UNIVERSIDADE FEDERAL DO PIAUÍ – UFPI
CAMPUS SENADOR HELVIDIO NUNES DE BARROS-CSHNB
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

CLASSIFICAÇÃO DE NÓDULOS PULMONARES UTILIZANDO VETORES DE QUANTIFICAÇÃO

Ivo Alves de Oliveira

PICOS – PIAUÍ
2016

IVO ALVES DE OLIVEIRA

CLASSIFICAÇÃO DE NÓDULOS PULMONARES UTILIZANDO VETORES DE QUANTIFICAÇÃO

Monografia submetida ao Curso de Bacharelado de Sistemas de Informação como requisito parcial para obtenção de grau de Bacharel em Sistemas de Informação.

Orientadora: Prof. Ma. Alcilene Dalília de Sousa

PICOS – PIAUÍ
2016

FICHA CATALOGRÁFICA
Serviço de Processamento Técnico da Universidade Federal do Piauí
Biblioteca José Albano de Macêdo

O482c Oliveira, Ivo Alves de.

Classificação de nódulos pulmonares utilizando vetores de quantificacão / Ivo Alves de Oliveira.– 2016.
CD-ROM : il.; 4 ¾ pol. (37 f.)
Monografia (Curso Bacharelado em Sistemas de Informação) – Universidade Federal do Piauí, Picos, 2016.
Orientador(A): Profª. Ma. Alcilene Dalíllía de Sousa

1. Câncer de Pulmão-Extração e Classificação. 2. Imagens Médicas. 3. Sistemas Auxiliares de Computador. I. Título.

CDD 005

CLASSIFICAÇÃO DE NÓDULOS PULMONARES UTILIZANDO VETORES DE QUANTIFICAÇÃO

IVO ALVES DE OLIVEIRA

Monografia aprovada como exigência parcial para obtenção do grau de Bacharel em Sistemas de Informação.

Data de Aprovação

Picos – PI, 18 de fevereiro de 2016

Alcilene Dalília de Sousa
Profª. Ma. Alcilene Dalília de Sousa
Orientador

Patrícia Medyna L L Dru.
Profª. Ma. Patrícia Medyna Lauritzen de Lucena Drumond
Membro

Prof. Me. Antonio Oseas de Carvalho Filho
Membro

Aos meus pais Francisco Alves Ferreira e Maria Salomé Pereira de Oliveira e a meus irmãos pela eterna confiança, força e eterno incentivo.

*A*GRADECIMENTOS

Agradeço primeiramente a DEUS pela oportunidade e pelo privilégio que nos foram dados em compartilhar tamanha experiência e, ao frequentar este curso, perceber e atentar para a relevância de temas que não faziam parte, em profundidade, das nossas vidas.

A minha Orientadora Prof.ª Ma. Alcilene Dalília de Sousa pelo incentivo, simpatia e presteza no auxílio às atividades e discussões sobre o andamento e normatização desta Monografia de Conclusão de Curso.

Particularmente a Prof.ª Ma. Patrícia Medyna Lauritzen de Lucena Drumond, por sua vocação inequívoca, por não poupar esforços como interlocutor dos alunos e por suprir eventuais falhas e lacunas.

Aos colegas de classe pela espontaneidade e alegria na troca de informações e materiais numa rara demonstração de amizade e solidariedade.

Aos Professores Me. Romuere Rodrigues Veloso e Silva e Flávio Henrique Duarte de Araújo pela disponibilidade na resolução de dúvidas e por sua dedicação ao auxiliarmos em assuntos pertinentes ao trabalho apresentado.

Aos demais idealizadores, coordenadores e funcionários da Universidade Federal do Piauí-Campus Senador Helvídio Nunes de Barros.

A todos os professores pela dedicação, entusiasmo, carinho e disponibilidade que demonstraram ao longo do curso.

A minha família e amigos pela paciência em tolerar a minha ausência.

E, finalmente, a todas as pessoas com as quais convivi em todo o período que estive ligado à Universidade Federal do Piauí. Pois ajudaram na minha formação tanto profissional quanto como pessoa e me fizeram refletir sobre muitos conceitos. Aos que me ajudaram em momentos difíceis e partilharam, também, de momentos felizes.

“Inovação se distingue entre um líder e um seguidor”
Steve Jobs.

“Não ache que as coisas da vida são fáceis, tudo na vida tem que ser lutado; e quando conquistares uma coisa fácil desconfie, pois ela não é tão fácil como parece”
Silvio Santos.

RESUMO

O câncer de pulmão é o tipo de câncer que tem a maior taxa de mortalidade no mundo e a menor taxa de sobrevida depois de diagnosticado. Uma das formas de detecção e diagnostico dos candidatos a nódulos pulmonares é através dos Sistemas Auxiliados por Computador (CAD), o qual utiliza como uma de suas etapas a extração e seleção de características dos possíveis candidatos a nódulos pulmonares. Nesta etapa, são retiradas características que ajudam a identificar os candidatos a nódulos utilizando técnicas, como o algoritmo de vetores de quantificação. O presente trabalho descreve o estudo e o aperfeiçoamento de técnicas de extração de características, obtendo melhores resultados na sensibilidade de 98.3% e a diminuição na incidência dos falsos positivos. Além de obter uma classificação com uma sensibilidade de 97% utilizando o classificador SVM.

Palavras-chave: Câncer de Pulmão, Imagens Medicas, Extração e Classificação.

ABSTRACT

Lung cancer is the cancer that has the highest mortality rate in the world and the lowest rate of survival after diagnosed. One of the ways of detection and diagnosis of lung nodules is through Computer Aided Systems (CAD), which it uses as one of its extraction and selection steps of characteristics of possible candidates for lung nodules. In this step, are dropped features that help identify candidates for nodules using techniques such as vector quantification algorithm. This paper describes the study and improvement of techniques of extraction of features, obtaining better results in sensitivity of 98.3% and the decrease in the incidence of false positives. In addition to getting a rating with a sensitivity of 97% using SVM classifier.

Keywords: Lung Cancer, Medical Images, Extraction and Classification.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CAD	Sistema Auxiliado Por Computador.
DICOM	*Digital Imaging Communications in Medicine.*
DRS	*Diverse Random Subspace.*
FP	Falso Positivo.
HRS	*Hybrid Probabilistic Sampling.*
INCA	Instituto Nacional do Câncer.
IRDI	*Image Database Resource Initiative*.
LCA	*Lung Cancer Alliance*.
LIDC	*Lung Image Database Consortium*.
NCI	*National Cancer Institute of EUA*.
SVM	Maquina de Vetor de Suporte.
TC	Tomografia Computadorizada.
VP	Verdadeiro Positivo.
WEKA	*Waikato Environment for Knowledge Analysis.*

SUMÁRIO

1 INTRODUÇÃO

Câncer é o nome dado a mais de 100 doenças que tem em comum o crescimento desordenado das células, que podem invadir órgãos e tecidos, e que em alguns casos mais graves pode-se espalhar por diversas regiões do corpo. Ele é considerado um dos cânceres mais letais, o câncer de pulmão é uma das doenças que mais mata em todo o mundo, tendo um aumento em sua incidência de 2% a cada ano (INCA, 2016).

