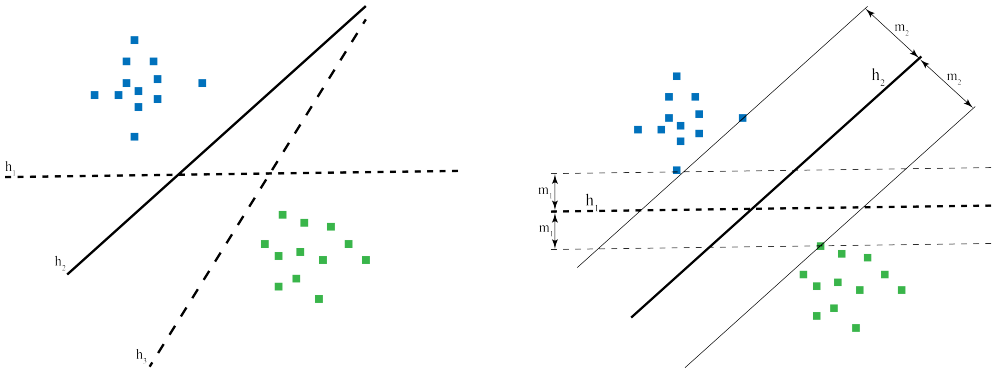


Figura 3: (a)Três classificadores possíveis, (b)Dois hiperplanos e suas margens.

Para se lidar com classes que não são linearmente separáveis, utiliza-se as funções *kernels* que modificam o espaço de atributos transformando o problema em linearmente separável. Além do linear, pode-se encontrar na literatura *kernels* do tipo polinomial, RBF(*Radial Basis Function*) e sigmoidal(HSU; CHANG; LIN, 2010).

3.2. Haar-like em cascata

Feições Haar-*like* são atributos extraídos de imagens e possuem esse nome devido a similaridade com *wavelets* Haar. Em (PAPAGEORGIOU; OREN; POGGIO, 1998) foi proposto a utilização dessas feições ao invés de operar diretamente com os níveis de cinza ou de cor dos pixels(VIOLA; JONES, 2001).

O processo começa somando-se o valor dos pixels nas regiões positivas e negativas do filtro, ou seja, a região branca e preta respectivamente que podem ser vistas na Figura 4. O resultado da subtração da região positiva pela região negativa é utilizado para categorizar as sub-regiões em uma imagem.

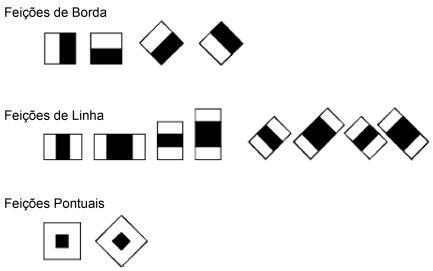


Figura 4: Feições Haar-*like*(LIENHART; KURANOV; PISAREVSKY, 2003)

Uma única feição é considerada um classificador fraco, mas quando colocada em cascata, a combinação se torna um classificador forte (Figura 5) com um alto poder de discriminação, capaz de detectar estruturas independente a iluminação, cor ou escala(VIOLA; JONES, 2004).

Apesar de parecer que o método executa uma busca exaustiva, a arquitetura interna possibilita uma rejeição precoce com o mínimo de avaliação possível, diminuindo, assim, drasticamente o custo computacional. Isto é baseado no fato que a maioria das janelas de detecção são negativas e existem 'poucas' janelas que conseguem passar por todas as etapas. Portanto, o poder computacional é focado nas janelas que possuem a maior probabilidade de ser positivas, uma vez que elas já passaram pelos estágios iniciais da árvore de decisão(CHEN; GEORGANAS; PETRIU, 2007)(VIOLA; JONES, 2001).

