Airton Gaio Júnior

Um método para classificação de opinião em vídeo combinando expressões faciais e gestos

Manaus

Março, 2017

Airton Gaio Júnior

Um método para classificação de opinião em vídeo combinando expressões faciais e gestos

Dissertação apresentada ao Instituto de Computação da Universidade Federal do Amazonas, para a obtenção do Grau de Mestre em Informática.

Universidade Federal do Amazonas - UFAM

Instituto de Computação - IComp

Programa de Pós-Graduação em Informática

Orientadora: Profª. Drª. Eulanda Miranda dos Santos

Manaus

Março, 2017

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Gaio Junior, Airton
J143i Um método para classificação de opinião em vídeo combinando expressões faciais e gestos / Airton Gaio Junior. 2017
73 f.: il. color; 31 cm.

Orientadora: Eulanda Mirada dos Santos
Dissertação (Mestrado em Informática) - Universidade Federal do Amazonas.

1. Reconhecimento multimodal de opinião. 2. Expressões faciais e corporais. 3. Codificadores. 4. Fusão baseada em decisão. I. Santos, Eulanda Mirada dos II. Universidade Federal do Amazonas III. Título

FOLHA DE APROVAÇÃO

"Um método para classificação de opinião em vídeo combinando expressões faciais e gestos"

AIRTON GAIO JÚNIOR

Dissertação de Mestrado defendida e aprovada pela banca examinadora constituída pelos Professores:

Profa. Eulanda Miranda dos Santos - PRESIDENTE

Prof. José Reginaldo Hughes Carvalho - MEMBRO INTERNO

Prof. José Luiz de Souza Pio - MEMBRO EXTERNO

Prof. Waldir Sabino da Silva Júnior - MEMBRO EXTERNO

Manaus, 05 de Abril de 2017

Este trabalho é dedicado às crianças adultas que,
quando pequenas, sonharam em se tornar cientistas.

Agradecimentos

Quero agradecer principalmente ao Grande Arquiteto do Universo pela dádiva da vida, à minha família e aos meus amigos que compreenderam, acreditaram e incentivaram a mim durante todo esse tempo.

Agradecimento especial à Profª. Drª. Eulanda Miranda dos Santos, minha orientadora e exemplo profissional, por ter me guiado com maestria durante todo o processo e pela confiança. Estendo esse agradecimento a todos os professores do IComp em especial aos ligados a área de Aprendizado de Máquinas e Visão Computacional que contribuíram de sobremaneira para que a produção desta dissertação fosse possível.

Agradeço ao Instituto de Computação - IComp, a Universidade Federal do Amazonas - UFAM e ao Instituto Federal de Educação - IFAC pela oportunidade oferecida.

“Nem tudo que se enfrenta pode ser modificado,
mas nada pode ser modificado até que seja enfrentado.”
(Albert Einstein)

Resumo

Um grande número de pessoas compartilha suas opiniões através de vídeos, gerando uma gama de dados incalculável. Esse fenômeno tem despertado elevado interesse de empresas em obter, a partir de vídeos a percepção do grau de sentimento envolvido na opinião das pessoas. E também tem sido uma nova tendência no campo de análise de sentimentos, com importantes desafios envolvidos. A maioria das pesquisas que abordam essa problemática utiliza em suas soluções a combinação de dados de três fontes diferentes: vídeo, áudio e texto. Portanto, são soluções baseadas em modelos complexos e dependentes do idioma, ainda assim, apresentam baixo desempenho. Nesse contexto, este trabalho busca responder a seguinte pergunta: é possível desenvolver um método de classificação de opinião que utilize somente vídeo como fonte de dados, e que obtenha resultados superiores ou equivalente aos resultados obtidos por métodos correntes que usam mais de uma fonte de dados? Como resposta a essa pergunta, é apresentado neste trabalho um método de classificação de opinião multimodal que combina informações de expressão facial e de gesto do corpo extraídas de vídeos *on-line*. O método proposto utiliza codificação de características para melhorar a representação dos dados e facilitar a tarefa de classificação, a fim de predizer a opinião exposta pelo usuário com elevada precisão e de forma independente do idioma utilizado nos vídeos. Com objetivo de testar o método proposto foram realizados experimentos com três bases de dados públicas e com três *baselines*. Os resultados dos experimentos mostram que o método proposto é em média 16% superior aos *baselines* em termos de acurácia e ou precisão, apesar de utilizar apenas dados de vídeo, enquanto os *baselines* utilizam vídeo, áudio e texto. Como forma de demonstrar portabilidade e independência de idiomas do método proposto, este foi treinado com instâncias de uma base de dados que tem opiniões expressas exclusivamente em inglês, e testado em uma base de dados cujas opiniões são expressas exclusivamente no idioma espanhol. O percentual de 82% de acurácia alcançado nesse teste indica que o método proposto pode ser considerado independente do idioma falado nos vídeos.

Palavras-chave: Reconhecimento multimodal de opinião, Expressões faciais e corporais, Codificadores, Fusão baseada em decisão.

Abstract

A large amount of people share their opinions through videos, generates huge volume of data. This phenomenon has lead companies to be highly interested on obtaining from videos the perception of the degree of feeling involved in people's opinion. It has also been a new trend in the field of sentiment analysis, with important challenges involved. Most of the researches that address this problem propose solutions based on the combination of data provided by three different sources: video, audio and text. Therefore, these solutions are complex and language-dependent. In addition, these solutions achieve low performance. In this context, this work focus on answering the following question: is it possible to develop an opinion classification method that uses only video as data source and still achieving superior or equivalent accuracy rates obtained by current methods that use more than one data source? In response to this question, a multimodal opinion classification method that combines facial expressions and body gestures information extracted from online videos is presented in this work. The proposed method uses a feature coding process to improve data representation in order to improve the classification task, leading to the prediction of the opinion expressed by the user with high precision and independent of the language used in the videos. In order to test the proposed method experiments were performed with three public datasets and three baselines. The results of the experiments show that the proposed method is on average 16% higher that baselines in terms of accuracy and precision, although it uses only video data, while the baselines employ information from video, audio and text. In order to verify whether or not the proposed method is portable and language-independent, the proposed method was trained with instances of a dataset whose language is exclusively English and tested using a dataset whose videos are exclusively in Spanish, applied in the conduct of the tests. The 82% of accuracy achieved in this test indicates that the proposed method may be assumed to be language-independent.

Keywords: Multimodal opinion recognition, Face and body expressions, Encoders, Decision-based fusion.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ANN	*Artificial Neural Network.*
AUC	*Area Under the Curve.*
BoVW	*Bag of Visual Words Model.*
BoW	*Bag-of-Words.*
DNN	*Deep Neural Network.*
ELM	*Extreme Machine Learn.*
FACs	*Facial Action Coding System.*
FV	*Fisher Vetor.*
GMM	*Gaussian Mixture Model.*
HMM	*Hidden Markov Model.*
HOG	*Histogram of Oriented Gradients.*
MAE	*Mean Absolute Error.*
MFCCs	*Mel-Frequency Cepstral Coefficients.*
MHI	*Motion History Image.*
MOSI	*Multimodal Opinion-level Sentiment Intensity.*
MOUD	*Multimodal Opinion Utterances Dataset.*
MPEG	*Moving Picture Experts Group.*
MPQA	*Multi-Perspective Question Answering.*
NAQ	*Normalized Amplitude Quotient.*
PCA	*Principal Component Analysis.*
RGB	*Red Green Blue.*
ROC	*Receiver Operating Characteristic.*
RP	Reconhecimento de Padrões.
SVM	*Support Vector Machine.*
SVR	*Support Vector Regression.*

TFPTaxa de Falso Positivo.

TVPTaxa de Verdadeiro Positivo.

VJ *Viola-Jones.*

VLAD *Vector of Locally Aggregated Descriptors.*

WEKA *Waikato Environment for Knowledge Analysis.*

Lista de símbolos

α	Letra grega minúscula alfa.
γ	Letra grega minúscula gama.
θ	Letra grega minúscula teta.
Θ	Letra grega maiúscula teta.
λ	Letra grega minúscula lambda.
μ	Letra grega minúscula miu.
π	Letra grega minúscula pi.
σ	Letra grega minúscula sigma.
τ	Letra grega minúscula tau.
Φ	Letra grega maiúscula fi.
Ω	Letra grega maiúscula ômega.
$\in$	Pertence.

Sumário

1 Introdução

A análise de sentimentos é uma habilidade inerente aos seres humanos sendo considerada um instrumento de fundamental importância na interação natural entre as pessoas (PALEARI; CHELLALI; HUET,2010). A comunicação dos sentimentos ocorre pela combinação de multimodalidade, ou seja, linguagem, tom da voz, expressão facial, gesto da cabeça, movimento do corpo, postura e poses, que são processados em um refinado mecanismo humano de fusão de dados. No entanto, este mecanismo não é inerente às máquinas, as quais são menos hábeis para reproduzir e compreender estas habilidades (GUNES; PICCARDI,2005).

As pesquisas desenvolvidas na área de análise sentimentos tornaram-se recorrentes nos últimos anos, possibilitando às máquinas a interpretar sentimentos, emoções e até mesmo o afeto das pessoas (GUNES et al.,2013). Compreender o grau de sentimento expresso por pessoas a respeito de suas opiniões com base em vídeos é uma nova tendência no campo de análise de sentimentos, a qual enfrenta importantes desafios. Como forma de exemplificar esses desafios podemos citar a diversidade nas expressões da opinião manifestada por pessoas de diferentes lugares e idiomas, bem como, ambientes ruidosos presentes na maioria dos vídeos gravados no *mundo real*.

Atualmente, um grande número de pessoas compartilha suas opiniões, histórias e comentários através de postagens de vídeos em sites como o *Youtube*, *Vine* e *Vimeo*. Dentre esses, certamente o *Youtube* é o site mais popular, o qual e recebe via *upload* mais de 300 horas de vídeo a cada minuto, formando, portanto, uma incalculável massa de dados para o processamento desta questão. Um grande número de empresas, investidores e consumidores têm observado este fenômeno com muita atenção no intuito de desenvolver melhores aplicações de mineração de opiniões baseada em vídeos *on-line* (ZADEH,2015). A partir disso, há expectativa por parte dos interessados em obter uma percepção do grau de sentimento envolvido na opinião das pessoas, expressa em massa por esses meios de comunicação *on-line*, de maneira livre a respeito de produtos, serviços ou qualquer assunto considerado relevante, fazendo com que os interessados no assunto possam oferecer, por exemplo, uma recomendação mais adequada e talvez criar perfis mais robustos destas pessoas.

A maioria dos trabalhos que tratam de análise de opinião, reconhecendo minimamente as intensidades posititva, negativa e neutra, utiliza a combinação de dados multimodais como textos, vídeos e áudios. O objetivo da utilização de mais de uma modalidade é aumentar o rigor e a acurácia das estimativas. Muitos estudos voltados à problemática de análise da opinião comprovam o potencial da fusão de dados, apresentando melhores resultados com a fusão das modalidades quando comparados aos resultados das modalidades isoladas (MORENCY; MIHALCEA; DOSHI, 2011), (ROSAS; MIHALCEA; MORENCY,2013), (PORIA et al.,2016), (ZADEH et al.,2016). Apesar disso, os resultados obtidos com o uso das fusões relatados nestas pesquisas, em geral, são abaixo de 80% na taxa de acerto e/ou precisão.

Ademais, lidar com informações de diferentes fontes de dados, em geral, aumenta con-

sideravelmente a complexidade do modelo e dificulta a replicação devido a várias questões, as quais podemos citar: **1)** a tarefa de obtenção de dados multimodais pode não ser trivial, como no caso de arquivos textos contendo as falas, que na maioria das vezes são transcritos manualmente; e **2)** o dados de fontes diferentes podem não estar corretamente sincronizados, dificultando todo o processo de fusão das modalidades. Em termos de vídeo, há desafios como a diversidade na expressão facial e gestos corporais manifestados por pessoas de diferentes lugares e línguas, bem como, ambiente ruidoso encontrado na maioria dos vídeos gravados com tipos de equipamentos diversos em ambientes sem nenhum tipo de controle, tendo variados planos de fundo, iluminação e escalas. Outro fator desfavorável aos modelos apresentados na literatura é a utilização de *software* comerciais como parte da solução, principalmente na etapa de extração automática de características.

Tendo em vista toda a problemática envolvida na análise de opinião e seus desafios destacados acima, bem como a forma complexa com que as pesquisas existentes na literatura têm proposto soluções, é possível desenvolver um método de classificação de opinião que utilize somente a fonte de dados de vídeo e que obtenha resultados superiores ou equivalente aos resultados obtidos por métodos correntes que usam mais de uma fonte de dados?

Muito embora a maioria maciça das pesquisas atuais trabalhe normalmente com três fontes diferentes de dados para analisar opinião, especificamente áudio, vídeo e texto, nossa pesquisa opta por simplificar o modelo ao propor um método de análise de opinião baseado em informações extraídas das modalidades de expressão facial e do gesto do corpo provenientes somente de vídeo, com intuito de construir um novo quadro de análise de sentimento multimodal visando a predição da opinião exposta pelo usuário. A escolha destas modalidades para representar o grau de sentimento expresso na opinião de uma pessoa está embasada no estudo deGunes et al. (2013), o qual relata que a utilização dos dados referentes aos gestos do corpo pode ser útil, pois, quando somos incapazes de dizer o estado emocional da face, que normalmente é usada nestes processos, ainda assim, podemos ler claramente a ação a partir da visão do corpo. Corroboram com esta afirmação os estudos apresentados porPanning et al.(2012) eGunes et al.(2013), os quais visam o reconhecimento dos sentimentos com as modalidades da expressão facial e gestos corporais obtendo bons resultados. Além do mais, pelo fato do método não utilizar dados de texto com as falas das opiniões e nem os arquivos de áudio, nós acreditamos que seja possível a construção de um modelo independente/invariável ao idioma falado pelas pessoas no vídeo.

O método proposto é dividido em três macro etapas: **1)** Extração de Características; **2)** Codificação de Características; e **3)** Fusão de decisão. Na primeira etapa, nós empregamos descritores de características clássicos e largamente difundidos na literatura, como o *Motion History Image - MHI* que, resumidamente, representa o movimento de uma opinião capturada em uma sequência de vídeo; e *Histogram of Oriented Gradients - HOG*, que extrai informações relacionadas à orientação do gradiente de uma imagem. O diferencial de nosso trabalho está na segunda fase, com o uso de codificação das características geradas pelos descritores clássicos. Em nosso método, os vetores de características gerados são codificados por técnicas baseadas no conceito de *Bag of Visual Words Model - BoVW*, o qual vem sendo amplamente usado em visão computacional e tem sido adotado como importante modelo de representação de imagens, sobretudo em problemas envolvendo reconhecimento de ações humanas, conforme é possível

comprovar no completo estudo desenvolvido porPeng et al.(2014).

Com base em recentes pesquisas, a abordagem de codificação *Fisher Vetor - FV* é a mais precisa no reconhecimento de ações humanas e oferece vantagens significativas, como por exemplo, FV contribui para o aumento na taxa de classificação mesmo que um classificador simples como *Support Vector Machine - SVM* com *kernel* linear, ou seja, sem necessidade de ajustes de parâmetros, seja utilizado como relatam as pesquisas deBiliński(2014) eOneata, Verbeek e Schmid(2013). Diante desses resultados, FV é utilizado neste trabalho e comparado ao método de codificação *Vector of Locally Aggregated Descriptors - VLAD*, ambos são baseados no conceito BoVW.

Por fim, dois vetores de características são gerados: um com dados da expressão facial e outro com dados de gestos. Cada vetor de características é utilizado para representar os dados e, dessa forma, treinar um classificador. Dado que dois classificadores são obtidos, a última fase de nosso método envolve a fusão da decisão dos dois classificadores.