Entre as principais causas estão fatores hereditários e o uso do tabaco (cigarro), sendo que o tabaco é indicado como o principal fator, abrangendo cerca de 90% dos casos. Sua taxa de sobrevida é baixa, sendo ela de 7 a 10% em países subdesenvolvidos e de 13 a 21% em países desenvolvidos. E tem-se a estimativa de 28.220 novos casos em 2016, sendo que destes 17.330 em homens e 10.890 em mulheres (INCA, 2016).

O profissional da área médica responsável pela análise das imagens e detecção de possíveis nódulos pulmonares, necessita de muito esforço e tempo para fazer a identificação de características que o ajudem em um possível diagnóstico. Contudo, muitos especialistas utilizam Sistemas Auxiliados por Computador (CAD) para diminuir o tempo total de processamento e ter uma segunda opinião na hora de proferir o diagnóstico sobre a existência ou não de um nódulo pulmonar.

As imagens de Tomografia Computadorizada(TC) proporcionam uma melhor visualização da área analisada e são amplamente usadas para detecção de estruturas não pertencentes aquela região. A extração de características se beneficia das imagens de TC, pois permitem a visualização das estruturas do pulmão com um maior realce e ajuda na obtenção de medidas, como a esfericidade, o que influencia na hora da classificação dos candidatos a nódulos, pois analisa estruturas que tem características que indicam a presença de possíveis nódulos.

Um sistema CAD é geralmente composto de três etapas, segmentação, extração e seleção de características, e classificação. Na etapa de segmentação é selecionada a região de interesse na imagem dando um realce nas estruturas, na extração de características é feita a análise e retirada de pontos e/ou características

importantes da imagem avaliada e na classificação dá-se um indicação sobre estruturas que podem ou não ser nódulo pulmonar.

Todas as etapas são essenciais na detecção precisa dessas estruturas. A etapa de extração e seleção de características é a etapa em que alguns aspectos e medidas matemáticas, como o diâmetro da região, são extraídas da imagem com o intuito de encontrar parâmetros que possam ser comparados. Além de um número de características suficientes para se ter uma menor taxa de erro na descoberta de possíveis candidatos a nódulos.

A extração e seleção de características é o processo no qual faz-se a retirada de características de uma imagem, com o objetivo de encontrar evidências que indique a existência de possíveis padrões que indiquem candidatos a nódulos pulmonares, com uma margem de erro pequena na condição de minimizar a incidência de falsos positivos (FP), que podem ser identificados como possíveis nódulos. Esse processo busca encontrar padrões de forma, textura ou cores que evidenciem potenciais candidatos.

Após o processo de extração de características é feita a classificação das estruturas que forem indicadas como possíveis candidatos a nódulos. Para tanto pode-se utilizar a ferramenta Weka (KRISHNAIAH *et. al.*, 2013), que contém uma biblioteca com implementações de classificadores de características e tem suporte para a adição de novos classificadores.

O trabalho visa a implementação e otimização das técnicas de extração de características, a fim de diminuir os custos do processo. E realizar uma extração que consiga retirar o máximo de características necessárias para uma classificação mais correta e com uma menor taxa de falsos positivos.

1.1 Objetivo

O presente trabalho teve como objetivo o estudo e a implementação de técnicas que fazem a extração e seleção de características de candidatos a nódulos pulmonares, a fim de determinar um conjunto de características nessas estruturas de forma automática para que a classificação torne-se mais precisa, minimizando a incidência de falsos positivos.

1.2 Organização do Trabalho

O trabalho foi organizado, a fim de dar um entendimento ao leitor sobre o problema e os passos seguidos para se chegar a metodologia proposta. E está organizado em 5 (cinco) capítulos.

No Capítulo 2 é feito todo o embasamento teórico sobre o tema estudado. Bem como são tratados temas pertinentes ao estudo, além de conter uma revisão bibliográfica de todo o material de pesquisa.

No Capítulo 3 é descrita a metodologia utilizada no trabalho, assim como as métricas e modelos testados, a fim de obtermos a forma mais adequada de tratamento sobre o tema proposto.

No Capítulo 4 são expostos todos os resultados e qual seu impacto na literatura, visto que foram necessários alguns testes antes de se chegar ao resultados apresentados.

No Capítulo 5 é feita a discursão e conclusão sobre os resultados obtidos, além da projeção de trabalhos futuros que podem ser gerados a partir do estudo apresentado.

2 ESTADO DA ARTE

Para o desenvolvimento do trabalho foi feito o estudo das técnicas de extração e seleção de características, que será descrita neste capítulo. Este capítulo visa apresentar todo o embasamento teórico necessário para compreensão do tema. E tópicos que servirão para demonstrar os resultados alcançados.

2.1 Câncer de Pulmão

O Pulmão é um órgão esponjoso e elástico formado por milhões de alvéolos que se enche de ar (LAURENCE, 2005). Tem um volume de aproximadamente 25 cm de comprimento, sendo o pulmão direito maior em largura e mais curto em altura do que o esquerdo, e 700g de peso. Sua principal função é fornecer ao nosso sangue oxigênio, que é transportado para as células do corpo (NETTER, 2000). Uma representação do pulmão pode ser vista na Figura 1.

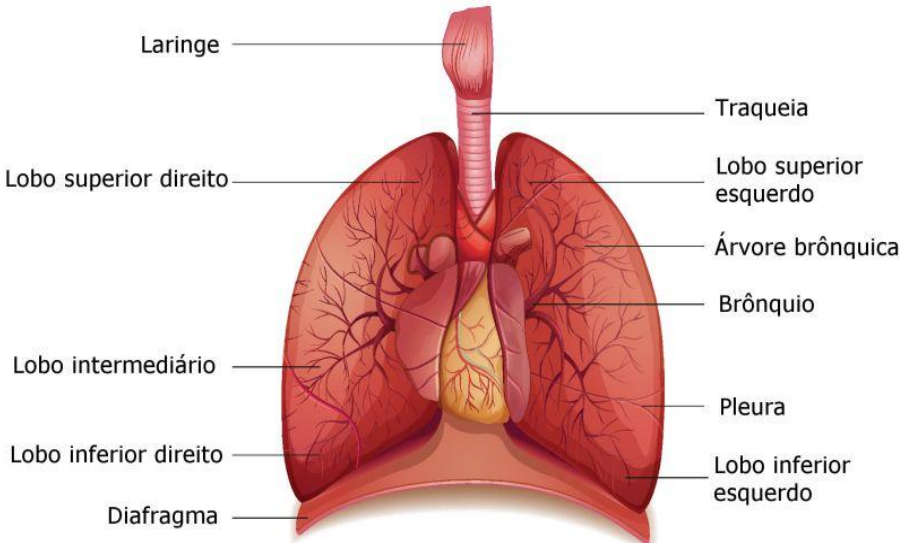


Figura 1. Pulmão Humano (NETTER, 2000).