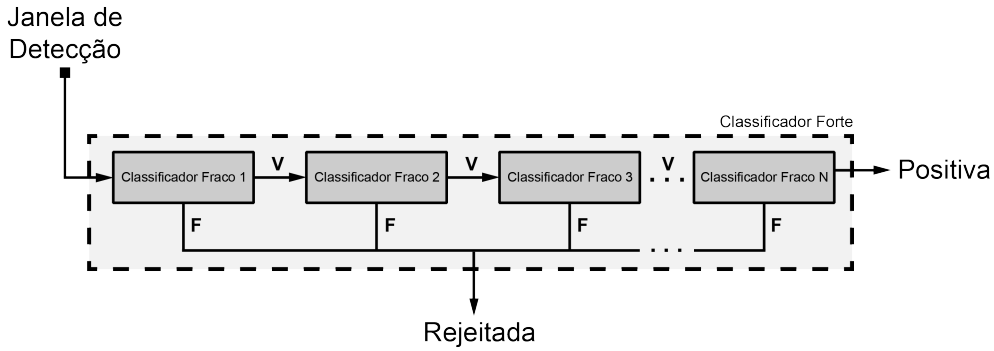


Figura 5: Haar-*like* em cascata

Em qualquer janela dentro da imagem, existem um número enorme de feições Haar-*like*. É necessário portanto, durante a fase de treinamento, focar em um conjunto pequeno de feições cruciais, no intuito de melhorar significativamente o velocidade de classificação sem afetar a precisão. AdaBoost(FREUND; SCHAPIRE, 1995), um algoritmo de aprendizado muito eficaz e com alto poder de generalização, é então responsável por resolver esse problema(VIOLA; JONES, 2001).

4. Metodologia

A escolha em identificar campos de futebol foi feita devido à maior facilidade na montagem de um *dataset* desse alvo, devido a popularidade do esporte ao redor do mundo e de ser de fácil visualização a olho nu em imagens aéreas ou orbitais. A maioria das imagens foram obtidas através do aplicativo Google Maps (Google Maps, 2012), pois ele é de fácil uso, acesso e disponibiliza imagens orbitais em alta resolução de grande parte do mundo, no entanto as imagens possuem marcas d'água com o logotipo do Google e ano de captura. Também foram utilizadas imagens aéreas e Ikonos. A intenção era criar um conjunto heterogêneo de amostras em relação ao sensor utilizado. O conjunto de treinamento utilizado para ambas abordagens continha 1064 imagens positivas e 859 imagens negativas, onde as imagens positivas eram compostas por imagens quadradas que continham campos de futebol rotacionados, como pode-se ver na Figura 6. Na fase de treinamento as amostras positivas e negativas tinham o tamanho da janela de detecção, mas as imagens positivas eram previamente redimensionadas para ter exatamente o tamanho da janela de detecção, enquanto as imagens negativas permaneciam do mesmo tamanho e tinham amostras sucessivamente extraídas.

No classificador Haar-*like* em cascata a janela de detecção utilizada foi de 50 por 50 pixels, a árvore possuía 20 níveis e na fase de treinamento foi utilizado o Gentle AdaBoost e o mínimo de acerto permitido era de 99,9%(LIENHART; KURANOV; PISAREVSKY, 2003).

Na abordagem HOG+SVM utilizou-se o descritor com tamanho de 128 por 128 pixels, com blocos de 16 por 16 pixels, células de 8 por 8 pixels e o histograma de orientação utilizado possuía 9 divisões. A janela de detecção possuía o mesmo tamanho do descritor. O *kernel*

utilizado pelo SVM era do tipo linear e o parâmetro de penalidade *C* utilizado no treinamento era de 0,01 como proposta em (DALAL; TRIGGS, 2005).

Os métodos utilizados foram todos implementados com a biblioteca OpenCV(OpenCV, 2012) exceto na fase de treinamento do SVM que foi utilizado o *software* SVMlight(SVMlight, 2012).

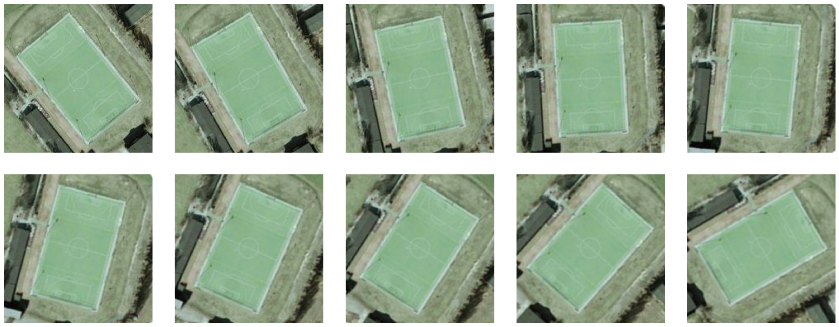


Figura 6: Exemplo de amostras do conjunto de treinamento positivo(Google Maps, 2012)