Nós utilizamos a forma de fusão de dados conhecida na literatura como fusão baseada em decisão, visando a integração dos resultados obtidos das modalidades das expressões faciais e gestos corporais como forma de melhorar os resultados em termos de classificação.

Objetivos

Esta seção descreve o objetivo geral e os objetivos específicos deste trabalho.

Objetivo Geral

Desenvolver um método de classificação de opinião multimodal baseado em informações combinadas das modalidades de expressão facial e de gesto do corpo provenientes de vídeos *on-line*, que utilize codificação de características para melhorar a representação dos dados e facilitar a tarefa de classificação, a fim de predizer a opinião exposta pelo usuário com elevada precisão e de forma independente do idioma utilizado nos vídeos. Para alcançar o objetivo proposto é necessário desenvolver atividades especificas, a saber:

1. Implementar estratégias de extração de características de face e gesto do corpo a partir de sequências de vídeo a fim de realizar o reconhecimento de emoções em vídeos *on-line*.

2. Comparar e avaliar métodos de codificação de vetores de características na predição da opinião.

3. Realizar a fusão das modalidades de face e corpo com uso de técnicas apropriadas que melhorem o desempenho da classificação.

4. Avaliar a portabilidade do método utilizando base de dados diferentes e verificar sua independência com relação ao idioma falado.

Contribuições do trabalho

Embora existam pesquisas abordando a problemática de classificação de opinião multimodal, o método proposto neste trabalho apresenta algumas contribuições importantes, a saber: **1)** utiliza as modalidades de face e de corpo provenientes somente de vídeos em sua solução; **2)** Pelo fato de não utilizar dados de áudio e texto, comumente empregados em outras pesquisas, não se restringe ao idioma falado; **3)** emprega método de codificação, o qual contribui sobremaneira ao processo de classificação, pois possibilita que elevadas taxas de classificação correta sejam obtidas, mesmo com o emprego de classificadores simples, sem a necessidade de ajustes de parâmetros.

Organização do Documento

O presente documento está organizado da seguinte forma: no capítulo2discorremos sobre a fundamentação teórica necessária para o entendimento da pesquisa. No capítulo3, os trabalhos relacionados são apresentados. No capítulo4a arquitetura do método empregada no desenvolvimento da pesquisa é discutida. O capítulo5apresenta os experimentos, as bases de dados investigadas e os resultados obtidos comparando-os com os *baselines*. Por fim, no capítulo 6apresentamos a conclusão, bem como, as propostas de trabalhos futuros.

2 Referencial Teórico

Neste capítulo são apresentados alguns dos principais conceitos necessários para o entendimento de problemas que envolvem a análise e o reconhecimento de opinião com base em dados multimodais, especificamente dados oriundos de expressões da face e de gestos do corpo extraídos de sequências de vídeos, dado que esse tipo de sistema é o foco deste trabalho.

A seção2.1apresenta conceitos relacionados a reconhecimento de padrões e descreve as etapas envolvidas no processo. Em seguida, o procedimento geral adotado em tarefas de reconhecimento de opinião em vídeo e algumas das técnicas que compõem essa atividade são descritas. Na seção2.2introduzimos o algoritmo *Viola-Jones* para a detecção da face, olhos, boca e nariz. A seção2.3descreve as técnicas de extração de características utilizadas na literatura, que são investigadas neste trabalho. Imediatamente após, apresentamos a redução de dimensionalidade com uso de *Principal Component Analysis - PCA*, na seção2.4. Na seção2.5são descritos os métodos de codificação: *Fisher Vector - FV* e *Vector of Locally Aggregated Descriptors - VLAD*. Logo após, na seção2.6nós apresentamos os conceitos do classificador SVM empregado nesta pesquisa. Por fim, a seção2.7discorre sobre a fusão de dados multimodais.

Reconhecimento de Padrões - RP

SegundoTrucco e Verri(1998), Reconhecimento de Padrões - RP envolve a descrição do objeto ou modelo disponível, pois, não se pode reconhecer o que não se conhece ainda. Um exemplo de processo de RP é a comparação de dados desconhecidos com um banco de dados de modelos conhecidos. Ao ocorrer a correspondência do modelo a um subconjunto de dados, por exemplo, configurações particulares de contornos ou formas específicas, dizemos que o objeto foi encontrado e coincide com os dados do modelo.

Os passos fundamentais no processamento do RP são relatados porGonzalez e Woods (2010), e, estão divididos em várias etapas, algumas, porém, podem variar de acordo com o problema investigado. A Figura1apresenta o diagrama com as principais etapas presentes em um sistema de RP. A seguir são apresentadas as descrições das etapas exibidas no diagrama:

- A primeira etapa se refere à *aquisição de imagens*. A aquisição pode ser tão simples quanto receber uma imagem ou uma sequência de imagens que já estejam em formato digital adequado.
- A etapa de *pré-processamento* em geral é responsável por uma série de processos. E exemplificando podemos citar: a conversão e transformação dos formatos dos dados, o redimensionamento de imagens, os filtros de redução dos ruídos, o realce de objetos específicos, de forma que, os resultados sejam mais adequados se comparados ao original, a compressão que lida com técnicas de redução do armazenamento, entre outros.

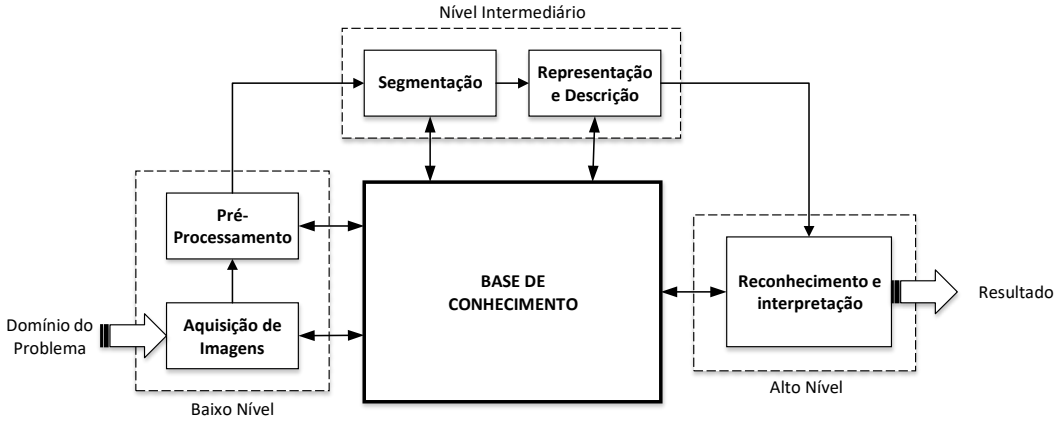


Figura 1 – Passos fundamentais no RP. Fonte: Adaptado de (GONZALEZ; WOODS,2010)

- A *segmentação* divide uma imagem pré-processada em suas partes ou objetos constituintes. A segmentação autônoma é uma das tarefas mais difíceis de todo o processo. Um procedimento de segmentação bem-sucedido aumenta as chances de sucesso na solução do problema que requer que os objetos sejam individualmente identificados. Por outro lado, algoritmos de segmentação fracos ou inconsistentes quase sempre ocasionam falhas no processamento. Em geral, quanto maior a precisão da segmentação, maiores serão as chances de sucesso no reconhecimento do objeto.

- A etapa de *representação e descrição* quase sempre parte da etapa anterior, a qual normalmente provê dados primários em forma de *pixels* correspondendo tanto à fronteira de uma região como a todos os pontos dentro dela. De qualquer forma, em ambos os casos, é necessário converter dados a uma forma adequada para o processamento computacional. A descrição, também conhecida como *seleção de características*, lida com a extração de atributos que resultam em alguma informação quantitativa de interesse ou que possam ser utilizados para diferenciar uma classe de objetos de outras classes.

- Na última etapa, *reconhecimento e interpretação*, é atribuído o rótulo a um objeto com base em sua descrição. A complexidade do processo de rotulação varia conforme o acerto na escolha dos valores analisados nas etapas anteriores.

- A *base de conhecimento* ilustrada na área central do diagrama se refere à base de dados do conhecimento adquirido. Esse conhecimento pode ser tão simples quanto o detalhamento de regiões de imagem na qual se sabe que a informação de interesse pode ser localizada, limitando, dessa forma, o processo de busca. Serve também para orientar a operação de cada etapa de processamento, além de controlar a interação entre as etapas.

Existem vários exemplos reais de aplicações de RP. Detecção de faces é uma dessas aplicações, conforme descrito na próxima seção.

Detecção de Face e de seus componentes: olhos, boca e nariz

Esta seção está embasada no artigo deLobban(2008), o qual destaca o uso do método de detecção de faces *Viola-Jones - VJ*. Segundo o autor, a essência do processo de detecção de faces de VJ é escanear uma sub-janela capaz de detectar faces através de uma dada imagem de entrada. Nesta pesquisa, optou-se pelo uso do algoritmo VJ para a detecção de face, pois, apresenta capacidade de processar imagens de forma extremamente rápida e alcança altas taxas de detecção.

A abordagem padrão dos demais algoritmos de detecção de objetos é redimensionar a imagem de entrada em diferentes tamanhos e em seguida executar o detector, enquanto *Viola-Jones* redimensiona o detector em vez da imagem de entrada. Em princípio pode parecer que ambas as abordagens são igualmente demoradas, entretanto, a solução de VJ cria uma escala de detector invariante que requer o mesmo número de cálculos independentemente da dimensão. Este detector é construído usando uma imagem integral e algumas características retangulares simples, obtendo assim, um detector de rosto extremamente rápido.

O primeiro passo do algoritmo *Viola-Jones* para a detecção de face é transformar a imagem de entrada em uma imagem integral. Isto é feito muito rapidamente usando uma representação intermediária para a imagem chamada de imagem integral. A imagem integral na posição x, y contém a soma dos *pixels* acima e à esquerda de x,y inclusive (VIOLA; JONES,2001):

$$ii(x, y) = \sum_{x^t \le x, y^t \le y} i(x^t, y^t)$$

onde $ii(x, y)$ é a imagem integral e $i(x^t, y^t)$ é a imagem original. Usando o seguinte par de recorrências:

$$s(x, y) = s(x, y - 1) + i(x, y),$$

$$ii(x, y) = ii(x - 1, y) + s(x, y)$$

sendo $s(x, y)$ a soma da linha acumulada, $s(x, -1) = 0$ e $ii(-1, y) = 0$ a imagem integral podendo ser calculada em uma passada pela imagem original. Isto permite o cálculo da soma de todos os *pixels* dentro de qualquer dado retângulo usando apenas quatro valores. Estes valores são os *pixels* na imagem integral que coincidem com os cantos do retângulo na imagem de entrada, conforme demostra a Figura2.

O algoritmo então analisa e determina uma sub-janela usando recursos constituídos por dois ou mais retângulos. A Figura3ilustra os diferentes tipos de características utilizados por *Viola-Jones*. SegundoLobban(2008), podemos compreender essas características como sendo a forma como o computador percebe uma imagem de entrada. Com este processo, há expectativa que algumas características obtenham valores maiores para uma face, facilitando assim a busca, se comparado com operações realizadas diretamente nos *pixels* brutos.

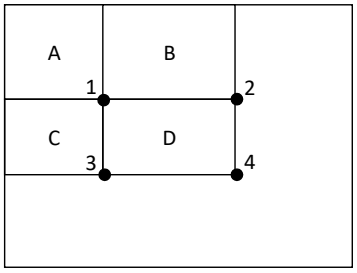


Figura 2 – A soma dos *pixels* dentro do retângulo *D* pode ser calculada com quatro referências de matriz. O valor da imagem integrante na localização 1 é a soma dos *pixels* no retângulo *A*. O valor na localização 2 é $A + B$, no local 3 é $A + C$, e a localização 4 é $A + B + C + D$. A soma dentro de *D* pode ser calculada $4 + 1 - (2 + 3)$. Imagem original em (VIOLA; JONES,2001)

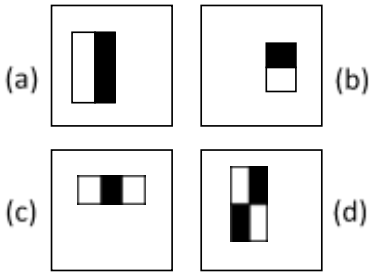


Figura 3 – Exemplo de características do retângulo mostrado em relação à janela de detecção. O valor da soma dos *pixels* que se encontram dentro dos retângulos brancos é subtraído da soma dos *pixels* nos retângulos escuros. Duas características do retângulo são mostradas em (a) e (b). A imagem em (c) mostra uma característica de três retângulos, e (d) um recurso de quatro retângulos. Imagem original em (VIOLA; JONES,2001)

O próximo passo do algoritmo de VJ é utilizar uma versão modificada do algoritmo de AdaBoost[1]. O algoritmo de aprendizagem é projetado para selecionar as características de um único retângulo que melhor separe os exemplos positivos dos negativos. Para cada característica é determinada a função de classificação de limiar ótimo, de modo que o número mínimo de exemplos seja classificado erroneamente.

Por fim, utiliza-se uma cascata de classificadores fortes que, ao invés de encontrar faces, descarta *não-faces*, pois é mais rápido descartar *não-faces* do que encontrar uma face, mesmo em imagens contendo muitos rostos.

De acordo comLobban(2008), o classificador é constituído em etapas, cada uma contendo um classificador forte. O trabalho de cada etapa é determinar se uma determinada sub-janela não é definitivamente uma face. Quando uma sub-janela é classificada como *"não-face"* por uma determinada fase, é imediatamente descartada, por outro lado, uma sub-janela classificada como talvez sendo uma face é passada para a próxima etapa na cascata, resultando que uma dada sub-janela tenha maior possibilidade de conter realmente uma face. A Figura4ilustra o conceito

[1] AdaBoost é uma máquina de aprendizagem com algoritmo de impulsionamento (*boosting*) capaz de construir um classificador forte por meio de uma combinação ponderada de classificadores fracos.

com duas fases.

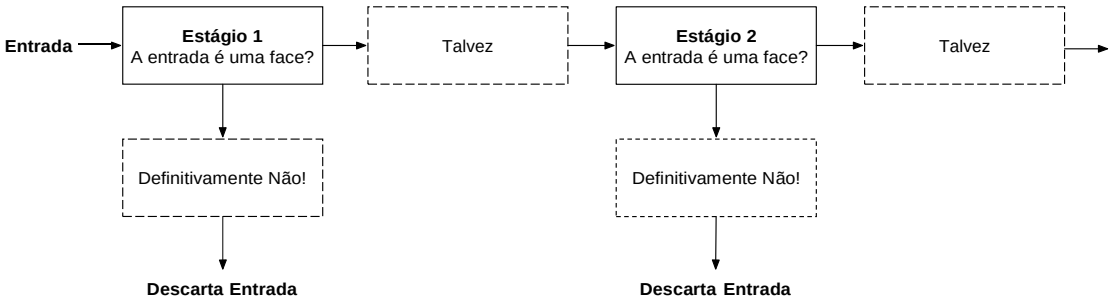


Figura 4 – Cascata de classificadores. Imagem original em (LOBBAN,2008)

Extração de Características

As próximas subseções descrevem técnicas de extração de características importantes para o desenvolvimento desta pesquisa.

Motion History Image - MHI

Esta seção está embasada no artigo deTian et al.(2012), o qual descreve o *Motion History Image - MHI*, como um modelo simples obtido com base nas diferenças entre as imagens. O MHI representa dentre uma sequência de imagens $S\{x_1, x_2, x_3, ..., x_n\}$, todo o movimento ocorrido nos objetos, formando um modelo do movimento de S.