A Figura 1, ilustra a organização do pulmão humano, bem como todas as suas estruturas. Além do câncer, o pulmão pode contrair outras doenças, tais como pneumonia (inflamação dos pulmões causada por infecção por bactéria), enfisema,

pleurite, tuberculose, e bronquite. E também pode ocorrer o acumulo de massa no pulmão, e algumas dessas doenças influenciam na hora da detecção dos nódulos pulmonares.

O câncer de pulmão consiste no crescimento desordenados das células pulmonares, e este crescimento pode ocorrer na região justa pleural (colado a parede do pulmão, chamada de parênquima). Ou na região interna que compreende toda a parte interna do pulmão. Uma estrutura pulmonar para ser reconhecida como nódulo, primeiro passo na análise sobre a existência ou não do câncer, tem que ter diâmetro máximo de 3 centímetros e estar rodeado de tecido pulmonar normal.

2.2 Processamento de Imagens

Processamento de imagens digitais é a área da tecnologia que visa a obtenção de conhecimento computacional a partir do processamento de uma imagem digital. Onde uma imagem 2D digital nada mais é do que uma função bidimensional, f(x, y), em que x e y são coordenadas. E que a amplitude de f em algum dos pares de coordenadas (x, y) é chamada de nível de intensidade de cinza para aquele ponto. Sendo que uma imagem pode ser representada em tons de cinza, binaria (preto e branco) e em formato *Red, Green and Blue*(RGB), em que cada fatia da imagem RGB tem um valor que vai de 0 a 255. A representação de uma imagem digital é mostrada na Figura 2.

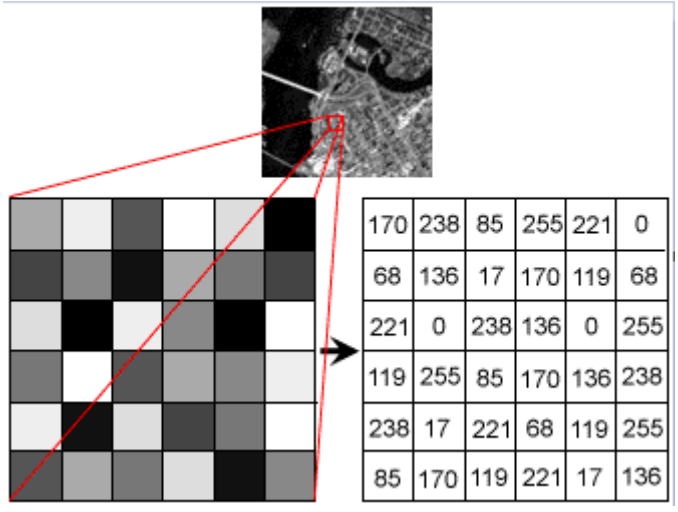


Figura 2. Imagem Digital (GONZALEZ, 2002).

Na Figura 2, pode-se ver a representação de uma imagem digital a partir de uma fatia da imagem original. A primeira aplicação do processamento de imagens aconteceu na década de 20 e foi utilizada para que fossem melhoradas imagens transmitidas por cabos submarinos entre Londres e Nova York para jornais.

Na época o interesse maior na área dava-se no sentido do melhoramento de informações para visualização humana e o processamento de dados para percepção automática através de maquinas. Hoje o processamento de imagens é utilizado para inúmeras funcionalidades, tais como melhoramento de imagens médicas, melhoramento de imagens transmitidas entre dois pontos e processamento de imagens em empresas fotográficas.

Porém, o processamento de imagens muitas vezes é confundido com a visão computacional e a computação gráfica. Contudo, as três vertentes servem a propósitos diferentes. No qual o processamento de imagens serve para melhoramento e pré-processamento de uma imagem e tem como entrada e saída uma imagem.

Já a visão computacional é utilizada para tirar conhecimento de uma imagens a partir de sua análise e tem como entrada uma imagem e a saída é modelo matemático, ou seja, informações. E a computação gráfica tem como proposito a construção de uma imagem e que tem como entrada dados acerca da imagem a ser desenvolvida. Nas Figuras 3,4 e 5 é mostrado o fluxograma dos 3 sistemas.

Figura 3. Sistema de computação gráfica.

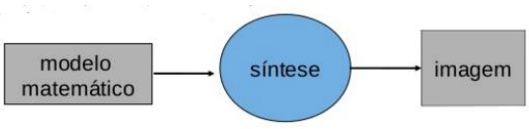


Figura 4. Sistema de visão computacional.

Figura 5. Sistema de Processamento digital de imagens.

Os sistemas de processamento de imagens digitais seguem algumas etapas, sendo elas; a aquisição de imagens, no qual a imagem é obtida através de algum método de obtenção de imagens, como a Tomografia Computadorizada. O pré-processamento em que é feita uma filtragem e melhoramento da imagem, a segmentação, representação e descrição, e reconhecimento e interpretação. Sendo que todas essas etapas levam em consideração o domínio do problema e giram em torno de uma base de conhecimento.

Na Figura 6, vê-se todos as etapas de um sistema de processamento digital de imagens. Antes de iniciarmos qualquer ação neste sistema, temos que ter o entendimento do problema no qual será aplicado o sistema, ou seja, ter uma boa base sobre o domínio do problema.

Após o entendimento completo do problema, as imagens a serem analisadas serão submetidas ao sistema que é composto pelas fases de aquisição das imagens que compõem o problema. As fases seguintes são o pré-processamento, a segmentação, a representação e descrição e o reconhecimento e interpretação do problema para que sejam gerados os resultados do sistema. Vale ressaltar que todas essas fases são importantes para a geração de uma base de conhecimento sobre o assunto, que pode ser reutilizada em trabalhos que tratem do mesmo domínio do problema.