5. Resultados

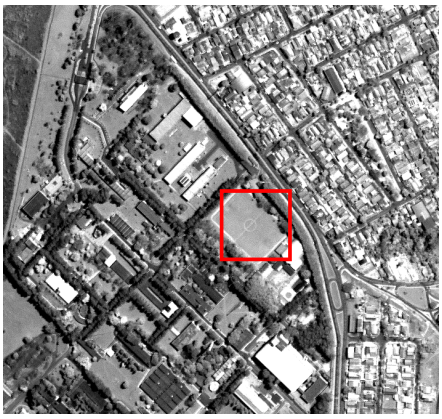


Figura 7: Resultado de detecção utilizando a abordagem HOG+SVM

Utilizou-se duas métricas(FAWCETT, 2006) para medir o desempenho da classificação de um conjunto de imagens obtidas por diversos sensores. As duas métricas são: a precisão, Equação 4, e a sensibilidade, Equação 5.

$$P = \frac{VP}{VP + FP} \times 100 \tag{4}$$

$$S = \frac{VP}{VP + FN} \times 100 \tag{5}$$

onde verdadeiros positivos(VP) são campos de futebol reconhecidos corretamente, falsos positivos(FP) são regiões da imagem em que foram classificadas erroneamente como campo de futebol e falsos negativos(FN) são campos de futebol não reconhecidos.

Foram processadas 13 imagens que continham no total 23 campos de futebol. As imagens utilizadas eram do tipo 'Google Maps', Ikonos e Zeiss(imagem aérea). Os resultados e os

índices de performance podem ser vistos na Tabela 1. Os tempos médios gastos no treinamento e na classificação obtidos em um ambiente Linux com um processador AMD Athlon X2 Dual-Core QL-65 2,1GHz e 4GB de RAM podem ser visto na Tabela 2. O tempo de classificação leva em conta o número de pixels das imagens em relação ao tempo necessário para processá-las.

Método	**VP**	**FP**	**FN**	**Precisão**	**Sensibilidade**
HOG+SVM	17	7	6	70,8%	73,9%
Haar-*like*	14	3	9	82,4%	60,9%

Tabela 1: Resultados de classificação

Método	**Treinamento**	**Classificação**
HOG+SVM	3 minutos	$4,62 \cdot 10^{-6}$ segundo/pixel
Haar-*like*	7 dias	$0,13 \cdot 10^{-6}$ segundo/pixel

Tabela 2: Tempo médio de treinamento e classificação

6. Conclusão

Em ambas as abordagens pode-se notar que as marcas d'água nas imagens do Google Maps provocaram pouca ou quase nenhuma interferência nos procedimentos de treinamento e classificação, isso se deve em grande parte às respectivas arquiteturas internas.

Apesar da abordagem Haar-*like* em cascata ter detectado em números absolutos menos campos de futebol do que HOG+SVM, ela se mostrou ser mais precisa, ou seja, somente 17,6% dos campos detectados eram falsos positivos enquanto na outra abordagem essa porcentagem é de 26,1%. Os falsos positivos na abordagem HOG+SVM obedeciam um certo padrão, objetos de formato retangular ou que ao seu centro possuíam um circulo bem destacado, já no Haar-*like* eles não seguiam uma regra.

Em questão de velocidade de classificação, pode-se observar que o método Haar-*like* em cascata é cerca de 35 vezes mais rápido do que a outra abordagem, porém sua fase de treinamento é absurdamente mais demorada.

Referências

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*. New York, NY, USA: ACM, 1992. (COLT '92), p. 144–152. ISBN 0-89791-497-X.

BRECKON, T. et al. Autonomous Real-time Vehicle Detection from a Medium-Level UAV. *Proc. 24th International Conference on Unmanned Air Vehicle Systems*, 2009.