O modelo de movimento em tempo real representa a temporalidade de camadas diferentes de imagens consecutivas em uma imagem estática. A intensidade do *pixel* é uma função da história de movimento nesse local, em que, valores mais brilhantes correspondem a um movimento mais recente. A informação de movimento direcional pode ser medida diretamente dos gradientes de intensidade no modelo MHI. Quando comparados ao fluxo óptico, os modelos MHI são mais eficientes nos cálculos do gradiente, assim como também mais robustos devido ao fato da informação sobre o movimento em MHI ser principalmente relacionada aos contornos dos objetos em movimento. Assim, em movimentos indesejados, as regiões do interior de contornos de objetos são ignoradas.

Para gerar um modelo MHI, usamos uma substituição simples e um operador de decaimento. Na posição (x, y) e tempo t, a intensidade do modelo $\mathrm{MHI}_\tau(x, y, t)$ é calculada desta forma:

$$\mathrm{MHI}_\tau(x, y, t)$$

$$\begin{cases} \tau, & se D(x, y, t) = 1 \\ max(0, MHI_\tau(x, y, t-1) - 1), & caso\ contrário \end{cases}$$

onde $D(x, y, t)$ é uma imagem binária de diferenças entre os quadros, e τ é a duração máxima de movimento. O modelo MHI é convertido para uma imagem em tons de cinza com a intensidade máxima de 255 *pixels*.

Histogram of Oriented Gradients - HOG

O artigo deDalal e Triggs(2005) foi usado como base para esta seção, o qual relata que o descritor de características *Histogram of Oriented Gradients - HOG* é baseado na avaliação dos histogramas locais normalizados das orientações do gradiente da imagem em uma grade densa.

A ideia principal do método é que a aparência e a forma do objeto local podem ser frequentemente bem definidas por meio da distribuição de gradientes de intensidade local, ou ainda, pela direção dos contornos, mesmo que não haja um conhecimento preciso das posições de inclinação ou pontos correspondentes. Em geral, o método é implementado através da divisão da imagem em janelas de pequenas regiões espaciais, conhecidas como células, cada célula acumula um local 1-D do histograma de gradientes orientados ou a orientação das bordas sobre os*pixels* das células. A combinação destes histogramas forma a representação. Ao todo as etapas envolvidas na construção do descritor são: **1)** detecção do espaço em escala; **2)** atribuição de orientação; e **3)** extração dodescritor.

> Como forma de melhorar o método com relação à invariância da iluminação, como o sombreamento, entre outros, é executada uma normalização de contraste, ou seja, acumula-se uma medida do histograma local em blocos de regiões espaciais e usa-se os resultados para normalizar todas as células do bloco. Estes blocos descritores normalizados são de fato o HOG.

O resultado obtido a partir do uso de extratores de características como HOG e MHI, é um conjunto formado normalmente por um número elevado de características. Entretanto, representar os dados por meio de vetores de características com elevadas dimensões pode prejudicar o processo de classificação. Com isso, torna-se necessário o uso de métodos de redução de dimensionalidade, conforme descrito na próxima seção.

Redução da Dimensionalidade

O objetivo principal na redução da dimensionalidade é a obtenção de um conjunto menor de características que represente com precisão o conjunto original, considerando que o novo conjunto deve captar o máximo de variância possível do antigo conjunto de dados (FORSYTH; PONCE,2002). Nesta pesquisa utilizamos *Principal Component Analysis - PCA* para realizar a tarefa de redução da dimensionalidade. Esta técnica será abordada na próxima subseção.

Principal Component Analysis - PCA

De acordo comForsyth e Ponce(2002), a ideia do PCA é utilizar um sistema de coordenadas especial que depende da massa de pontos, construída da seguinte maneira: deve-se colocar o primeiro eixo na direção de maior variância dos pontos para maximizar a variância no eixo, o segundo eixo deve ser perpendicular ao primeiro e assim por diante. Para o caso de duas dimensões, não há escolha, ou seja, a direção é determinada pelo primeiro eixo, entretanto, para três dimensões, a direção pode estar em qualquer lugar no plano perpendicular ao primeiro eixo. No caso de dimensões maiores, há diversas escolhas, embora sempre seja perpendicular ao primeiro eixo. Com esta restrição, a escolha do segundo eixo deve ser feita de forma que

maximize a variância ao longo do eixo, e assim por diante, ocorre a escolha de cada eixo para maximizar a variância restante.

Para obter o PCA é necessário calcular a matriz de covariância das coordenadas centrada na média obtida nos pontos e diagonalizá-la para encontrar os autovetores. Esses serão os eixos do espaço transformado, ordenados por autovalores, pois, cada autovalor retorna a variação ao longo do seu eixo.

Dados os pontos $x_1, x_2, \ldots, x_n \in \mathrm{R}^p$ define-se a reconstrução do dado em R^q para R^p como:

$$f(\lambda) = \mu + v_q \lambda.$$

Neste modelo de grau q, a média é $\mu \in \mathrm{R}^p$ e v_q é uma matriz $p \times q$, com q sendo vetores unitários ortogonais. Finalmente, $\lambda \in \mathrm{R}^q$ são os pontos de dados de menor dimensionalidade projetados (BELL; POP,2008).

A equação abaixo ilustra um exemplo hipotético do uso do PCA recebendo as características de entrada e as transformando em 4 componentes principais mais importantes de PCA no espaço de caraterísticas, reduzindo assim, as dimensões.

$$\begin{bmatrix} a^i & a^1_{i+1} & \ldots \\ a^1_i & a^1_{i+1} & \\ a^1_2 & a_2 & \ldots \\ \ldots & \ldots & \ldots \end{bmatrix} \rightarrow PCA \rightarrow \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix}.$$

Em nosso método, após a etapa de redução de dimensionalidade são empregados os metódos de codificação *Fisher Vector - FV* e *Vector of Locally Aggregated Descriptors - VLAD*, descritos na seção2.5, antes da interpretação, a qual é a última etapa da tarefa de reconhecimento de padrões, conforme descrito na seção2.1. Neste trabalho, a etapa de interpretação será realizada por meio de algoritmos de aprendizagem de máquina, mais precisamente, algoritmos de classificação, também conhecidos como classificadores. Apesar de haver uma ampla variedade de classificadores disponíveis na literatura, *Support Vector Machine* tem sido utilizado com muito sucesso em aplicações envolvendo detecção de faces e classificação de gestos, exemplificando podemos citar: (MAAOUI; ABDAT; PRUSKI,2014), (CHEN et al.,2013) e (HUSSAIN; MONKARESI; CALVO,2012). SVM, empregado neste trabalho, é descrito resumidamente na seção 2.6.

Métodos codificadores

Os métodos codificadores são baseados no conceito *Bag of Visual Words Model - BoVW*, originalmente proposto para a recuperação de documentos através de *Bag-of-Words - BoW*, e atualmente usado amplamente em visão computacional e adotado como principal modelo de representação de imagens de agrupamento de descritores locais. Em geral, um *framework BoVW* tradicional contém 03 etapas: **1)** extração de características locais; **2)** a produção de um

dicionário visual com um algoritmo de agrupamento, como por exemplo, *K-means* e *Gaussian Mixture Model - GMM* ; e **3)** processo de codificação do vetor. Nesta pesquisa, dois métodos de codificação são abordados: *Fisher Vector - FV* e *Vector of Locally Aggregated Descriptors - VLAD*, descritos nas subseções2.5.1e2.5.2, os quais têm obtido os melhores resultados dentre todos os codificadores existentes, evidenciados no completo estudo apresentado porPeng et al. (2014).

Fisher Vector - FV

Os conceitos descritos nesta seção sobre o método de codificação FV são baseados nos trabalhos dePerronnin e Dance(2007) ePerronnin, Sánchez e Mensink(2010), os quais relatam que FV é uma representação de imagem obtida por meio do agrupamento de características locais da imagem. É frequentemente usado como um descritor de imagem global na classificação visual. Nesta pesquisa, a codificação FV foi empregada para codificar os dados dos descritores da face e do gesto do corpo separadamente.

Uma especificidade desta codificação é o uso exclusivo do GMM na construção do dicionário visual. O GMM é um modelo probabilístico usado para representar subpopulações normalmente distribuídas dentro de uma população global. Os modelos de mistura em geral não requerem saber se a subpopulação pertence a um conjunto de dados específico, permitindo que o modelo aprenda as subpopulações automaticamente. Uma vez que a atribuição de subpopulação não é conhecida, esta constitui uma forma de aprendizagem não supervisionada. Busca demonstrar a partir do GMM uma amostra para uma primeiro subconjunto de índice $k \in \{1, \ldots, N\}$, com probabilidade prévia π_k e então amostras do vetor $x \in \mathrm{R}^d$ da k-ésima distribuição gaussiana $p(x/\mu_k, \bar{}_k)$, onde μ_k e $\bar{}_k$ são respectivamente a média e a covariância da distribuição. O GMM é completamente especificado pelos parâmetros $\Theta = (\mu_k, \Sigma_k, \pi_k : k = 1, \ldots, K)$. A densidade $p(x/\Theta)$ induzida nos dados de treino é obtida marginalizando a seleção do subconjunto k, obtendo:

$$p(x/\Theta) = \sum_{k=1}^{K} \pi_k p(x_i/\mu_k, \square_k),$$

$$p(x/\mu_k, \square_k) = \frac{1}{\sqrt{(2\pi)^d \det \bar{}_k}} \exp\left(-\frac{1}{2}(x - \mu_k)^T \square_k^{-1} (x - \mu_k)\right).$$

O processo para que o GMM se ajuste aos conjuntos de dados $X = (x_1, \ldots, x_n)$ é geralmente feito maximizando a função de log-verossimilhança dos dados:

$$A(\Theta; X) = E_{x \sim \hat{p}} [\log p(x/\Theta)] = \frac{1}{n} \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k p(x_i/\mu_k, \square_k)$$

onde $\hat{p}$ é uma distribuição empírica dos dados. Dados os conceitos matemáticos do GMM, podemos descrever FV da seguinte forma: seja $I = (x_1, \ldots, x_n)$ um conjunto de vetor de características de dimensão D, por exemplo descritores de HOG, extraídos de uma imagem; seja $\Theta = (\mu_k, \Sigma_k, \pi_k :$

$k = 1, ..., K$) os parâmetros do *Gaussian Mixture Model - GMM* ajustados e associados a cada vetor x_i para um modo k na mistura como uma força dada pela probabilidade a posteriori:

$$q_{ik} = \frac{exp\left[-\frac{1}{2}(x_i - \mu_k)^T \Sigma_k^{-1}(x_i - \mu_k)\right]}{\sum_{t=1}^{K} exp\left[-\frac{1}{2}(x_i - \mu_t)^T \Sigma_t^{-1}(x_i - \mu_t)\right]}.$$

Para cada modo k, considere os vetores de média e desvio de covariância.

$$u_{jk} = \frac{1}{N\sqrt{\pi_k}} \sum_{i=1}^{N} q_{ik} \frac{x_{ji} - \mu_{jk}}{\sigma_{jk}},$$

$$v_{jk} = \frac{1}{N\sqrt{2\pi_k}} \sum_{i=1}^{N} q_{ik} \left[\left(\frac{x_{ji} - \mu_{jk}}{\sigma_{jk}}\right)^2 - 1\right].$$

Onde $j = 1, 2, ..., D$ abrange as dimensões do vetor. A codificação FV da imagem I, demonstrada na equação abaixo, é o empilhamento dos vetores u_k e v_k para cada um dos modos de K nas misturas Gaussianas.

$$\Phi(I) = \begin{bmatrix} \vdots \\ u_k \\ \vdots \\ v_k \\ \vdots \end{bmatrix}.$$

O vetor de saída resultante da codificação tem um tamanho fixo de $2 \times k \times d$, onde k é o número de agrupamentos indicados na etapa anterior e d é o número de dimensão do vetor de características. A abordagem de codificação FV produz vetores que melhoram significativamente o desempenho, com relativamente poucos agrupamentos, devido ao fato de que os dados codificados são melhor separados espacialmente, assim, possibilitando o uso de simples classificadores, como separadores lineares, sem a necessidade de ajustes de parâmetros, ainda assim, gerando excelentes resultados em estudos que trabalham com o reconhecimento de ações humanas (BILIŃSKI,2014).

Vector of Locally Aggregated Descriptors - VLAD

O codificador VLAD é um método de codificação e agrupamento de características semelhante ao FV, pois, codifica um conjunto de descritores de características locais $I = (x_1, ..., x_n)$ extraídos de uma imagem usando um dicionário construído usando um método de agrupamento. No caso de VLAD, os dicionários podem ser gerados pelo algoritmo de agrupamento *K-means* ou ainda GMM. ConformeBiliński(2014) relata, VLAD acumula o residual de cada característica local com respeito à sua palavra visual atribuída. Então, combina cada característica local com a sua palavra visual mais próxima. Finalmente, para cada agrupamento, armazena a soma das diferenças dos descritores atribuídos ao agrupamento e do centroide do agrupamento. Nesta pesquisa, fizemos uso do *K-means* na construção do dicionário visual para a codificação VLAD.

Podemos definir os fundamentos de *K-means* da seguinte maneira: dado n pontos $x_1, ..., x_n \in \mathrm{R}^d$, o objetivo de *K-means* é encontrar K centros $c_1, ..., c_n \in \mathrm{R}^d$ e atribuições $q_1, ..., q_n \in \{1, ..., K\}$ dos pontos aos centros de modo que a soma das distâncias seja:

$$E(c_1, ..., c_k, q_1, ..., Q_n) = \sum_{i=1}^{n} \| x_i - c_{q_i} \|_p .$$

O *K-means* é amplamente utilizado em visão computacional, por exemplo, na construção de vocabulários de características visuais (palavras visuais). Nestas aplicações, o número n de pontos a agrupar e/ou o número K de aglomerados é muitas vezes elevado. Infelizmente, a minimização do objetivo E é, em geral, um problema combinatório difícil, pelo que são procuradas soluções localmente ótimas ou aproximadas.

Esclarecido o mecanismo do *K-means* utilizado na etapa de construção do dicionário visual, apresentamos a fundamentação matemática para a codificação VLAD conforme a pesquisa deJÉGOU et al.(2010). Seja q_{ik} a força da associação do vetor de dados x_i ao agrupamento μ_k, tal que $q_{ik} \geq 0$ e $\sum_{k=1}^{K} q_{ik} = 1$. A associação pode ser *suave*, por exemplo, obtida como as probabilidades posteriores dos agrupamentos de GMM, ou *dura* obtida por quantificação de vetor com *K-means*. O μ_i são os centros do agrupamento, vetores com a mesma dimensão que os dados x_i. VLAD codifica a característica x considerando os resíduos:

$$v_k = \sum_{i=1}^{N} q_{ik} (x_i - \mu_k).$$

Os resíduos são empilhados em conjunto para se obter o vetor codificado, conforme demonstra a equação abaixo:

$$\hat{\Phi}(I) = \begin{bmatrix} \vdots \\ v_k \\ \vdots \end{bmatrix}.$$

De forma geral, os vetores VLAD, representados pelo $\hat{\Phi}$ acima são normalizados antes do uso no classificador. O tamanho do vetor de saída codificado com VLAD é $k \times d$, onde k é o número de agrupamentos indicados na etapa anterior e d é o número de dimensão do vetor de características.

Support Vector Machine (SVM)

Introduzido em 1992 por Boser, Guyon, e Vapnik, o classificador SVM é considerado robusto para a classificação de dados com alta-dimensionaliade, pois, possui alta precisão e flexibilidade na modelagem de diversas fontes de dados e vem sendo amplamente utilizado em diversas áreas da computação (BEN-HUR; WESTON,2010). Os conceitos descritos nesta seção estão embasados no Livro *Data Mining Practical Machine Learning Tools and Techniques* de Witten, Frank e Hall(2011).