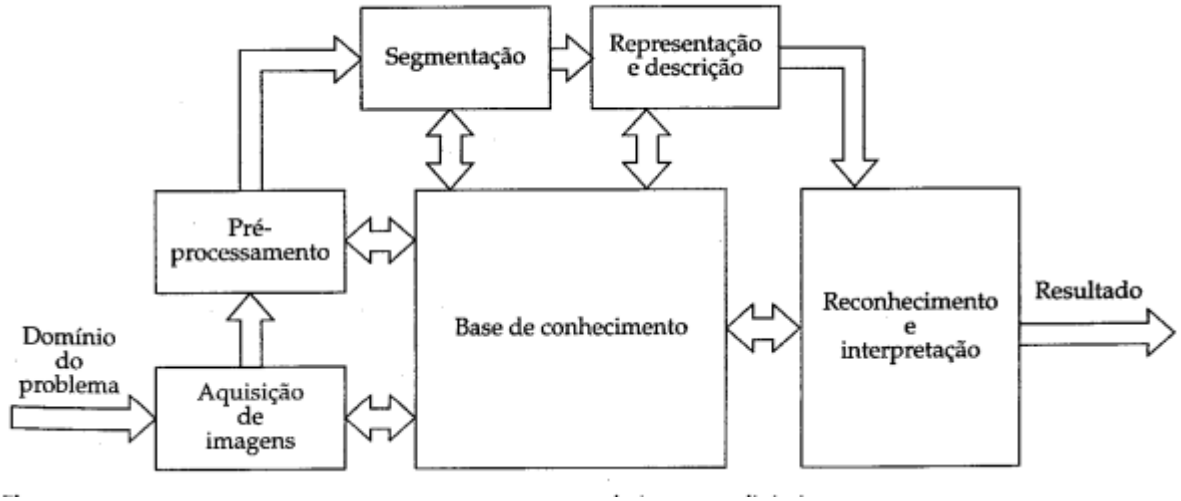
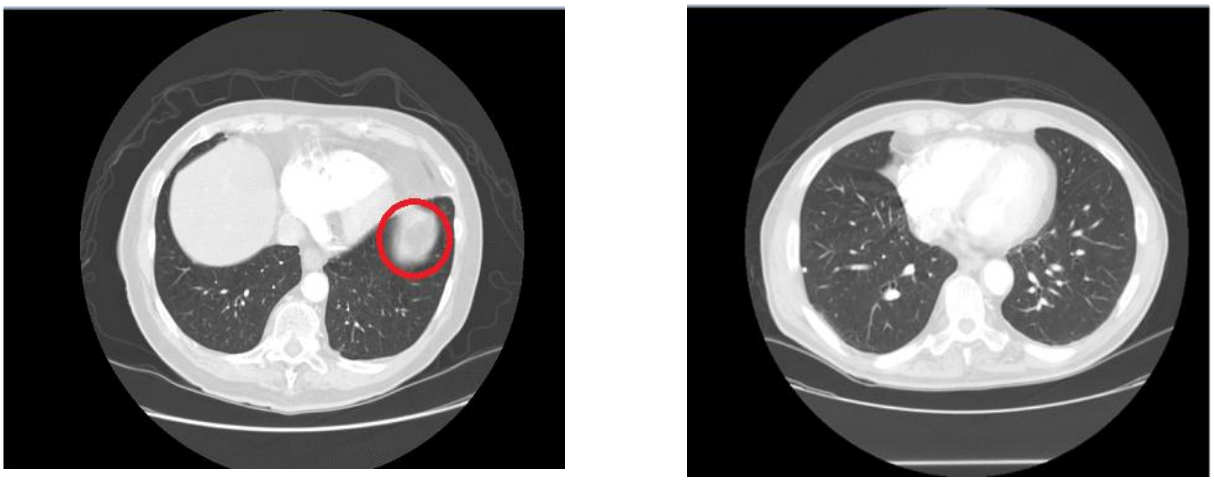


Figura 6. Processamento digital de imagens(GONZALEZ,2002).

2.3 Base de Imagens

As imagens utilizadas neste trabalho foram conseguidas através da base de imagens LIDC-IRDI (ARMATO, 2011). Essa base tem imagens que seguem um padrão e tem maior qualidade por serem imagens de Tomografia Computadorizada do Tórax, sendo citada em grande parte dos trabalho que tratam do assunto. A Figura 7 a apresenta duas imagens da base LIDC-IRDI, uma de um pulmão com nódulo e outra de um saudável.



(a) (b)

Figura 7. Imagens da base LIDC-IRDI (ARMANTO, 2011) (a) Pulmão com nódulo; (b) Pulmão Saudável.

A base LIDC-IDRI é fornecida pelo *National Cancer Institute of EUA* (NCI) e as imagens estão no formato DICOM (*Digital Imaging Communications in Medicine*), possuem 16 bits por *voxel* (representação tridimensional de um *pixel*) e dimensões de 512 X 512. A base contém ainda um arquivo com informações sobre os nódulos e as marcações feitas por quatro especialistas.

Percebe-se, pela análise da Figura 7 que o pulmão contém várias estruturas e que a análise manual de cada uma delas pelo especialista é um trabalho muito dispendioso e que consome muito tempo.

Pode-se notar, também, que na Figura 7a além das estruturas do pulmão existem duas áreas mais claras na imagem. A região delimitada pelo círculo vermelho trata-se de um nódulo, que foi identificado e catalogado pelos quatro especialistas que analisaram a imagem. Já a outra região mais clara, pode ter sido proveniente de outras enfermidades que atingem o pulmão, ou até mesmo pode ser uma acumulo de massa.

A Figura 7b, segundo a marcação dos especialista, trata-se de um pulmão saldável. Não havendo nenhuma estrutura nodular, ou seja, somente estruturas características do pulmão.

2.4 Transformada Karhumen-Loève

A transformada K-L é uma transformada ortogonal, por criar uma representação matemática e por transformar um modelo de representação complexo em um modelo representativo simples, ou seja, simplificando modelos. E que trabalha com funções não-senoidal que visa a compressão de um grupo amostral e a extração de características matemáticas de uma base ou de diversas bases de conhecimento. Ela é usada quando se pretende criar um modelo de dados a ser seguido em um processo de análise ou na criação de uma coleção de características de um determinado grupo analisado.

Ela foi usada visando a compressão das imagens sem a perda de qualidade, visto que as imagens utilizadas tem uma alta resolução, o que torna o processamento das mesma mais demorado. E foi usada também para obtenção de algumas

características pertinentes ao trabalho, como o diâmetro mínimo de um candidato a nódulo pulmonar e outras medidas e parâmetros matemáticos.

2.5 Vetores de Quantificação

Os vetores de quantificação são construídos com base em dois tipos de características, a saber, as de alto nível e as de baixo nível. Sendo que, as características de alto nível estão ligadas a reconstrução ou reorganização da imagem, a fim de fornecer um maior entendimento sobre as estruturas fornecidas pela imagem da região analisada (BAGCI *et al*., 2012).

Como as características de alto nível fazem a reestruturação da imagem, são comumente utilizadas na fase de segmentação da imagem. Já as características de baixo nível compreendem aspectos mais intrínsecos as estruturas obtidas nas imagens. Alguns destes aspectos são a textura, forma, entre outros. Essas características são tidas como contextuais e são essenciais na classificação das estruturas presentes na imagem (HAO et al., 2013).

A criação dos vetores dar-se pela divisão das características em seus respectivos grupos, ou seja, de alto e baixo nível. Essa divisão consome muito tempo e processamento. Tendo em vista, que pode ser necessário dividir a imagem em blocos para tornar a construção dos vetores mais eficiente. A produção dos vetores pode ser feita utilizando uma imagem segmentada visando uma maior efetividade na caracterização dos vetores. Ou utilizar uma imagem não segmentada, visando a utilização do vetor de alto nível na segmentação da imagem.

2.6 Descritores de Forma

Os descritores de forma são utilizados para análise de imagens diversas, como as imagens medicas. E são divididas em dois grupos descritores globais e estruturais, a primeira classe comtempla a forma como um todo, já o segundo grupo se baseia em figuras geométricas para a identificação da forma.

Neste trabalho foram utilizados descritores de forma globais, pois apesar dos nódulos terem formato esférico, eles também tem regiões especuladas em seu perímetro e o uso de descritores estruturais poderia deixar candidatos em potencial fora do grupo de amostragem.

2.7 Suport Vector Machine

As *Support Vector Machine* (SVM) são sistemas de aprendizado que utilizam um espaço de hipótese de funções lineares em um espaço de muitas dimensões (CRISTIANINI, 2000). Foi idealizada por Vapnik e consiste em um poderoso método de análise e classificação, que vem sendo utilizado em vários ramos de conhecimento.