BRECKON, T. et al. Human identity recognition in aerial images. *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, 2010.

CHEN, Q.; GEORGANAS, N. D.; PETRIU, E. M. Real-time Vision-based Hand Gesture Recognition Using Haar-like Features. In: *Instrumentation and Measurement Technology Conference Proceedings, 2007. IMTC 2007. IEEE*. [S.l.: s.n.], 2007. p. 1–6.

DALAL, N.; TRIGGS, B. Histograms of Oriented Gradients for Human Detection. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Washington, DC, USA: IEEE, 2005. (CVPR '05, v. 1), p. 886–893. ISBN 0-7695-2372-2. ISSN 1063-6919.

FAWCETT, T. An introduction to ROC analysis. In: *ROC Analysis in Pattern Recognition*. New York, NY, USA: Elsevier Science Inc., 2006. v. 27, n. 8, p. 861–874.

FELZENSZWALB, P. F. et al. Object Detection with Discriminatively Trained Part-Based Models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, IEEE, Los Alamitos, CA, USA, v. 32, n. 9, p. 1627–1645, set. 2010. ISSN 0162-8828.

FREUND, Y.; SCHAPIRE, R. A decision-theoretic generalization of on-line learning and an application to boosting. In: VITáNYI, P. (Ed.). *Computational Learning Theory*. [S.l.]: Springer Berlin / Heidelberg, 1995, (Lecture Notes in Computer Science, v. 904). p. 23–37.

GASZCZAKA, A.; BRECKON, T.; HANA, J. Real-time People and Vehicle Detection from UAV Imagery. *Proc. SPIE Conference Intelligent Robots and Computer Vision XXVIII: Algorithms and Techniques*, 2011.

Google Maps. *Google*. 2012. `http://maps.google.com`.

GRITTI, T. et al. Local Features based Facial Expression Recognition with Face Registration Errors. *IEEE International Conference on Automatic Face and Gesture Recognition*, 2008.

HSU, C. W.; CHANG, C. C.; LIN, C. J. *A Practical Guide to Support Vector Classification*. 2010.

LIENHART, R.; KURANOV, A.; PISAREVSKY, V. Empirical Analysis of Detection Cascades of Boosted Classifiers for Rapid Object Detection. *Pattern Recognition*, p. 297–304, 2003.

LOWE, D. G. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, Springer Netherlands, Hingham, MA, USA, v. 60, n. 2, p. 91–110, nov. 2004. ISSN 0920-5691.

OpenCV. 2012. `http://opencv.willowgarage.com/wiki/`.

PALIY, I. Face Detection Using Haar-like Features Cascade and Convolutional Neural Network. *International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science*, Lviv-Slavsko, Ukraine, 2008.

PAPAGEORGIOU, C.; OREN, M.; POGGIO, T. A General Framework for Object Detection. *International Conference on Computer Vision*, 1998.

REBOUçAS, R.; SHIGUEMORI, H. Acompanhamento de objetos móveis em imagens aéreas. *I Simpósio de Ciência e Tecnologia do IEAv*, 2012.

RODRIGUES, R. et al. Color and Texture Features for Landmarks Recognition on UAV Navigation. *Anais do XIV Simpósio Brasileiro de Sensoriamento Remoto*, 2009.

SVMlight. 2012. `http://svmlight.joachims.org/`.

THEODORIDIS, S.; KOUTROUMBAS, K. *Pattern Recognition, Third Edition*. Orlando, FL, USA: Academic Press, Inc., 2006. ISBN 0123695317.

VIOLA, P.; JONES, M. Rapid object detection using a boosted cascade of simple features. In: *2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Los Alamitos, CA, USA: IEEE Comput. Soc, 2001. v. 1, p. 511–518. ISBN 0-7695-1272-0. ISSN 1063-6919.

VIOLA, P.; JONES, M. Robust Real-Time Face Detection. *International Journal Computer Vision*, 2004.

WILSON, P. I.; FERNANDEZ, J. Facial feature detection using Haar classifiers. *J. Comput. Small Coll.*, Consortium for Computing Sciences in Colleges, , USA, v. 21, n. 4, p. 127–133, 2006. ISSN 1937-4771.