O classificador SVM usa modelos lineares para separar fronteiras de classes não lineares. Este artifício simples, funciona da seguinte forma: transforma-se a entrada usando um mapeamento não-linear, em outras palavras, o espaço de características original é transformado em um novo espaço. Devido ao mapeamento não-linear, o modelo linear construído no novo espaço pode representar uma fronteira de decisão não-linear no espaço original.

SVM tenta encontrar um tipo especial de modelo linear que maximize a margem de separação entre as classes. O hiperplano de separação máxima é a fronteira que maximiza a distância de separação entre as classes, como demonstrado na Figura5. Nessa figura, duas classes são ilustradas, as quais são representadas por círculos preenchidos e não preenchidos, respectivamente.

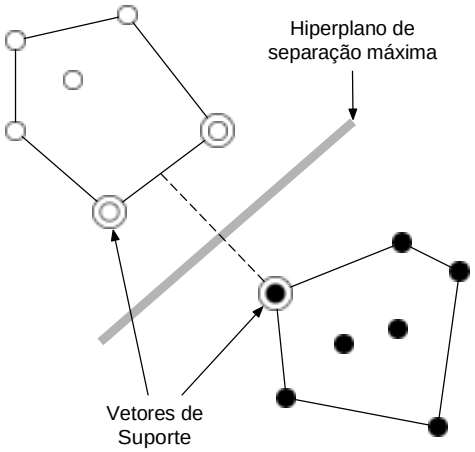


Figura 5 – Margem de separação máxima. Imagem orginal em (WITTEN; FRANK; HALL, 2011).

De acordo comWitten, Frank e Hall(2011), um hiperplano de separação de duas classes pode ser escrito: $x = w_0 + w_1a_1 + w_2a_2$ no caso de dois atributos, onde a_1 e a_2 são os valores de atributos e há três pesos w, a serem aprendidos. Entretanto, a equação que define o hiperplano de separação de margem máxima pode ser escrita de outra forma, em termos dos vetores de suporte. Escrevendo o valor de classe y de uma instância de treino como 1 (para sim) ou -1 (para não). O hiperplano de separação de margem máxima pode ser obtido como:

$$x = b + \sum \alpha_i y_i a(i) \bullet a \qquad (2.1)$$

onde, y_i é o valor da classe da instância de treinamento $a(i)$, enquanto b e α_i são parâmetros numéricos determinados pelo aprendizado do algoritmo. Note que $a(i)$ e a são vetores. O vetor a representa uma instância de teste. Os vetores $a(i)$ são vetores de suporte, assim como os circulados na Figura5.

Uma desvantagem dessa versão de SVM é o problema da complexidade computacional. Supondo que o espaço transformado é de alta dimensionalidade, tanto os vetores de suporte transformados quanto a instância de teste terão muitos componentes. De acordo com a equação 2.1, toda vez que uma instância é classificada, seu produto escalar com todos os vetores de suporte deve ser calculado. Esse procedimento se torna bastante custoso computacionalmente

em um espaço de alta dimensão produzido pelo mapeamento não-linear. Entretanto, é possível calcular o produto de pontos antes do mapeamento não linear ser executado, por meio do uso de funções *kernel*.

Este trabalho tem como objetivo classificar dados provenientes de duas fontes de informação: face e gestos. Com isso, as principais formas de fusão de dados multimodais são discutidas na próxima seção.

Fusão de dados

Embora na literatura a fusão de dados possa ser realizada por dois tipos, a saber: fusão em nível de características, também conhecida como recurso, e fusão em nível de decisão ou classificação, neste trabalho abordarmos somente a forma de fusão baseado na decisão, descrita com detalhe na subseção2.7.1.

Fusão em Nível de Decisão

A fusão em nível de decisão concentra esforços na utilização de conjuntos de classificadores e suas combinações (WAGNER et al.,2011). O principal objetivo deste método de fusão é aproveitar a redundância de um conjunto de classificadores independentes para alcançar maior robustez através da combinação de seus resultados (PLANET; IRIONDO,2012).

De acordo comSANTOS, Sabourin e Maupin(2008), este método baseia-se no pressuposto que classificadores independentes podem cometer erros diferentes. Então, combinar as decisões pode levar à uma melhora do desempenho global do sistema. Métodos como *bagging*, *boosting* e *randomization* são frequentemente usados para a geração dos membros do conjunto, enquanto que os métodos de votação por maioria, soma, produto, máximo e mínimo são funções usadas para combinar as decisões. Os resultados destes classificadores são tidos em conta para o processo de tomada de decisão final. O termo de fusão em nível de decisão resume uma variedade de métodos concebidos a fim de fundir as decisões dos membros do conjunto em uma única decisão.

Uma desvantagem dessa abordagem apontada porWitten, Frank e Hall(2011), é que torna-se difícil analisar os resultados obtidos, pois, um conjunto de classificadores pode compreender dezenas ou mesmo centenas de modelos individuais e, apesar de um bom desempenho ser alcançado, não é fácil entender em termos intuitivos quais fatores contribuem mais decididamente para as melhores decisões.

Na seção2.7.1.1a seguir descrevemos de forma resumidas as principais técnicas para combinar as decisões embasada no artigo deAlmeida, Cavalcanti e Ren(2014):

2.7.1.1 Combinadores de Decisão

Os graus de suporte dados por um padrão de entrada x podem ser interpretados de diferentes maneiras, as duas mais comuns são: o valor de confiança em uma sugestão de classificação e a estimativa das probabilidades a posteriori para as classes. Dado que o padrão $x \in R^n$ seja um vetor de características e $\Omega = \{\omega_1, \omega_2, \ldots, \omega_L\}$ seja o conjunto das classes do problema, cada classificador c_i no conjunto $C = \{c_1, \ldots, c_L\}$ tem L saídas de graus de suporte. Assumindo que

todas as L saídas são valores no intervalo $[0, 1]$, então $c_i : R^n \rightarrow [0, 1]^L$ trabalha no nível de medição. Portanto, denota-se $d_{i,l}(x)$ o suporte que o classificador ci dá à hipótese que x é da classe ω_l. Quanto maior o suporte, maior a chance de ser da classe ω_l.

Os combinadores usam a matriz de perfil de decisão DP(x) para encontrar o suporte geral de cada classe para a entrada x com o maior valor de suporte. Isso é feito através de $\mu_l(x) = F\,[d_{1,l}(x), \ldots, d_{M,l}(x)]$, o nível de suporte geral para ω_l, obtido após aplicar uma função de combinação F de expressões algébricas aos suportes individuais $d_{i,l}(x)$ da classe ω_l é dado pelo conjunto. Sendo assim, a decisão final pode ser obtida por $h_{final}(x) = argmax\ \mu_l(x)$. As funções de combinação F podem computar o suporte geral da classe de diferentes maneiras como visto a seguir:

Média: calcula a média dos suportes para cada classe, e a decisão final $h_{final}(x)$ é dada pela classe com maior média:

$$\mu_l(x) = \frac{1}{M} \sum_{i=1}^{M} d_{i,l}(x).$$

Soma: realiza a soma dos suportes para cada classe, e a decisão final $h_{final}(x)$ é dada pela classe com maior soma. Essa regra é equivalente à média:

$$\mu_l(x) = \sum_{i=1}^{M} d_{i,l}(x).$$

Produto: multiplica os suportes para cada classe, e atribui a decisão final $h_{final}(x)$ para a classe que tiver o maior produto:

$$\mu_l(x) = \prod_{i=1}^{M} d_{i,l}(x).$$

Máximo: encontra o suporte máximo para cada classe, e atribui a decisão final $h_{final}(x)$ para a classe com o maior suporte máximo:

$$\mu_l(x) = \max_{l} \{d_{i,l}(x)\}.$$

Mínimo: encontra o suporte mínimo para cada classe, e atribui a decisão final $h_{final}(x)$ para a classe com o menor suporte mínimo:

$$\mu_l(x) = \min_{l} \{d_{i,l}(x)\}.$$

Mediana: encontra a mediana dos suportes para cada classe, e atribui a decisão final $h_{final}(x)$ para a classe com o maior mediana dos suportes:

$$\mu_l(x) = \operatorname{med}_{l} \{d_{i,l}(x)\}.$$

Nesta seção apresentamos as regras de fusão baseada em decisão mais difundidas na literatura. A regra empregada em nosso método é apresentada com maiores detalhes na seção 4.6. Neste capítulo foram discutidos conceitos importantes que envolvem reconhecimento de padrões, extração de características, métodos de codificação, formas de fusão e classificação de dados oriundos das modalidades da face e gestos do corpo para o entendimento deste trabalho. O capítulo3apresenta pesquisas correlatas a este trabalho, resultado de uma revisão sistemática da literatura.

3 Trabalhos Correlatos

Este trabalho está relacionado ao campo de pesquisa de análise e reconhecimento de opinião com base em informações multimodais, apoiado pelas áreas de visão computacional e aprendizagem de máquina. Em geral, trabalhos que tratam desta problemática utilizam 3 modalidades de diferentes fontes de dados, são elas: informações de vídeo, áudio e texto contendo as transcrições das falas. Também utilizam técnicas de fusão de modalidades com objetivo de melhorar o desempenho de suas pesquisas.

A seguir, descrevemos resumidamente 5 principais pesquisas relacionadas com este trabalho. A organização dos trabalhos está disposta em ordem de relevância. A análise dos trabalhos e um quadro resumo comparativo são apresentados na seção3.1.

Uma das primeiras pesquisas a trabalhar com análise de opinião multimodal foi a abordagem desenvolvida porMorency, Mihalcea e Doshi(2011), a qual combina recursos audiovisuais e de texto para analisar e reconhecer opiniões positivas, negativas ou neutras. Uma importante contribuição dos autores foi a introdução de um novo conjunto de dados, denominado de *Youtube Dataset*, coletado a partir de vídeos do *Youtube*. A base tem ao todo 47 vídeos com pessoas que expressam suas opiniões no idioma inglês, a respeito de diversos tópicos, sendo 20 pessoas do sexo feminino e 27 do sexo masculino, com idades na faixa de 14 a 60 anos e de diferentes origens étnicas.

As principais características extraídas da fonte de texto foram palavras polarizadas, como por exemplo, bom, ruim, etc., as quais foram compiladas a partir de um dicionário de dados *Multi-Perspective Question Answering - MPQA* (MIHALCEA; BANEA; WIEBE,2007) que consiste em palavras carregadas com sentimento positivo ou negativo. Características visuais foram obtidas com a utilização de *software* comercial chamado Okao Vision[1], o qual retorna de forma automática a intensidade do sorriso e a direção do olhar usando ângulos horizontais e verticais registrados em graus. A partir da fonte de áudio, os autores extraíram de forma automática momentos com silêncio (pausas) e entonação da voz com uso do *software* público OpenEAR[2]. Os autores empregaram o *Hidden Markov Model - HMM* no processo de classificação com a opção de teste conhecida como *leave-one-out*.

Observando a métrica da Medida-F adotada no trabalho pelos autores, a pesquisa alcançou os resultados das modalidades isoladas de 0,439 com dados provenientes de vídeo, 0,430 com dados de texto e 0,419 com dados de áudio. Porém, o melhor resultado foi obtido com a fusão das modalidades com 0,553, portanto, superior aos resultados alcançados com uso das modalidades separadamente.

O método de análise de opinião multimodal proposto porZadeh et al.(2016) introduz uma importante contribuição com a elaboração da base de dados *Multimodal Opinion-level Sentiment Intensity - MOSI*. Composta por um total de 93 vídeos coletados do *Youtube*, os

[1] Disponível no endereço eletrônico: https://plus-sensing.omron.com/technology/
[2] Disponível no endereço eletrônico: https://sourceforge.net/projects/openart/

vídeos da base MOSI apresentam 89 pessoas diferentes, sendo 41 mulheres e 48 homens, com idades aproximadas entre 20 e 30 anos, de diferentes etnias e que expressam suas opiniões sobre variados temas no idioma inglês. Ao todo, a base contempla 7 tipos de intensidade da opinião, a saber: fortemente positiva, positiva, fracamente positiva, neutra, fracamente negativa, negativa e fortemente negativa.

Nesse estudo, os autores buscam compreender o padrão de interação entre as palavras (texto) e os gestos visuais. Assim como no trabalho mencionado anteriormente, três modalidades são empregadas: textos, vídeos e áudio, além da fusão entre elas. As características utilizadas no estudo foram a transcrição das falas, para os dados de texto, utilizando o conceito de *Bag-of-Words* simples, mais de 32 características de áudio, incluindo entonação, *Mel-Frequency Cepstral Coefficients - MFCCs* e *Normalized Amplitude Quotient - NAQ*, usando um repositório colaborativo de análise de voz para tecnologias de fala (DEGOTTEX et al.,2014), e por fim, o sorriso, olhar severo, aceno e sacudir de cabeça para os dados de vídeo (WOOD et al.,2015). Os experimentos foram classificados com uso de SVM e *Deep Neural Network - DNN*, com validação cruzada com 5 partes, obtendo as maiores acurácias tanto para texto (0,65) e para vídeo (0,61), quanto para áudio (0,57), foram obtidos por SVM. A fusão das modalidades foi realizada em nível de características, por meio de uma técnica simples de concatenação dos vetores, a qual obteve maior acurácia com 0,71, também com uso do SVM.

Assim como as pesquisas anteriores, o estudo deRosas, Mihalcea e Morency(2013) aborda o reconhecimento de análise de opinião combinando textos, áudio e vídeos. Uma grande contribuição desta pesquisa foi a criação da base de dados *Multimodal Opinion Utterances Dataset - MOUD*. Essa base é composta por um conjunto de 105 vídeos coletados a partir do *Youtube*, nos quais pessoas expressam suas opiniões majoritariamente a respeito de filmes, produtos cosméticos e livros. Os vídeos incluem 21 homens e 84 mulheres, com idades na faixa etária aproximada de 15 a 60 anos, oriundas de países que falam o idioma espanhol. A base é rotulada com 3 diferentes classes de opinião: positiva, negativa eneutra.

Características provenientes do áudio como duração da pausa, variação do tom, média do volume e a intensidade da voz foram extraídas automaticamente com uso do *software* OpenEAR. Informações relacionadas à modalidade de vídeo foram captadas com a utilização de *software* comercial chamado Okao Vision, que retorna automaticamente a intensidade de sorriso e a direção do olhar registradas por ângulos horizontais e verticais. Para informações de texto, os autores fizeram uso de uma abordagem de *Bag-of-Words* com as transcrições das falas para construir o vocabulário e extrair o vetor de características. A fusão das modalidades foi baseada em características e executada por concatenação simples dos vetores.

O algoritmo de classificação empregado na pesquisa foi SVM com *kernel* linear com validação cruzada com 10 partes. Novamente, a fusão das modalidades, com taxa de acurácia igual a 0,75, superou os resultados de acurácia alcançados com o uso das modalidades isoladamente: 0,649 para texto, 0,610 para vídeo e 0,467 para áudio. Além disso, os autores investigaram a portabilidade do método utilizando um segundo banco de dados contendo 37 opiniões sobre telefones celulares expressadas por pessoas em inglês, capturadas do site *Expotv.com*. A fusão das modalidades alcançou a maior acurácia com, 0,648, seguida de 0,540 para texto e vídeo

separadamente e 0,486 para a modalidade de áudio isolada.

O estudo dePoria et al.(2016) utilizou o mesmo conjunto de dados proposto por Morency, Mihalcea e Doshi(2011) na análise de opinião multimodal. Três classes de opiniões foram investigadas: positiva, negativa e neutra. A pesquisa extrai dos quadros de vídeo 66 pontos característicos da face, com uso do *software* comercial de reconhecimento facial Luxand FSDK 1.7[3] e calcula a distância média desses pontos para formar o vetor de característica final da modalidade do vídeo. Do mesmo modo, as características de áudio, como o tom e a intensidade da voz, também são extraídas de cada um dos segmentos de áudio automaticamente com uso do *software* público openEAR. As informações de textos são extraídas das transcrições das falas seguindo uma heurística própria baseada em conceitos de análise de sentimentos.