A SVM é uma técnica de classificação que utiliza aprendizado supervisionado, ou seja, ela utiliza classes que indiquem a classificação de determinada estrutura analisado diferentemente do aprendizado não supervisionado que o algoritmo tem que aprender sem ter uma base de conhecimento preestabelecida.

A SVM usa para classificação de estruturas o conceito de hiperplano, que nada mais é do que uma região de separação em um espaço multidimensional, em que o número de dimensões usadas podem ser infinitas. Desta forma a SVM é capaz de dividir em classes um determinado grupo amostral de forma que minimize o erro ao máximo. Podemos ver como dar-se a divisão das classes no hiperplano, na Figura 8.

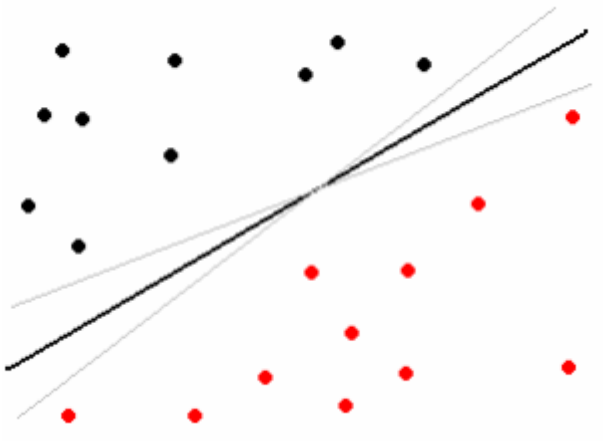


Figura 8. Separação de classes em um hiperplano na SVM (SOUSA, 2011).

2.8 Trabalhos Relacionados

No processo de extração e seleção de características são analisadas características de forma ou de textura da imagem. Sendo que todas as informações obtidas na extração de características de uma imagem específica precisam ser armazenadas em arquivo de texto para posterior classificação, baseada nas características obtidas. A seguir, são mostrados alguns trabalhos que serviram de base para a construção da fundamentação teórica e para desenvolvimento da metodologia apresentada neste trabalho.

Cao *et al.* (2014) descreve a utilização dos método HRS *(Hybrid Probabilistic Sampling*) e DRS (*Diverse Random Subspace*) que usam dados probabilísticos para fazer a extração de características com base na faixa de forma e intensidade que mais evidenciam possíveis candidatos a nódulos pulmonares. Em seguida, ele faz a junção dos dois algoritmos com o intuito de corrigir falhas que ocorrem em um algoritmo e que no outro são executadas de forma mais plena. Ainda é feita uma comparação entre o método proposto (HRS-DRS) e os algoritmos separadamente. Ao final o autor comenta sobre a melhora ocorrida com o uso dos dois algoritmos trabalhando em conjunto. Porém, a abordagem proposta consome muito tempo na geração dos dados probabilísticos e necessita de máquinas com grande poder de processamento.

Hao *et al*. (2013) detalha o método de extração de características através de vetores de quantificação. Ele mostra como se dá a criação dos vetores e para que os vetores de alto e baixo nível são usados, respectivamente segmentação e extração de características. E aliados a esse método de extração é utilizado uma filtragem baseada em regras, que serve para diminuir a incidência de falsos positivos (FP). Contudo, caso as regras de filtragem e criação dos vetores não estejam bem definidas, o método pode remover estruturas que sejam candidatos a nódulos.

Araujo (2011) descreve a importância dos descritores na recuperação de informações em uma imagem. Ele inicia o trabalho falando sobre sistemas de recuperação baseados em conteúdo e como a extração de características é importante neste tipo de sistema. O autor frisa que nestes sistemas são criados vetores de características, com base em elementos primitivos de cor, textura e forma. E que os vetores são utilizados na recuperação e comparação das imagens em um sistema que se utiliza do processamento de imagens.

Carvalho filho (2013) descreve a construção de uma metodologia para detecção automática de nódulos pulmonares solitários. Inicialmente, o autor, faz uma pequena introdução ao tema. Em seguida, descreve como o uso do técnica *Quality Threshold Clustering*(QT) é utilizada no processo de segmentação das imagens de TC. Na fase de extração e seleção de características dos candidatos a nódulos pulmonares, o autor descreve com clareza com as características de forma e textura serão retiradas, bem como os métodos usados para a extração, tal como a desproporção esférica e o histograma. E por fim, ele descreve como a classificação das características é feita, através do uso da máquina de vetor de suporte.

Rey *et al.* (2010) descreve como é feita a extração de características dos candidatos a nódulos através dos algoritmos baseados em redes neurais, como pode-se organizar uma rede para que possamos ter uma maior acurácia na extração de características. Neste método pode-se criar várias camadas intermediárias entre a camada de entrada, na qual são colhidas as imagens de entrada, e a camada de saída que reunirá todas as características que as camadas intermediárias coletaram. Porém, nesta abordagem há uma problemática, pois não temos como mensurar quantas camadas intermediárias serão necessárias para coletar todas as características pretendidas.

Sousa (2011) descreve uma metodologia automática para detecção de

nódulos pulmonares, a fim de que o sistema faça a detecção de nódulos pulmonares sem a necessidade de que haja uma intensa intervenção do especialista. O trabalho começa com uma pequena introdução ao assunto, dá uma explanada em conceitos pertinentes ao trabalho. O autor fala, também, sobre a técnica de esqueletonização volumétricos de objetos baseada em afinamento conexo. E descreve como é feita a extração de cada estrutura isoladamente, a separação dos vasos e nódulos e como é feita a classificação dos candidatos a nódulos, além de como é feita a avaliação de desempenho da técnica proposta.

Marar (2013) descreve como a transformada Karhumen-Loève é utilizada no reconhecimento de padrões numéricos. Inicialmente ele faz uma pequena introdução sobre as transformadas e seus tipos. Em seguida, o autor descreve o processo de extração de características via transformadas ortogonais e via a transformada K-L. Por fim é descrito o processo de utilização de transformadas na construção de classificadores.

Contudo, as técnicas explicitadas na literatura deixam a desejar no processo de extração de características, provocando uma incerteza na interpretação das características e, às vezes, colhendo informações desnecessárias da imagem ou de pouca relevância. Porém, a maior dificuldade encontrada nos algoritmos de extração que compõe os sistemas CAD é não ter um padrão de forma e textura definidas, que ajudem o profissional da área médica ou os próprios sistemas CAD’s a montarem um perfil dos possíveis candidatos a nódulos.

Com presente trabalho, espera-se contribuir com o processo de extração e seleção de características dos candidatos a nódulos pulmonares, afim de diminuir a incidência de falsos positivos e otimizar o processo. Através da otimização da técnica de vetores de quantificação e da filtragem baseada em regras, espera-se também contribuir com os profissionais da área medica e diminuir a incerteza no indicação de nódulos pulmonares.

3 MÉTODO PROPOSTO

O método proposto consiste na criação de vetores de quantificação, os quais guardam características de alto e baixo nível. Porém neste trabalho foi usado apenas o vetor de baixo nível, pois é o vetor utilizado no processo de extração de características (HAO *et al.*, 2013). Foi adicionado, também, a este vetor aspectos importantes de forma e textura que puderam ser observados em outras técnica de extração, tais como o contorno das estruturas nodulares e a intensidade do pixel mais intenso nestas estruturas. O processo de um sistema CAD pode ser visto na Figura 9, com mais ênfase ao processo de extração e seleção de características.

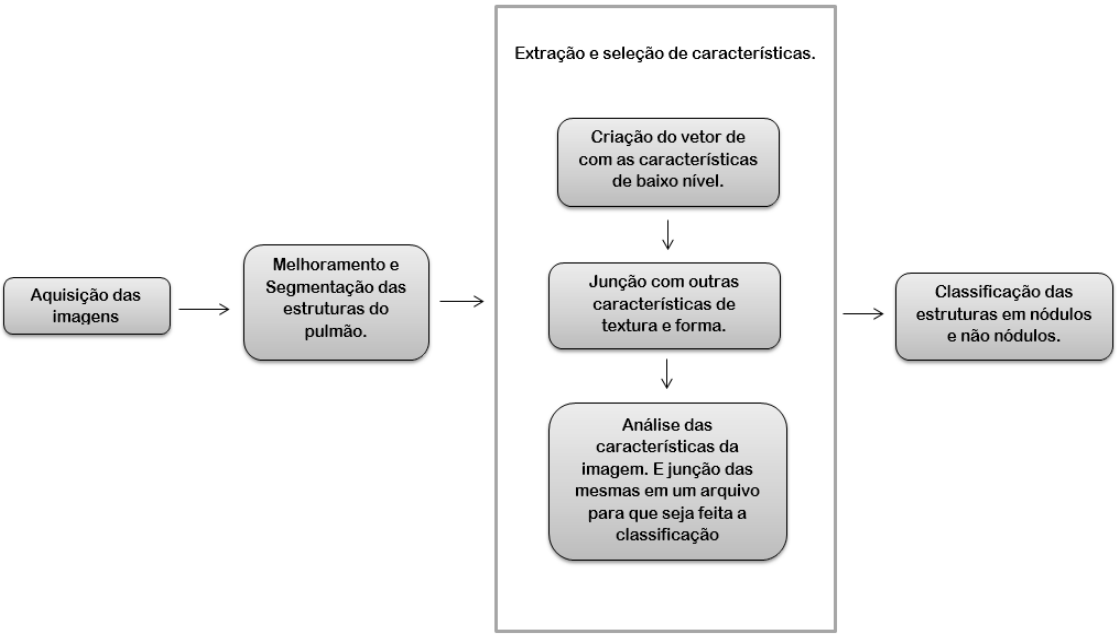


Figura 9. Fluxograma do processo de um CAD.

Percebe-se que a Figura 9 ilustra desde a aquisição das imagens até sua classificação. Analisando o processo, percebe-se que o processo de aquisição do vetor de baixo nível acontece através do uso de imagens já segmentadas, afim de termos uma vetor mais conciso.

3.1 Extração de Características

A técnica de vetores de quantificação baseia-se na criação de vetores de características, os quais comtemplam características de cor, textura, entre outras já citadas neste trabalho, tal como diâmetro mínimo e contorno das estruturas nodulares. Para a criação dos vetores foi tomado como base dados adquiridos através da criação de um histograma e de dados coletados de outras bases de conhecimentos, tal como os citados por (CARVALHO FILHO, 2013). A Figura 10 mostra a representação de um histograma obtido com a análise de uma das imagens fornecidas pela base.

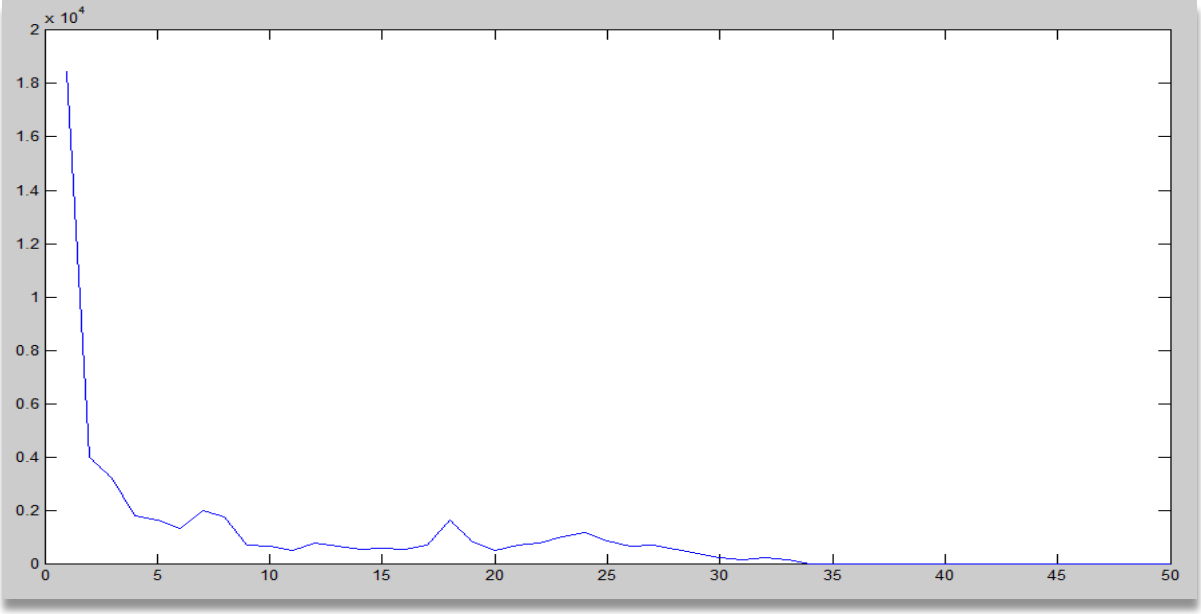


Figura 10. Histograma utilizado para a extração de características.

O histograma mostrado na Figura 10 representa a distribuição dos tons de cinza na região pulmonar, em que no eixo vertical temos a quantidade de ocorrências e no eixo horizontal temos os níveis de cinza presentes na imagem analisada.

No processo de construção do vetor utilizado na extração de características, os aspectos da imagem são divididos em dois vetores, um de alto nível, utilizado para a segmentação, e um de baixo nível, utilizado efetivamente na extração. A partir da criação do vetor de baixo nível, o mesmo é aplicado nas imagens, previamente segmentadas, sendo iniciado o processo de comparação e extração

das características, tal como forma e intensidade.

Após a extração é aplicada uma filtragem baseada em regras, intensidade dos pixels, visando refinar o resultado obtido e diminuir a incidência de falsos positivos, proporcionando assim, uma maior acurácia na seleção de características da imagem dos possíveis candidatos a nódulos.

A partir deste ponto, foi reproduzido os resultados conseguidos por (HAO *et al.*, 2013), e então foi iniciado o processo de otimização na criação dos vetores, em que foi feita a adição de características obtidas a partir da análise dos resultados de outras técnicas estudadas e descritas na revisão de literatura, como por exemplo, a esfericidade e o diâmetro da estrutura analisada. Além de terem sido feitas otimizações nos parâmetros utilizados na filtragem baseada em regras.