Dentre os algoritmos de aprendizagem de máquina testados pelos autores, estão: *Naive Bayes*, SVM, *Artificial Neural Network - ANN* e *Extreme Machine Learn - ELM*. Embora por uma pequena margem, o ELM com opção de teste de validação cruzada com 10 partes obteve o melhor resultado em relação aos demais. Os resultados das classificações separadas referentes às modalidades de texto, áudio e vídeo, observando a medida de precisão, foram: 0,619, 0,652 e 0,681 respectivamente. Os autores realizaram fusão das modalidades em nível de características, alcançando a maior precisão relatada com 0,782, e fusão baseada em decisão, que alcançou 0,753. Em geral, a pesquisa produziu uma acurácia de 68,60%.

A pesquisa deZadeh(2015), traz inicialmente, um estudo do comportamento verbal e não-verbal, investigando diferentes padrões de correspondência estrutural e de interação entre texto, áudio e vídeo, tendo como objetivo classificar a subjetividade e a intensidade da opinião. O autores fizeram uso do mesma base de dados elaborada porZadeh et al.(2016), permitindo investigar a intensidade da opinião, definidas na base como: fortemente positiva, positiva, fracamente positiva, neutra, fracamente negativa, negativa e fortemente negativa.

A partir de arquivos de textos contendo as transcrições das falas, os autores identificaram as polaridades sentimentais das palavras utilizando modelos de linguagem baseados em *n-gramas*, formando o vetor de característica para a modalidade. As características extraídas das *Facial Action Coding System - FACs* e pose da cabeça, e, 48 elementos da voz incluindo MFCCs e NAQ provenientes das fontes de vídeo e áudio respectivamente formaram os vetores finais de características.

O algoritmo *Support Vector Regression* linear foi empregado na geração do modelo de análise de opinião. A pesquisa definiu como métrica o *Mean Absolute Error - MAE*, alcançando 1,24 para dados visuais, 1,18 para dados verbais (texto e áudio). A fusão dos dados verbais e visuais alcançou o melhor resultado com 1,14. Portanto, de maneira recorrente, em todos os trabalhos descritos neste capítulo, o melhor resultado foi obtido com a fusão das modalidades, sendo esta baseada em características, por meio de uma técnica simples de concatenação de vetores. Essas análises mais gerais são apresentadas na próxima seção.

[3] Disponível no endereço eletrônico: https://www.luxand.com/index.php

Análise e Resumo Comparativo

Ao avaliar os trabalhos destacados neste capítulo, é possível observar que os métodos existentes normalmente combinam 3 modalidades distintas, áudio, texto e vídeo, tendo como objetivo obter resultados melhores se comparados com os resultados alcançados com as modalidades isoladas. Como consequência, são modelos complexos e de difícil reprodução. É importante notar também que a maioria dos estudos faz uso de algum *software* privado, com custo de aquisição elevado, principalmente na etapa de extração automática de características das modalidades, fato que também inviabiliza em alguns casos a reprodução desses métodos. Além disso, por trabalharem com arquivos de áudio para obter informações de intensidade, tom e entonação de voz, bem como com arquivos textuais contendo as transcrições das falas, essas soluções tornam-se restritas a um idioma, dificultando sua portabilidade para outros idiomas. Além disso, um grande inconveniente de utilizar a modalidade de texto é obrigar que os dados tenham as transcrições das falas, feitas em sua grande maioria manualmente, sincronizadas com as demais fontes, dificultando a aquisição das informações completas para o modelo.

Apesar das bases de dados investigadas nesses trabalhos serem rotuladas e estarem transcritas, essa situação normalmente não ocorre em aplicações reais. Entretanto, essas bases são formadas por vídeos do *mundo real*, os quais trazem uma gama de desafios, como por exemplo, ruídos, diferentes escalas e plano de fundos variados, que dificultam a solução do problema. Em geral, as pesquisas apresentam melhores resultados com a fusão das modalidades, mesmo assim, abaixo de 80% de taxa de acerto. Portanto, trata-se de um problema difícil, para o qual não são atingidas altíssimas taxas de acurácia. A Tabela1apresenta os estudos supracitados de forma resumida, em função das características extraídas, base de dados e algoritmo de classificação usados na predição da opinião expressa.

Tabela 1 – Comparativo dos estudos sobre análise de opinião multimodal.

ID	Banco de Dados	Caraterísticas Extraídas Modalidades			Algoritmos de Classiftcação
		Vídeo	Áudio	Texto	
1	*Youtube Dataset.*	Intensidade do sorriso; direção do olhar.	Momentos com silêncio (pausas) e entonação da voz.	Palavras polarizadas.	HMM.
2	*Multimodal Opinion-level Sentiment Intensity - MOSI.*	Sorriso, olhar severo, aceno e sacudir de cabeça.	+32 características, incluindo entonação, MFCCs e NAQ.	BoW.	SVM e DNN.
3	*Multimodal Opinion Utterances Dataset - MOUD.*	Intensidade do sorriso e direção do olhar.	Duração da pausa, variação do tom, média do volume e intensidade da voz	BoW.	SVM.
4	*Multimodal Opinion-level Sentiment Intensity - MOSI.*	FACs e pose da cabeça.	48 elementos da voz, incluindo MFCCs e NAQ.	BoW.	SVR.
5	*Youtube Dataset.*	Distâncias de 66 pontos da face.	Tom e a intensidade da voz.	BoW.	*Naive Bayes*, SVM, ANN e ELM.

De maneira diferente, nosso trabalho trata com somente duas modalidades: expressão da face e gesto do corpo, provenientes de uma única fonte de dados, com objetivo de simplificar o

Tabela 2 – Identificação do estudos correlatos.

ID	Estudo	Baseline
1	(MORENCY; MIHALCEA; DOSHI,2011)	✓
2	(ZADEH et al.,2016)	✓
3	(ROSAS; MIHALCEA; MORENCY,2013)	
4	(ZADEH,2015)	✓
5	(PORIA et al.,2016)	

modelo. Outra vantagem que o uso de apenas uma fonte de dados proporciona é a facilidade da portabilidade do método independentemente do idioma falado pela pessoa que expressa a opinião no vídeo, conforme demonstrado na seção5.3.4. Além disso, adotamos o uso de métodos de codificação, que, de acordo com a nosso revisão bibliográfica, não foram aplicados no contexto de fusão de dados multimodais com foco em análise de opinião, no entanto, são amplamente utilizados no reconhecimento de atividades humanas (PENG et al.,2014). Os experimentos realizados seguindo nosso método são descritos na seção5.3e apresentam resultados superiores aos *baselines* citados na Tabela2. Antes, porém, o método proposto é descrito no próximo capítulo.

4 Método Proposto

Este capítulo descreve a estrutura e o funcionamento do método desenvolvido neste trabalho. A Figura6exibe um diagrama com a arquitetura do método dividido em fases e ilustra o funcionamento de cada módulo que o compõe e como eles se inter-relacionam. Os componentes descritos nesse diagrama são detalhados a seguir.

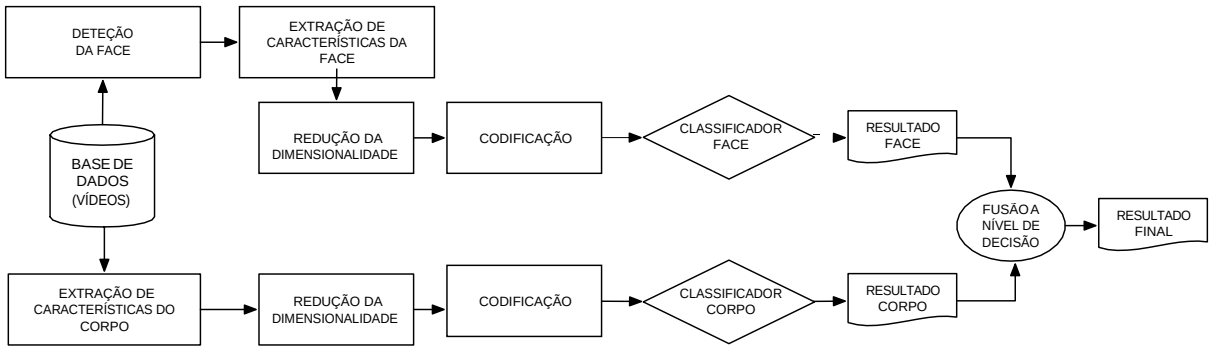


Figura 6 – Arquitetura geral empregada no reconhecimento de opiniões multimodal. PCA é empregado para redução de dimensão; e codificadores são empregados nos descritores da face e do corpo. Um método de classificação é então treinado para aprender a classificar opiniões utilizando modalidades de face e corpo separadamente. Por fim, uma estratégia de fusão é empregada para combinar a saída dos dois classificadores individuais para, dessa forma, produzir uma única classificação para o dado de entrada.

Conforme pode ser observado na Figura6, o processo ilustrado nessa arquitetura é tradicional em problemas de reconhecimento de padrões. Inicialmente, os vídeos de opinião são considerados como entradas para o processo, dos quais derivam as modalidades de face e gestos do corpo. Como são utilizadas informações referentes à faces e gestos do corpo de cada indivíduo, os vídeos são submetidos aos processos de detecção de faces, bem como, ao processamento de extração de características. Neste trabalho, o módulo de extração de características da face é realizado por meio do uso do descritor HOG. Da mesma forma, os vídeos são submetidos à etapa de extração de características de gesto de corpo. Neste trabalho, esse módulo é feito com a combinação de dois descritores: MHI e HOG. Inicialmente nós sintetizamos todo o movimento do vídeo em um único quadro com o uso do MHI, em seguida a síntese do movimento é descrita com o HOG.

Na sequência, a quantidade de características obtidas, tanto de faces quanto de gestos do corpo é reduzida pelo processo de PCA. Posteriormente, o método aplica a codificação dos descritores. Conforme mencionado no capítulo2, nós testamos dois codificadores: VLAD e FV. Para tanto, há a necessidade de criar dicionários visuais utilizando algoritmos de agrupamento, *Kmeans* e GMM foram utilizados. O propósito de utilizar os codificadores é criar uma separação espacial mais bem definida para os vetores de características, consequentemente, facilitando a tarefa de classificação. Em seguida, as informações codificadas são classificadas com o emprego de um algoritmo de aprendizagem de máquina. Dentre as muitas possibilidades disponíveis na

literatura, nós utilizamos SVM sem ajustes de parâmetros, ou seja, em sua versão com *kernel* linear. Por fim, os resultados da classificação das modalidades face e corpo são combinados com técnica de fusão, objetivando uma melhor predição da opinião expressa no vídeo. Na etapa da fusão das modalidades deste trabalho é usada a técnica de fusão baseada em decisão, com a regra de máxima probabilidade estimada. Cada etapa da arquitetura é descrita em maiores detalhes, a seguir.

Entrada de Dados

Nós acreditamos que o método proposto possa ser robusto para classificar opiniões expressas em vídeos do *mundo real*, com qualidades e tamanhos variados, gravados com diversos tipos de equipamentos e com planos de fundo e iluminações diferenciadas. Entretanto, é importante destacar que as bases de vídeos investigadas neste trabalho têm como principal características a existência de somente uma pessoa expressando sua opinião sobre um determinado assunto qualquer, invariavelmente posicionada de frente para a câmera. Dessa forma, é possível ver no mínimo o rosto da pessoa por completo.

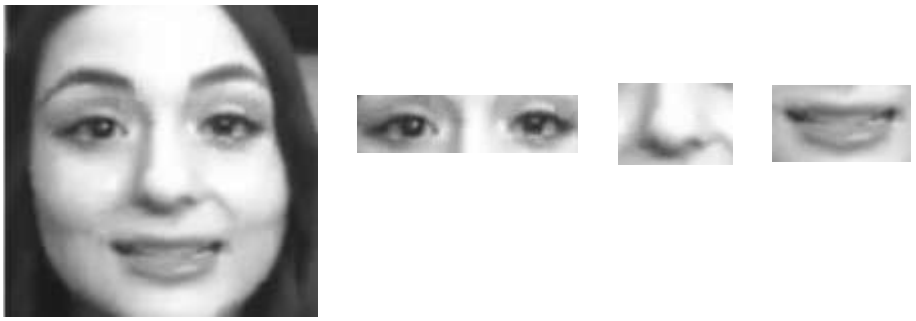
Modalidade Face

A literatura mostra em diversos estudos que para utilizar adequadamente a modalidade da face, as seguintes etapas principais precisam ser realizadas: detecção de faces e extração de características (GUNES et al.,2013). A primeira delas é referente à detecção da face e de seus componentes, como olhos, boca e nariz, os quais devem ser detectados e segmentados dos quadros dos vídeos. Na etapa seguinte, as regiões segmentadas na fase anterior serão submetidas à extração de características, objetivando retirar informações discriminantes para classificação desta modalidade. Esses dois processos são descritos a seguir.

Detecção de Face

Por sua eficiência e velocidade na detecção da face, o algoritmo *Viola-Jones - VJ* foi escolhido para esta tarefa, descrito com maiores detalhes na seção2.2. Inicialmente, os quadros do vídeo são capturados e convertidos do sistema de cor RGB para escala de cinza, então, são repassados como parâmetro de entrada para o algoritmo que rastreia a face e retorna as informações dos limites referentes ao rosto, conforme ilustrado na Figura7(a). Os limites da face permitem recortar do quadro somente a área de interesse. Novamente utilizamos *Viola Jones*, desta vez, com parâmetros específicos para retornar os limites da boca, nariz e olhos, tendo a imagem da face rastreada anteriormente como entrada para o algoritmo. A Figura7ilustra um exemplo do uso do algoritmo na detecção de face, olhos, nariz e boca de um quadro do vídeo.

Como podemos observar, o resultado deste processo são blocos de imagens referentes às regiões dos olhos, nariz e boca, conforme pode ser visto na Figura7(b) (c) (d). Devido ao fato dos vídeos não serem controlados, podem ocorrer variados ângulos das faces das pessoas que expressam sua opiniões de forma espontânea no quadro. Nesses casos, é possível que o algoritmo de detecção recorte os blocos com tamanhos ligeiramente diferentes. Como medida



(a) (b) (c) (d)

Figura 7 – Ilustração da detecção da face e seus componentes obtidos pelo algoritmo *Viola Jones*. (a) detecção da imagem da face; (b) detecção dos dos olhos, (c) detecção do nariz; e (d) detecção da boca.

para resolver este problema, os blocos sofreram uma padronização nos tamanhos após a detecção, assim definida: olhos (25×95 *pixels*); nariz (35×42 *pixels*); e boca (34×55 *pixels*).

Os casos de oclusões da face ou seus componentes, que eventualmente podem acontecer durante a execução do vídeo, são tratados com a adição de um quadro vazio com o mesmo tamanho do bloco. Em seguida, os blocos são normalizados pela média e desvio padrão com objetivo de mitigar os ruídos. Por fim, uma imagem média normalizada destes blocos é obtida, a qual é utilizada como entrada para a fase de extração das características.

Figura 8 – Ilustração do resultado obtido pelo método *Viola Jones* na forma de uma face média de um vídeo.

Como forma de exemplificação, na Figura8é possível visualizar uma face média normalizada, a qual está pronta para ser submetida ao processo de extração de características, detalhado na próxima subseção.

Extração e Características

Nesta etapa, os componentes trabalhados na fase anterior são submetidos como entrada para o descritor de características HOG, o qual é baseado na avaliação dos histogramas locais normalizados das orientações do gradiente da imagem em uma grade densa, conforme descrito na subseção2.3.2. Levando em consideração o pequeno tamanho dos blocos (olhos, boca e nariz), nós definimos uma janela com tamanho de 4×4 *pixels* como parâmetro para o descritor. Janelas

de tamanhos menores obtêm mais informações, em contrapartida, geram vetores com tamanhos maiores. A Figura9ilustra um exemplo do descritor HOG sobre o bloco referente a olhos.

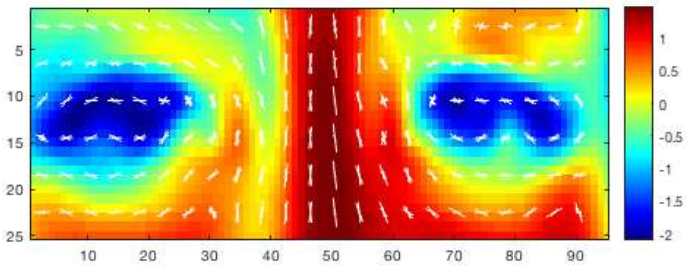


Figura 9 – Exemplificação do descritor HOG representado por asteriscos branco demonstrando as orientações do gradiente sobre o bloco de olhos aplicado com a paleta de cores *jet*.

Os eixos x e y da Figura9representam o tamanho do bloco e a escala representa os valores do gradiente. Por fim, temos como resultado desta etapa uma gama de características extraídas dos blocos que são concatenadas em um único e longo vetor, o qual será processado na próxima etapa dométodo.

Modalidade Gestos do Corpo

Os vídeos são submetidos ao processo de extração de características, entretanto, na modalidade gestos do corpo, não segmentamos área específica no quadro, simplesmente consideramos o quadro inteiro como informação de entrada para o processo. Como os vídeos possuem tamanhos variados, optamos pela padronização de todos os quadros no tamanho de 360×480 *pixels*. Neste trabalho, a extração de características do corpo se deu pela combinação de 2 descritores: MHI e HOG. Enquanto o MHI faz uma representação de uma sequência de movimento do corpo no vídeo, o descritor HOG extrai do MHI as orientações da distribuição do gradiente. Estas etapas são detalhadas nas próximas subseções.

Motion History Image - MHI

SegundoTian et al.(2012), a intensidade do *pixel* em uma imagem MHI representa a história de movimento naquele ponto, ou seja, valores mais brilhantes correspondem a um maior e mais recente movimento. O MHI foi empregado para representar o movimento do corpo de uma sequência de vídeo. O processo de geração da imagem MHI é bem simples. A sequência do vídeo do corpo é convertida do sistema de cor RGB para escala de cinza. Em seguida, são capturados todos os n quadros calculando a imagem diferença de todos os quadros, resultando em um conjunto $S = \{1, 2, 3, \ldots, n-1\}$de imagens diferença. Neste trabalho, foi definido o valor de 5 para o limiar (*threshold*) de intensidade, o qual controla a intensidade dos registros do movimento. Por fim, uma soma ponderada, considerando a escala de intensidade de 255, dos resultados das imagens filtradas é realizada, obtendo-se assim uma única imagem. A Figura10 exibe um exemplo da geração da representação do movimento de uma opinião capturada de uma sequência de vídeo.

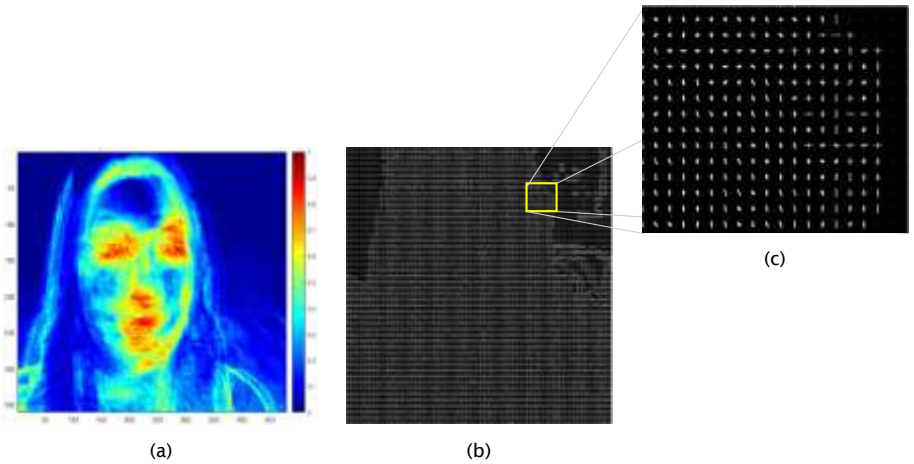
(a) (b)

Figura 10 – Demonstração da extração do MHI do Corpo. (a) Sequência de quadros com a expressão de uma opinião positiva a respeito de um livro; e (b) Representação do MHI com paleta de cores *jet* aplicada. Os valores do gradiente mais claros correspondem a um maior movimento do corpo, valores mais escuros representam menor movimento corporal.

O resultado desse processamento traz uma imagem gradiente que sintetiza o movimento de uma opinião expressa no espaço e no tempo, a qual será fornecida como entrada de dados para o próximo descritor.

Histogram of Oriented Gradients - HOG

ConformeDalal e Triggs(2005), a ideia principal do HOG é que a aparência e a forma do objeto local podem fornecer direção dos contornos utilizando a informação do gradiente. O descritor recebe uma imagem de entrada e extrai as características referentes ao histograma de gradientes orientados, podendo retornar um vetor ou matriz. De forma a complementar o processo de extração de característica do corpo com maior informação, esta pesquisa emprega o descritor HOG na representação do movimento gerado na etapa anterior. O exemplo da Figura 11demonstra a extração das características retiradas a partir do MHI.



(a) (b) (c)

Figura 11 – Demonstração da extração do HOG a partir do MHI. (a) quadro de representação do MHI com paleta de cores *jet* aplicada; (b) Representação visual do HOG extraído do MHI; e (c) Detalhe do HOG demonstrando as orientações do gradiente.

Da mesma maneira que na modalidade anterior, a janela com tamanho de 4×4 *pixels* foi escolhida para execução do HOG. Seguindo a arquitetura, a próxima tarefa para predição de uma

opinião é a redução de dimensionalidade, descrita na seção4.4. Em seguida as codificações FV e VLAD são empregadas, detalhadas na seção4.5. Por fim, deverá ser realizada a classificação e a fusão dos resultados. Essa última etapa é apresentada na Seção4.6.

Redução da Dimensionalidade

De acordo comForsyth e Ponce(2002) o PCA cria um sistema de coordenadas próprio de dimensão menor, o qual representa o sistema de coordenadas original. O processo de redução da dimensionalidade tem a finalidade principal de diminuir os vetores de características, garantindo a representatividade das informações extraídas, este processo está detalhado na seção2.4. Devido ao fato dos vetores de características extraídos nas etapas anteriores possuírem elevada dimensão, optou-se pela aplicação do PCA nas modalidades de face e gesto nesta pesquisa. O estudo definiu o limiar de 95% de componentes principais como garantia de manter as representatividades das informações extraídas.

Como resultado dessa etapa, obtivemos vetores de características com dimensão menor e com um alto grau de representatividade, os quais são as informações de entrada a serem entregue para a próxima etapa do método.

Codificadores

Os codificadores FV e VLAD, abordados com maiores detalhes nas subseções2.5.1e 2.5.2, são métodos baseados no conceito *Bag of Visual Words Model - BoVW*, um conceito originalmente proposto para a recuperação de documentos *Bag-of-Words - BoW*, que atualmente vem sendo usado amplamente em visão computacional e adotado como principal modelo de representação de imagens agrupando descritores locais. Em geral, um *framework BoVW* tradicional contém a extração local de características, a produção de um dicionário visual com um algoritmo de agrupamento, como por exemplo, *K-means* e GMM, além da codificação das características. Especificamente nesta etapa de codificações, nós fizemos uso do *Toolbox Matlab VLFeat open source*, uma biblioteca de código aberto que implementa algoritmos populares de visão computacional especializados em compreensão de imagem, extração de características locais e de correspondência, disponível no endereço eletrônico *http://www.vlfeat.org* e desenvolvida por Vedaldi e Fulkerson(2008).

Fisher Vector - FV

SegundoBiliński(2014), a codificação FV não representa características como uma combinação de palavras visuais, mas representa diferenças entre características e palavras visuais. Antes de realizar a codificação é necessário criar um vocabulário visual agrupando recursos locais extraídos dos vídeos de treinamento.

Uma particularidade da codificação FV é que o algoritmo de agrupamento usado para gerar o dicionário de palavras visuais deve ser exclusivamente o GMM. O objetivo do GMM é identificar a presença de subpopulações contidas em um conjunto total de dados de forma não supervisionada. O GMM é dependente de um único parâmetro: *k* - número de agrupamentos.

O parâmetro k é sensível aos dados, portanto, realizamos alguns experimentos para definir esse valor, conforme será descrito na seção5.2. Uma vantagem do FV em relação aos demais codificadores é que o dicionário gerado pelo GMM captura tanto estatísticas de primeira ordem quanto de segunda ordem. Fornecendo ao GMM os vetores conseguidos na etapa de redução de dimensionalidade e o número de agrupamentos, temos como retorno informações valiosas como os centróides dos agrupamentos e a matriz de co-variança. Em seguida, o FV utiliza essas informações para codificar a diferença entre os descritores agrupados e o vocabulário, aplicando operações derivadas sobre a probabilidade com respeito aos parâmetros de distribuição do vocabulário (BILIŃSKI,2014).

A abordagem de codificação FV produz um desempenho elevado com relativamente poucas palavras visuais, mesmo que classificadores lineares simples sejam usados. Uma desvantagem é que o vetor de saída resultante da codificação terá um tamanho fixo, precisamente de $2 \times k \times d$, onde k é o número de agrupamentos indicados na etapa anterior e d é o número de dimensões do vetor de características.

Vector of Locally Aggregated Descriptors - VLAD

O codificador VLAD é uma versão do FV, entretanto, este método só mantém estatísticas de primeira ordem (PENG et al.,2014). ConformeBiliński(2014) relata, VLAD acumula o residual de cada característica local com respeito à sua palavra visual atribuída. Então, ele combina cada característica local com a sua palavra visual mais próxima. Finalmente, para cada agrupamento, armazena a soma das diferenças dos descritores atribuídos ao agrupamento e do centroide do agrupamento. No caso de VLAD, os dicionários são gerados pelo algoritmo de agrupamento *K-means*. O tamanho do vetor de saída codificado com VLAD é $k \times d$, onde k é o número de agrupamentos indicados na etapa anterior e d é o número de dimensões do vetor de características.

Classificação e Fusão

Devido ao ganho na distribuição espacial que os codificadores produzem, não é necessário empregar complexos classificadores para obter bons resultados, tornando o processo de aprendizado mais simples. Por exemplo, em nossos experimentos nós utilizamos o classificador SVM com o *kernel* linear sem a necessidade de alteração de parâmetros adicionais. Os vetores de características codificados com VLAD e FV são fornecidos como entrada para o processo de treino e teste do classificador. Modelos diferentes são criados para as modalidades da face e gesto do corpo gerando resultados isolados. Conforme mencionado anteriormente, a fusão das modalidades pode ser feita em nível de características ou em nível de decisão. Porém, como no nosso método dois classificadores são treinados e, consequentemente, dois modelos são gerados, é necessário que os dois modelos sejam combinados em nível de decisão, para que apenas uma classe seja atribuída ao dado de entrada. Dentre as várias formas de fusão de classificadores, uma estratégia de fusão baseada na probabilidade a posteriori do classificador é utilizada neste trabalho, conforme o resultado de saída exemplificado na Tabela3. Todos os experimentos com classificadores foram

executados com o *software* Waikato Environment for Knowledge Analysis - WEKA (versão 3.9.0), disponível gratuitamente no endereço eletrônico *http://www.cs.waikato.ac.nz*.

Tabela 3 – Exemplo de resultado gerado por SVM para a modalidade da face.

inst	atual	predita	erro	prob
1	1	-1	1	0,388
2	1	1	0	0,804
3	0	0	0	0,827
4	-1	-1	0	0,988
:	:	:	:	:
n	...	...	...	...

Dentre todas as informações exibidas nas Tabela3, a que possui maior importância para a etapa de fusão é a probabilidade estimada de uma instância *x* ser da classe predita pelo classificador. Na primeira coluna da Tabela3temos os valores de identificação das instâncias, a coluna *atual* refere-se à classe correta da instância, em *predita* temos a classe predita pelo classificador, a coluna *erro* demonstra se o classificador acertou (0) ou errou (1) a predição, e na última coluna temos a probabilidade estimada.

A partir dos valores de probabilidade gerados pelos modelos das modalidades de face e gesto do corpo, empregamos a técnica de fusão destes resultados baseada em decisão, seguindo uma regra simples: **1)** - quando os dois classificadores predizem a mesma classe para a instância de entrada, tal classe é atribuída à instância; **2)** - quando os dois classificadores divergem, o classificador com maior probabilidade estimada é escolhido para atribuir a classe à instância de entrada. Dessa forma, será escolhido o classificador com maior confiança para tomar a decisão.

Como forma de exemplificar a regra adotada, ilustramos na Tabela4um pequeno exemplo da fusão das modalidade de face e corpo utilizando a regra exposta acima.

Tabela 4 – Ilustração da regra de fusão das modalidades da face e corpo.

Modalidade Face			**Modalidade Corpo**			**Fusão Face/Corpo**		
inst	**erro**	**prob**	**inst**	**erro**	**prob**	**inst**	**erro**	**prob**
1	1	0,665	**1**	1	0,876	**1**	1	0,876
2	1	0,345	**2**	0	0,982	**2**	0	0,982
3	0	0,893	**3**	0	0,879	**3**	0	0,893
4	0	0,544	**4**	1	0,675	**4**	1	0,675
:	:	:	**:**	:	:	**:**	:	:
n	...	...	***n***	...	...	***n***	...	...

Observando a Tabela4, notamos que o resultado final da fusão das modalidades de face e corpo nas instâncias 1, 2, e 4 foi decidido através do resultado obtido pelo classificador do corpo e a instância 3 pelo classificador da face.

Neste capítulo descrevemos e discutimos o método com detalhes para deixar clara a implementação da pesquisa. Diversos experimentos foram realizados e os resultados são descritos e discutidos no próximo capítulo.

5 Experimentos e Resultados

Neste capítulo são apresentados os resultados obtidos a partir dos experimentos realizados com o método descrito no capítulo4. Inicialmente, porém, são descritas as três bases de dados utilizadas nos experimentos. Em seguida, são apresentadas as definições de parâmetros necessários para a geração dos dicionários utilizados nas codificações FV e VLAD das modalidade de corpo e face. Por fim, são descritos na seção5.3os resultados obtidos com o classificador SVM linear, além de uma análise dos resultados apresentada na subseção5.3.4.

Bases de Dados

O método foi testado com três bases de dados, a saber: *Youtube Dataset*, desenvolvido porMorency, Mihalcea e Doshi(2011); *Multimodal Opinion-level Sentiment Intensity - MOSI* criado porZadeh et al.(2016); e *Multimodal Opinion Utterances Dataset - MOUD*, elaborado porRosas, Mihalcea e Morency(2013). Embora não sejam bases de dados controladas, todas as bases são rotuladas e contêm arquivos de vídeos, áudio e texto com as transcrições das falas. Este último obtido de forma manual.