Para a construção do método proposto foram necessários estudos sobre as métricas utilizadas por (HAO *et al.*, 2013), tal como dados sobre o intervalo de textura adotado para indicar nas imagens de Tomografia Computadorizada (TC) onde encontravam-se os possíveis candidatos a nódulos pulmonares. Outro parâmetro analisado foi a esfericidade que indica quanto esférica é a estrutura analisada, pois apesar de não ter forma definida geralmente os possíveis candidatos a nódulos são, de certa forma, um pouco esféricos.

Carvalho Filho (2013) descreve como é feita a obtenção de dados como a esfericidade e diâmetro mínimo da região analisada na imagem médica para que a estrutura seja considerada candidato a nódulo. Essa informação foi utilizada na construção dos vetores de quantificação e na retirada de informações da imagem para obtenção dessas medidas. Para que fossem posteriormente comparada e até tivesse seu valor integrado a filtragem baseada em regras, realizada após a extração das características.

Ao fim da otimização do algoritmo, ou seja, da introdução de características não abrangidas pelos vetores como a esfericidade, foram feitos testes, a fim de comparar a técnica proposta por (HAO *et al.*, 2013) e o melhoramento na criação dos vetores e na filtragem baseada em regras, além da inclusão de novos parâmetros realizada no presente estudo.

3.1.1 MOMENTOS INVARIANTES AFINS

Momentos invariantes afins é um descritor de forma deduzido por (FLUSSER, 1993), no qual os autores encontraram um novo conjunto de invariantes, estas entretanto, invariantes a transformações afins em geral. Este descritor foi utilizado neste trabalho por reconhecer formas que sofrem alterações afins, como irregularidades em sua superfície.

3.2 Classificação

Após a fase de extração de características foi feita a classificação das estruturas com base nos aspectos extraídos. Para tanto, foi utilizada a ferramenta *Waikato Environment for Knowledge Analysis* (Weka) (KRISHNAIAH *et al.*, 2013) que consiste em uma coletânea de algoritmos utilizados com base na interface do programa.

O classificador escolhido para ser utilizado na análise das características extraídas, foi a *Support Vector Machine* (SVM). A SVM foi escolhida por ter uma boa capacidade de generalização e robustez em grandes dimensões, além de ter uma teoria bem definida e se adequar ao problema (LORENA, 2003).

4 RESULTADOS E DISCUSSÕES

Para chegar aos resultados ilustrados na Tabela 1, foram feitas as alterações descritas na Seção 3 e testes com três bases de imagens já segmentadas, porém com duas formas de segmentações diferentes. A primeira (Base 1) é composta de 50 imagens de TC, no processo de segmentação das mesmas não foi utilizado o crescimento de região. A segunda base (Base 2) é composta de 155 imagens, as quais foram processadas utilizando o crescimento de região. E a terceira base (Base 3) é uma junção das duas primeiras, as imagens segmentadas obtidas na primeira e segunda base.

Tabela 1. Comparação entre o método de Hao e a otimização descrita por este trabalho.

Índices/ Algoritmos	Sensibilidade	Falso Positivo (FP)	Kappa (K)	Verdadeiro Positivo (VP)	Imagens Usadas
Algoritmo de Hao	96,9%	138,7	-	4335,49	205
Algoritmo Proposto (Base 1)	98,3%	39,0	0.703	2255,12	50
Algoritmo Proposto (Base 2)	98%	122,8	0.678	6017,2	155
Algoritmo Proposto (Base 3)	98%	135,0	0.912	6615,0	205

A Tabela 1 mostra dados relevantes, tal como o número de falsos positivos que são estruturas que podem ser identificadas como candidatos a nódulos, porém não são. Os verdadeiros Positivos (VP), são os reais candidatos a nódulos acertadamente encontrados.

E a sensibilidade, é o parâmetro que indica a taxa de acertos do algoritmo para a amostra fornecida, tendo como resultado a divisão dos verdadeiros positivos pelo somatório entre os verdadeiros positivos e os falsos positivos, sendo calculado com

base na Equação 1. Há também o índice Kappa que mede a concordância entre diferentes medidas, fornecendo indicações que ajudam a saber quanto legítimas são as interpretações encontradas.

$$Sensibilidade = \frac{VP}{\mathrm{VP} + \mathrm{FP}} \qquad (1)$$

O índice Kappa segue uma indicação que varia de 0 a 1, em que quanto mais se aproxima de 1 (um) mais confiáveis são os resultados obtidos. E quanto mais se aproximam de 0 (zero) menos confiáveis eles se tornam. Sendo que os valores menores que 0,4 são tidos como ruins, entre 0,4 e 0,6 são considerados aceitáveis, de 0,6 a 0,8 são considerados bons e acima de 0,8 são considerados excelentes.

Conforme mostrado na Tabela 1, podemos perceber que ouve uma diminuição significativa na incidência de falsos positivos, de 138,7 para 135,0 e consequente aumento dos verdadeiros positivos, quando comparamos os resultados do trabalho de (HAO *et al.*, 2013) e os conseguidos com o algoritmo proposto. Existe, também um aumento na sensibilidade passando de 96,9% para 98,3%, quando comparamos com os resultados obtidos quando usada a terceira base de imagens.

Percebemos que o maior índice de sensibilidade se deu quando foi utilizada a base de imagens 1. E que não foi possível mensurar o índice Kappa para os resultados conseguidos por (HAO *et al.*, 2013), tendo em vista que o mesmo não é fornecido pelo autor em seu estudo.

Foi conseguida uma maior precisão nos resultados, que os obtidos por (HAO *et al.*, 2013), independentemente de qual das 3 bases tenha sido utilizada. Notamos, também, que quando a base 1 foi incorporada a base 2, criando-se a terceira base, ouve uma queda na sensibilidade do método proposto, porém não muito expressiva.

A Tabela 1 mostra apenas os resultados conseguidos na fase de extração das características. Após a extração foi aplicado o classificador SVM nas características extraídas. Foi conseguido um resultado com uma sensibilidade de 98%. O índice de falsos positivos encontrado foi de 134 estrutura/exame e o índice Kappa chegou a 0,9, sendo que foram analisados 1010 candidatos a nódulos. Para tanto, utilizando-se as características extraídas a partir de imagens contidas na terceira base de

imagens.

Tabela 2. Resultados obtidos na fase de Classificação no Weka.

Índices/ Algoritmos	Falsos Positivos (FP)	Verdadeiro Positivo (VP)	Quantidade de estruturas analisadas	Imagens Utilizadas
Algoritmo Proposto (Base 1)	30.0	70.0	100	50
Algoritmo Proposto (Base 2)	40.0	270.0	310	155
Algoritmo Proposto (Base 3)	50.0	450.0	600	205

A Tabela 2 mostra os resultados obtidos após a classificação das características extraídas utilizado a metodologia proposta, através da classificação utilizando a máquina de vetor de suporte. Nesta Tabela estão contidos dados acerca dos falsos positivos, verdadeiros positivos, a quantidade de estruturas analisadas e o número de imagens/exames utilizados. A partir dos resultados obtidos na Tabela 2, pode-se concluir que ouve um aumento na sensibilidade da metodologia proposta, já que a quantidade de falsos positivos teve uma queda considerável.

.

5 CONCLUSÃO E TRABALHOS FUTUROS

Notou-se ao longo da pesquisa que a extração e seleção de características e a classificação são processos presentes na maioria dos sistemas que identificam os candidatos a nódulos em diversas áreas médicas. E também encontram-se em outras áreas que fazem análise de imagens e que envolvem processamento das mesmas.

O intuito deste trabalho foi o estudo e a implementação dos algoritmos de extração e seleção de características, a fim de verificar como os algoritmos estudados fazem este processo e quais os pros e contras de cada um deles. Ao final foi escolhido um para ser implementado e melhorado. A partir deste melhoramento viu-se que a técnica melhorada obteve um melhor resultado do que a original. Além de diminuir a incidência de falsos positivos.

Com os resultados conseguidos, espera-se ajudar os profissionais da área médica a diminuir o tempo de análise das imagens e dar uma segunda opinião sobre as estruturas analisadas. Espera-se, também, contribuir para pesquisas futuras a fim de diminuir o tempo para que se tenha o diagnóstico sobre o câncer de pulmão e consequentemente aumento da taxa de sobrevida.

5.1 Trabalhos Futuros

Como trabalhos futuros serão feitas a integração com as fases de segmentação e melhoramento das imagens. A fim de construir um sistema auxiliado por computado, para possibilitar a detecção mais precisos e tornar menos enfadonhos o processo para os profissionais da área, além de servir como uma segunda opinião para estes profissionais. Deve ser feita, também, a procura de novas bases de imagens, a fim de se validar a metodologia em outras bases. Além de ser feita a validação deste novo sistema.

Pode-se também fazer a integração de outras técnicas de extração de características, tal como a integração com outros tipos de algoritmos estudados. A fim de termos uma classificação exata e eficiente que a descrita nos trabalhos da

literatura.

5.2 Trabalhos Publicados

Como resultados alcançados foi conseguida uma aceitação na Conferência Latinoamericana de Informática (CLEI) em Arequipa no Peru de um artigo com o mesmo título deste trabalho. Houve, contudo, a apresentação deste trabalho em forma de banner no XXIV Seminário de Iniciação Cientifica da UFPI, que foi realizado na cidade de Teresina -PI. E na Mostra do Semiárido Digital, ocorrido na cidade de Picos-PI.

REFERÊNCIAS BIBLIOGRÁFICAS

ARMATO, S. G. (2011).**The lung image database consortium (LIDC) and image database resource initiative (IDRI): A complete reference database of lung nodules on CT scans**. Med. Phys., vol. 38, pp. 915-931.

ARAUJO, Lucas Moreno; OLIVEIRA, Fernando Luiz. **Estudo dos descritores de características primitivas.** Encontro de Computação e Informática do Tocantins.

BAGCI, Ulas; BRAY, Mike; CABAN, Jesus; YAO, Jiahhua, MOLLURA, Daniel J. .(2012). **Computer-assisted detection of infectious lung diseases: A review**. Computerized Medical Imaging and Graphics.72-84.

CARVALHO FILHO, A. O. de.(2013) . **Detecção automática de nódulos pulmonares solitários usando quality threshold clustering e mvs**. Dissertação de Mestrado na área de Ciência da Computação. (Programa de Pós-Graduação em Engenharia de Eletricidade) - Universidade Federal do Maranhão, São Luís.

CAO, Perg; YANG, Jinzhu; LI, Wei; ZHAO, Dazhe; ZAIANE, Osmar (2014). **Ensemble-based hybrid probabilistic sampling for imbalanced data learning in lung nodule cad**.

CRISTIANINI, Nello; TAYLOR, John S. **An Introduction to Cambridge University Press**, Mar. 2000.

FLUSSER, J.; SUK, T. Pattern Recognition by Affine Moment Invariants. **Pattern Recognition,** Oxford, V.26, p.167-174, 1993.

GONZALEZ, R. C.; WOODS, R. E. (2002) . **Digital Image Processing**. São Paulo: Edgard Blücher. 443 p.

HAN, Hao ; LI, Lihong ; MEMBER, Senior; HAN, Fangfang; SONG, Bowen; MOORE, Willian; LIANG, Zhengrong.(2013) .**Fast and Adaptative Detection of Pulmonary Nodules in Thoracic CT Images Using a Hierarchical Vector Quantization Scheme**.

Instituto Nacional do Câncer - INCA. Mistério da Saúde. Câncer de pulmão. Disponível em: <http://www2.inca.gov.br/wps/wcm/connect/cancer/site/oquee>. Acessado em: 10 de Janeiro de 2016.

KRISHNAIAH, V.; NARSIMHA, G.; CHANDRA, N. Subhash.(2013). **Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques**. ISBN-0975-9646

LORENA, Ana Carolina; DE CARVALHO, André C. P. L. F.(2003). **Introdução às maquinas de vetores suporte(Support Vector Machne)** . Nº 192. ISBN- 0103-2569.

LAURENCE, J. **Biologia: ensino médio**. Vol. Único – 1 ed. São Paulo: Nova geração, 2005.

Lung Cancer Alliance – LCA. Lung Cancer Symptoms. Disponível em: http://www.lungcanceralliance.org/get-information/what-is-lung-cancer/symptoms-of-lung-cancer/. Acesso em: 28 de Janeiro de 2016.

MARAR, João Fernando; FILHO, Edson Costa de Barros Carvalho.(2013). **Reconhecimento de Padrões Numéricos Através da Transformada Karhumen-Loève**.

NETTER, Frank H.. **Atlas de Anatomia Humana**. 2ed. Porto Alegre: Artmed, 2000.

REY, J.; TRUNDLE, P.; JIANG, J.(2010). **Medical image analysis with artificial neural networks**.617-631.

SOUSA, J. R. F. S. (2011).**Metodologia para detecção automática de nódulos pulmonares**. Dissertação de Mestrado na área de Ciência da Computação. (Programa de Pós-Graduação em Engenharia de Eletricidade) - Universidade Federal do Maranhão, São Luís.

TERMO DE AUTORIZAÇÃO PARA PUBLICAÇÃO DIGITAL NA BIBLIOTECA "JOSÉ ALBANO DE MACEDO"

Identificação do Tipo de Documento

( ) Tese

( ) Dissertação

(x) Monografia

( ) Artigo

Eu, **Ivo Alves de Oliveira**, autorizo com base na Lei Federal nº 9.610 de 19 de Fevereiro de 1998 e na Lei nº 10.973 de 02 de dezembro de 2004, a biblioteca da Universidade Federal do Piauí a divulgar, gratuitamente, sem ressarcimento de direitos autorais, o texto integral da publicação Classificação de Nódulos Pulmonares Utilizando Vetores de Quantificação de minha autoria, em formato PDF, para fins de leitura e/ou impressão, pela internet a título de divulgação da produção científica gerada pela Universidade.

Picos-PI 26 de fevereiro de 2016.

Ivo Alves de Oliveira.
Assinatura