Youtube Dataset

Composta por 47 vídeos adquiridos diretamente do site do *Youtube*. Trata de diversos tópicos, desde de comentários com opiniões sobre produtos, religião ou até mesmo sobre posicionamento político. Ao todo, são 20 pessoas do sexo feminino e 27 do sexo masculino, com idades na faixa de 14 a 60 anos, com diferentes origens étnicas, como por exemplo, Europeus, Americanos, Hispânico e Asiáticos, expressando suas opiniões no idioma exclusivamente inglês. A base de dados pode ser solicitada a partir do endereço eletrônico *http://projects.ict.usc.edu/youtube/*. ConformeMorency, Mihalcea e Doshi(2011) observam em sua pesquisa, por se tratar de vídeos do *mundo real*, essa base representa um problema desafiador para a análise de opinião, pois, engloba diferentes desafios a superar:

- Diversidade. Pessoas podem expressar sentimentos de forma variada, algumas pessoas são mais sutis do que outras. Os vídeos contêm pessoas de diferentes nacionalidades, gêneros e faixas etárias expressando opiniões sobre diversos temas.
- Multimodalidade. O conjunto de dados contêm vídeos multimodais nos quais as pessoas misturam expressões faciais, posturas corporais, entonação de voz e escolhas de palavras para expressar suas opiniões e/ou afirmar fatos diretamente para a câmera.
- Ambientes ruidosos. A base de dados traz uma variedade de ruídos presente na maioria das gravações do *mundo real*, pois, são de vídeos gravados por pessoas comuns em suas casas, escritórios ou ao ar livre, utilizando diferentes equipamentos.

Os vídeos estão no formato *Moving Picture Experts Group - MPEG-4* com um tamanho padrão de 360 × 480 *pixels*. O tempo de duração dos vídeos varia de 2-5 minutos. A Figura12 demonstra seis exemplos extraídos de quadros dos vídeos que compõem a base de dados.

Figura 12 – Exemplos de vídeos da base de dados *Youtube Dataset*.

Todos os 47 vídeos são rotulados por três pessoas em três classes: positiva (1), negativa (-1) e neutra (0). A tarefa de anotação foi associar um rótulo de sentimento que melhor resume a opinião expressa no vídeo. Apesar do pouco número de instâncias, no geral, os autores equilibraram de maneira considerada satisfatória a distribuição entre as classes: 32% positiva, 25% negativa e 43% neutra.

MOSI Dataset

Assim como a base de dados descrita anteriormente, a base de dados *Multimodal Opinion-level Sentiment Intensity - MOSI* é composta por vídeos que foram coletados do site *Youtube* com foco em *vídeo-blogs* populares usados por muitas pessoas para expressar opiniões sobre diferentes assuntos. Rotulada conforme a intensidade da opinião, apresenta um total de 7 classes definidas: fortemente positiva (+3), positiva (+2), fracamente positiva (+1), neutra (0), fracamente negativa (-1), negativa (-2) e fortemente negativa (-3). Entretanto, neste trabalho de pesquisa, nós centramos esforços na investigação somente de vídeos rotulados com as 3 classes básicas de opinião: positiva, negativa e neutra.

SegundoZadeh et al.(2016), uma grande vantagem deste tipo de vídeo é que eles geralmente contêm apenas uma pessoa, olhando diretamente para a câmera. Em contrapartida, alguns desafios se mostram na coleção, pois, os vídeos são gravados em diversas configurações, algumas pessoas utilizaram equipamentos com qualidade profissional, enquanto outras, menos profissionais, gravaram os vídeos com seus próprios dispositivos. Além disso, as pessoas presentes no vídeo são gravadas em diferentes distâncias, com plano de fundo e condições de iluminação variados entre os vídeos. O formato MPEG-4 foi mantido do site original, assim, como seus tamanhos originais foram preservados, resultando em vídeos com diferentes tamanhos. Os tempos de duração dos vídeos completos variam de 2-5 minutos. A Figura13ilustra seis quadros de exemplos capturados de vídeos que compõem a base de dados MOSI.

Um conjunto total de 89 diferentes pessoas foi selecionado nos vídeos, sendo, 41 mulheres e 48 homens com idades aproximadas entre 20 e 30 anos de diferentes etnias, por exemplo, Cau-

Figura 13 – Exemplos de vídeos da base de dados MOSI.

casianos, Afro-americano, Hispânico e Asiáticos, todos expressando suas opiniões exclusivamente em inglês. Embora a base tenha um total de 89 vídeos, os autores fragmentaram os vídeos em segmentos menores em conformidade com a intensidade da opinião, resultando num total de 1.298 vídeos utilizados no processamento dos experimentos.

Da mesma forma que na base de dados descrita na subseção5.1.1, a base de dados MOSI mantém uma distribuição de classes equilibrada, sendo: 39% da classe positiva, 35% negativa e 26% da classe neutra, portanto, considerada adequada para geração do modelo. A base de dados MOSI pode ser solicitada pelo endereço eletrônico *https://goo.gl/forms/vFfFCdP2Jua8Wwtm2*.

5.1.3 MOUD Dataset

A base *Multimodal Opinion Utterances Dataset - MOUD* é formada por um conjunto de 105 vídeos, nos quais pessoas expressam suas opiniões em sua maioria a respeito de filmes, produtos cosméticos e livros no idioma exclusivamente espanhol, coletados a partir do *Youtube*. O conjunto final dos vídeos inclui 21 homens e 84 mulheres selecionados aleatoriamente, com idades na faixa etária aproximada de 15 a 60 anos. Todas as pessoas presentes nos vídeos são de origem de países que falam espanhol, como por exemplo, México, Espanha ou países da América do Sul.

Os vídeos foram selecionados pelos autores com alguns critérios, a saber: **1)** a pessoa deve aparecer no vídeo de frente para a câmera; **2)** o rosto deve ser visível; e **3)** não existir qualquer música de fundo ou animação. A Figura14exibe exemplos de quadros capturados de seis vídeos pertencentes à base. A base encontra-se disponível gratuitamente no endereço eletrônico *http://lit.eecs.umich.edu/*.

Os vídeos foram convertidos para o formato MPEG-4 com o tamanho padronizado de 352 *×*288 *pixels*. Os tempos de duração dos vídeos variam de 2-8 minutos. É importante notar na Figura15que há um desequilíbrio na distribuição entre as classes, desfavorável à classe neutra, com apenas 10% do total de instâncias.

Os 105 vídeos foram fragmentados em segmentos menores conforme o enunciado da opinião, resultando num total de 482 vídeos, os quais foram utilizados no processamento desta pesquisa.

Na seção5.3são apresentados os resultados obtidos pelo método proposto e comparados

Figura 14 – Exemplos de vídeos da base de dados MOUD.

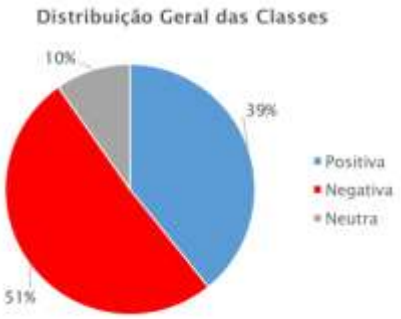


Figura 15 – Distribuição das classes positiva, negativa e neutra.

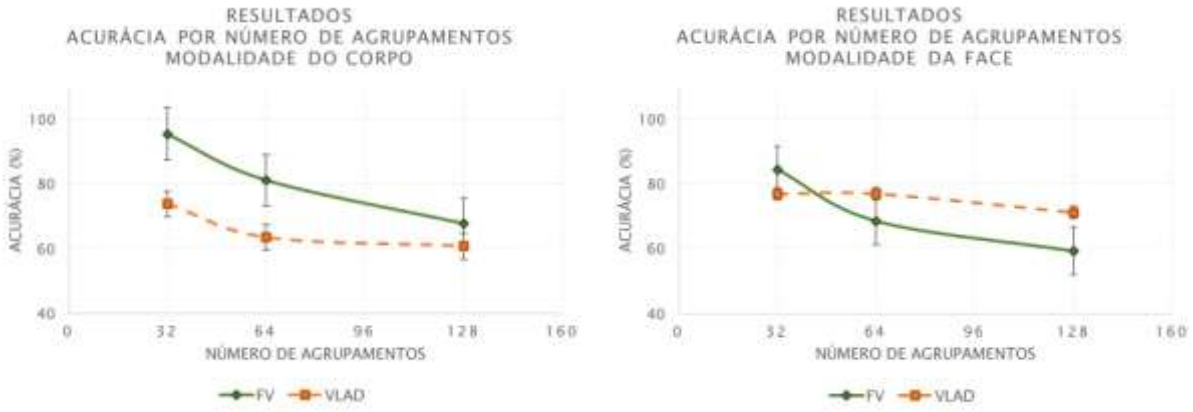
aos *baselines* listados na Tabela2, os quais utilizam em suas soluções três fontes de dados diferentes: vídeo, texto e áudio.

Definição de Parâmetros

Os experimentos com classificadores foram executados com o *software* Waikato Environment for Knowledge Analysis - WEKA, utilizando o algoritmo LibSVM configurado com os parâmetros *C* igual a 1 e *kernel* linear. Os parâmetros adicionais foram mantidos com os valores padrão do *software*. Em função dos métodos de codificação empregados em nossa pesquisa, tanto FV quanto VLAD necessitam de um pré-processamento realizado com os algoritmos de agrupamento *K-means* e o GMM para gerar os dicionários visuais. O objetivo dessa etapa é particionar *n* observações dentre *k* grupos, onde cada observação pertence ao grupo mais próximo da média, sendo o primeiro utilizado na codificação VLAD e o último com exclusividade na codificação FV.

Como forma de definirmos o número de *k* agrupamentos necessários para cada método de codificação realizamos os testes na base MOSI, pois, a mesma apresenta o maior número de instâncias dentre as bases investigadas. Três valores de agrupamentos foram investigados: 32, 64 e 128. A Figura16apresenta os resultados obtidos com as modalidades de corpo (16a) e face (16b) codificadas em FV e VLAD com diferentes valores de agrupamentos.

Tanto a linha sólida no gráfico, a qual representa a codificação FV, quanto a linha



(a) Resultados a partir dos dados do corpo. (b) Resultados a partir dos dados da face.

Figura 16 – Resultados da acurácia obtida para as codificações FV e VLAD nas modalidades de corpo e face em razão do número de agrupamentos.

tracejada representando a codificação VLAD, ilustram uma trajetória decrescente de acurácia à medida que aumenta o número de agrupamentos. Os resultados mostram que, seja na modalidade do corpo ou de face, aplicando qualquer codificador, a acurácia obtida com 32 agrupamentos supera os resultados com agrupamentos maiores. Assim, tomou-se como definição o valor de 32 para o parâmetro *k* em todos os demais experimentos neste trabalho.

Resultados dos Experimentos

Nesta seção descrevemos os resultados alcançados pelo método proposto nas bases de dados investigadas nas subseções:5.3.1,5.3.2e5.3.3. Os resultados obtidos pelo método proposto são comparados aos resultados obtidos pelos *baselines*. Por fim, apresentamos uma análise do desempenho do método proposto ao ser treinado com vídeos de um idioma e testado com vídeos de outro idioma. O objetivo dessa segunda série de experimentos é avaliar se o método é invariante ao idioma, ou seja, se é possível criar um modelo que apresente elevada taxa de acurácia independentemente do idioma com o qual foi treinado. Para tanto, a base de dados MOSI é usada para gerar o modelo e a MOUD para testar. O classificador SVM com *kernel* linear foi empregado em todos os experimentos.

Youtube Dataset

Para fins de comparação dos resultados obtidos nesta base de dados, utilizamos a estratégia de validação cruzada de 10 partes, em conformidade com *baseline* proposto porPoria et al.(2016). Entretanto, é importante salientar que não há garantias de que os mesmos conjuntos de treino e teste do *baseline* foram replicados em nossos experimentos devido a diversos fatores, como exemplo, o uso da semente para a seleção dos conjuntos pode não ter sido a mesma dos autores. As instâncias de treino e teste que compõem a base de dados foram divididas conforme a Tabela 1. A precisão é usada coo métrica de desempenho.

Conforme já mencionado no capítulo3, a pesquisa desenvolvida porPoria et al.(2016) faz uso das modalidades de vídeo, áudio e texto, além de tratar as fusão destas modalidades

Tabela 5 – Resumo de vídeos processados - Banco de Dados *Youtube Dataset*.

Face / Corpo	Nr. vídeo processados			Total
	Positiva	Negativa	Neutra	
Instâncias de treino	13	10	19	42
Instâncias de teste	2	2	1	5
Total	15	12	20	47

de duas formas: baseada em características e decisão. As modalidades de face e corpo tratadas em nossa pesquisa são provenientes somente de vídeo. Os resultados apresentados na Tabela6 demonstram que, tanto para o *baseline* quanto para o nosso método, a fusão das modalidades alcançou um melhor resultado se comparada com os resultados obtidos com as modalidades separadamente.

Tabela 6 – Resultados obtidos pelo método proposto na base *Youtube* comparados aos resultados do *baseline*.

Baseline	**Precisão**
Somente vídeo	0,68
Somente áudio	0,65
Somente texto	0,61
Fusão baseada em características*	0,78
Fusão baseada em decisão*	0,75
Nosso Método - Codiftcação VLAD	
Somente face	0,57
Somente corpo	0,32
Fusão baseada em decisão**	0,60
Nosso Método - Codiftcação FV	
Somente face	0,77
Somente corpo	0,71
Fusão baseada em decisão**	**0,84**

*Vídeo, áudio e texto. **Face e corpo.

A maior precisão apresentada pelo *baseline* foi de 0,78. Essa taxa foi obtida com fusão baseada em características, ao integrar as modalidades de áudio, vídeo e texto. Porém, pode-se observar na tabela que essa taxa é menor do que alcançamos em nosso método com a fusão das modalidades de face e corpo com a codificação FV, que chegou à precisão de 0,84. Se comparamos o resultado da precisão obtido com dados provenientes somente de vídeo pelo *baseline* (0,68), em relação às modalidades de face (0,77) e corpo (0,71) isoladas na codificação FV, mesmo assim, o nosso método supera os resultados.

A codificação VLAD apresentou resultados inferiores ao *baseline*. É importante considerar que a pouca quantidade de instâncias do banco de dados somada à divisão dos conjuntos de treino e teste resultante da validação cruzada faz com que somente 5 instâncias sejam usadas nos testes (ver Tabela5), podendo ocorrer por exemplo, que, com somente 2 instâncias preditas incorretamente pelo classificador tornem o resultado final em 0,60, o que dificulta a análise, pois, cada instância de teste representa 20% do total.

MOSI Dataset

Os resultados obtidos com a base MOSI são apresentados com uso da métrica da acurácia. Os experimentos foram realizados com validação cruzada de 5 partes, semelhante ao *baseline* desenvolvido porZadeh et al.(2016). Como na base anterior, os autores do *baseline* buscaram analisar opiniões empregando as modalidades de vídeo, áudio e texto, além de utilizar a técnica de fusão baseada em decisão e de terem testado dois tipos de classificadores: SVM com *kernel* linear e *Deep Neural Network - DNN*, porém, os melhores resultados relatados foram alcançados com uso de SVM. As instâncias de treino e teste que compõem a base de dados foram divididas conforme aTabela7.

Tabela 7 – Resumo de vídeos processados - Banco de Dados MOSI.

Face / Corpo	Nr. vídeo processados			Total
	Positiva	Negativa	Neutra	
Instâncias de treino	394	355	289	1.038
Instâncias de teste	110	98	52	260
Total	504	453	341	1.298

A Tabela8compara os resultados do *baseline*, ao nosso método. Observamos que, seja no *baseline* ou em nosso método, os experimentos demonstraram de forma recorrente que os resultados das fusões das modalidades alcançam maior acurácia se comparados com os resultados obtidos das modalidades isoladas.

Tabela 8 – Resultados da base MOSI comparados ao *baseline*.

Baseline	**Acurácia**
Somente vídeo	0,61
Somente áudio	0,57
Somente texto	0,65
Fusão baseada em decisão*	0,71
Nosso Método - Codiftcação VLAD	
Somente face	0,77
Somente corpo	0,74
Fusão baseada em decisão**	0,83
Nosso Método - Codiftcação FV	
Somente face	0,79
Somente corpo	0,92
Fusão baseada em decisão**	**0,94**

*Vídeo, áudio e texto. **Face e corpo.

Da mesma forma que os resultados da seção5.3.1, SVM treinado com dados representados via codificação com FV, com 0,94 de acurácia, superou a codificação com VLAD, com a qual foi obtida taxa de 0,83, na fusão das modalidades da face e do corpo. O *baseline* apresentou 0,61 de acurácia para dados provenientes de vídeo e 0,71 com fusão, sendo superado por nosso método com ambas codificações testadas.

MOUD Dataset

Os resultados dos experimentos a seguir são comparados ao trabalho elaborado por Rosas, Mihalcea e Morency(2013), os quais fizeram uso de SVM linear para classificação das modalidades de vídeo, áudio e texto com validação cruzada com 10 partes. A acurácia foi a métrica empregada. Nesta base de dados, os resultados da codificação VLAD foram superiores ao FV. É importante destacar que a codificação FV apresenta melhor desempenho na modalidade de gesto do corpo. Portanto, nossa hipótese para este fato é que a codificação FV na modalidade do corpo foi prejudicada, pois, observamos um grande número de segmentos de vídeos com pouco tempo de execução, inferiores a 10 segundos, o que pode ter contribuído para diminuir a representação do descritor MHI, e, consequentemente do descritor HOG. A Tabela9apresenta o total de segmentos de vídeos processados, bem como, a distribuição das classes no treino e no teste.

Tabela 9 – Resumo de vídeos processados - Banco de Dados MOUD.

Face / Corpo	Nr. vídeo processados			Total
	Positiva	Negativa	Neutra	
Instâncias de treino	163	226	44	433
Instâncias de teste	26	20	3	49
Total	189	246	47	482

Demonstramos por meio da Tabela10a comparação dos resultados obtidos pelo *baseline* em relação ao nosso método proposto. Tanto o *baseline* quanto nosso método apresentaram resultados com maior acurácia nas fusões de modalidades, se comparamos com os resultados obtidos das modalidades isoladamente.

Tabela 10 – Resultados da base MOUD comparados ao *baseline*.

Baseline	**Acurácia**
Somente vídeo	0,61
Somente áudio	0,47
Somente texto	0,65
Fusão baseada em decisão*	0,75
Nosso Método - Codiftcação VLAD	
Somente face	0,93
Somente corpo	0,90
Fusão baseada em decisão**	**0,95**
Nosso Método - Codiftcação FV	
Somente face	0,76
Somente corpo	0,73
Fusão baseada em decisão**	0,80

*Vídeo, áudio e texto. **Face e corpo.

Das três modalidades investigadas separadamente pelo método *baseline* de análise de opinião, a modalidade que obteve maior acurácia foi a de vídeo com 0,61, mesmo assim, ficou abaixo das modalidades de face e corpo usadas em nosso método, independentemente do tipo de codificação empregado em nossa pesquisa. O melhor resultado apresentado pelo *baseline* foi

obtido com a fusão baseada em decisão, precisamente 0,75, também inferior ao nosso método que atinge 0,80 com a codificação FV e 0,95 com a codificação VLAD.

Treino e teste em diferentes idiomas: MOSI *versus* MOUD

Como forma de avaliar com maior profundidade o método que construímos nesta pesquisa, e ainda, demonstrar portabilidade e independência de idiomas em dados opinião multimodal, executamos experimentos utilizando as instâncias do banco de dados MOSI (inglês) na geração do modelo de classificação e as instâncias do banco de dados MOUD (espanhol) para a realização dos testes. Ao todo foram 1.298 instâncias de treino e 482 de teste. O classificador SVM com *kernel* linear foi utilizado nos experimentos e a métrica adotada nesta seção foi a acurácia.

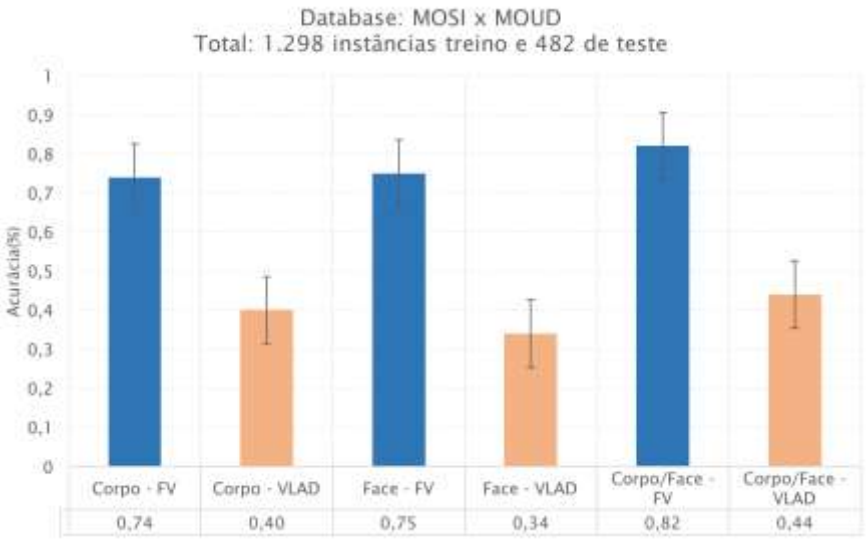
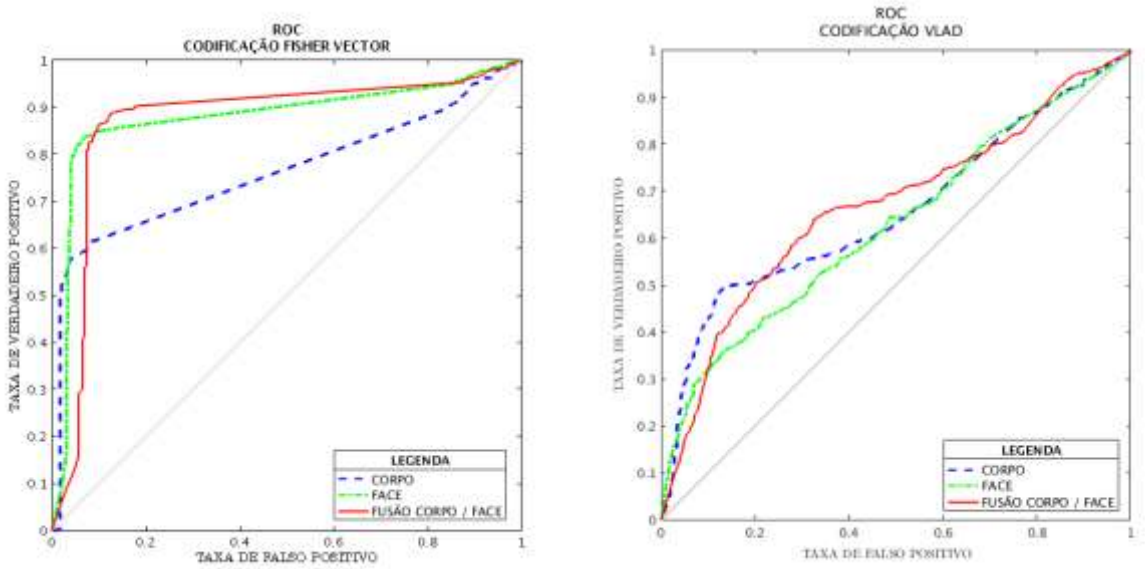


Figura 17 – Resultados dos testes feitos com os bancos de dados: MOSI e MOUD.

A Figura17mostra que os resultados obtidos ao usar a codificação VLAD foram taxas de acurácia com valores baixos, ficando abaixo de 0,50. Por outro lado, a codificação FV possibilitou que melhores resultados fossem obtidos, tanto com dados de corpo (0,74) quanto de face (0,75), porém, novamente o melhor resultado foi alcançado novamente com a fusão das modalidades da face e do corpo com 0,82 de acurácia.

De maneira recorrente, os resultados da fusão foram melhores, e para que possamos analisar e demonstrar o desempenho do método utilizado nesta seção, fizemos uso da *Receiver Operating Characteristic - ROC* como métrica. A curva é criada traçando a Taxa de Verdadeiro Positivo (TVP), também conhecida como Sensibilidade, contra a Taxa de Falso Positivo (TFP), conhecida como Especificidade. A análise ROC fornece ferramentas para selecionar modelos possivelmente ótimos e descartar os sub-ótimos, independente de contexto ou distribuição de classes. De forma simplista, as curvas são ótimas quando sua trajetória está mais próxima do valor 1 no eixo *y*, e indo em direção o valor 1 no eixo *x* pela parte superior. Quando as curvas estão muito próximas umas das outras, impossibilitando realizar uma análise visual mais apurada, podemos recorrer à métrica *Area Under the Curve - AUC* que mede a área sob a curva. O desempenho ótimo da curva alcança o valor de área de no máximo 1.



(a) Resultados da curva ROC com codificação FV.

(b) Resultados da curva ROC com codificação VLAD.

Figura 18 – Resultados da Curva ROC obtida para as codificações FV e VLAD nas modalidades de corpo e face das bases MOSI contra MOUD.

A Figura18(a) ilustra o desempenho do método com a codificação FV das modalidades de face - linha com traço e ponto na cor verde, corpo - linha tracejada em azul, bem como a fusão - linha sólida em vermelho. Visualmente é possível verificar na figura citada que a linha sólida vermelha da fusão encontra-se acima das demais, ou seja, a fusão apresenta um melhor desempenho se comparamos com as demais linhas, sendo que a linha da face vem logo abaixo e o corpo com um pior desempenho. Os cálculos de área sobre a curva (AUC) obtidos foram: Corpo (0,76); Face (0,86); e Fusão da Face e Corpo (0,87), confirmando que a fusão obteve um melhor desempenho mesmo que pequeno.

Na Figura18(b) percebemos que o desempenho obtido pela codificação VLAD foi inferior se compararmos com a codificação FV, mesmo assim, a fusão obtém um melhor desempenho, conforme os cálculos da AUC: Corpo (0,65); Face (0,62); e Fusão da Face e Corpo (0,66).

Em todos os experimentos os resultados da fusão das modalidades da face e do corpo superaram as modalidades isoladamente, mais uma vez sendo comprovada nesta seção. Em geral, a codificação FV obteve melhores resultados com a modalidade do corpo, enquanto VLAD, com a da face, o que colaborou para a diversificação dos resultados e, por consequência, houve uma melhora com a fusão das modalidades. Somente nos experimentos feitos com o banco de dados MOUD a codificação VLAD obteve um melhor desempenho sobre a FV, em todas os outros casos FV superou VLAD, resultado que confirma os estudo dePeng et al.(2014), o qual demonstra a superioridade de FV sobre os demais codificadores.

Pode-se concluir, portanto, que o método proposto apresenta algumas vantagens sobre as soluções correntes, pois, mesmo usando apenas informações de gesto e face, supera em média 16% os *baselines* que usam vídeo, texto, e áudio. Além disso, o método proposto usa extratores

de características clássicos e disponíveis publicamente, enquanto os *baselines* usam *software* proprietário. Em geral, os dados coletados de corpo e face produziram uma diversidade nos resultados dos classificadores, propiciando um melhor desempenho nos resultados da fusão. Por fim, os experimentos demonstram que é possível usar o método para identificar emoções em vídeos com idiomas diferente do idioma usado na base de treino, alcançando uma acurácia de 82%. A conclusão e trabalhos futuros são apresentados no próximo capítulo.

6 Conclusão

Esta dissertação de mestrado teve como objetivo desenvolver um método de classificação de opinião multimodal baseado em informações combinadas das modalidades de expressão facial e do gesto do corpo, provenientes de vídeos *on-line*. Embora existam pesquisas abordando essa problemática, elas apresentam modelos complexos, visto que em suas soluções além do emprego de vídeo como fonte de dados, utiliza-se também áudios e textos contendo as transcrições das falas. Devido a este fato, essas soluções restritas e dependentes do idioma falado nos vídeos. Além do mais, essas pesquisas usam *software* proprietários em pelo menos uma de suas etapas, o que dificulta a reprodução do modelo devido a custos financeiros. Ademais, os *baselines* apresentaram taxas de acurácia que podem ser consideradas baixas, pois seus melhores resultados, em geral, alcançam acurácia de 80%.

O método proposto neste trabalho foi essas pesquisas usam três base de dados e comparado a três diferentes *baselines*, superando-os em aproximadamente 16%. Durante a etapa de análise dos resultados foi possível perceber que o uso de codificadores melhora significativamente as taxas de acurácia do método proposto, mesmo com o emprego de classificadores menos robusto como SVM com *kernel* linear e, sem a necessidade de ajustes de parâmetros. Em geral, o codificador FV alcança melhores resultados se comparado ao VLAD. As modalidades de face e corpo produziram informações complementares para para fornecer o grau de opinião. Este fato, ajudou significativamente a fusão atingir os melhores resultados quando comparados com os resultados das modalidades isoladas, já que ofereceu diversidade na escolha, fato que favoreceu as as regras da fusão.

Consideramos que, por não utilizar dados de áudio e de texto, o método possibilita sua portabilidade independentemente do idioma falado no vídeo. Para corroborar com esta afirmação, experimentos demonstraram ser possível usar o método para classificar opinião em vídeo cujo idioma é diferente do idioma usado para o treino. Nesse contexto, o método proposto ainda obteve 82% de acurácia.

Este estudo apresenta algumas limitações, tais como: os vídeos devem conter somente uma pessoa expressando sua opinião sobre um determinado assunto qualquer, invariavelmente posicionada de frente para a câmera, devendo ser possível ver no mínimo o rosto da pessoa por completo.

Como futuras investigações pretendemos elaborar uma nova base de dados de opinião no idioma português, seguindo os protocolos das demais bases. Além de aprofundar os estudos incorporando outros graus de sentimento expressos na opinião como: fracamente positiva, fracamente negativa, fortemente positiva e fortemente negativa.

Referências

ALMEIDA, H. A. d. M. S.; CAVALCANTI, G. D. d. C. O.; REN, T. I. Seleção dinâmica de classificadores baseada em filtragem e em distância adaptativa. Universidade Federal de Pernambuco, 2014. Citado na página40.

BELL, C.; POP, A. *COS 424: Interacting with Data*. [S.l.], 2008. Citado na página35.

BEN-HUR, A.; WESTON, J. A user's guide to support vector machines. In: *Data mining techniques for the life sciences*. [S.l.]: Springer, 2010. p. 223–239. Citado na página38.

BILIŃSKI, P. T. *Human action recognition in videos* . Tese (Theses) — Université Nice Sophia Antipolis, dez. 2014. Disponível em:<https://tel.archives-ouvertes.fr/tel-01134481>. Citado 4 vezes nas páginas27,37,54e55.

CHEN, S. et al. Recognizing expressions from face and body gesture by temporal normalized motion and appearance features. *Image and Vision Computing*, v. 31, n. 2, p. 175–185, 2013. ISSN 02628856. Disponível em:<http://linkinghub.elsevier.com/retrieve/pii/S0262885612001023>. Citado na página35.

DALAL, N.; TRIGGS, B. Histograms of Oriented Gradients for Human Detection. *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1*, p. 886–893, 2005. ISSN 1063-6919. Disponível em:<citeulike-article-id:3047126$\delimiter"026E30F$nhttp://dx.doi.org/10.1109/CVPR.2005.177>. Citado 2 vezes nas páginas34e53.

DEGOTTEX, G. et al. Covarep x2014; a collaborative voice analysis repository for speech technologies. In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2014. p. 960–964. ISSN 1520-6149. Citado na página44.

FORSYTH, D. A.; PONCE, J. *Computer Vision: A Modern Approach*. [S.l.]: Prentice Hall Professional Technical Reference, 2002. ISBN 0130851981. Citado 2 vezes nas páginas34e54.

GONZALEZ, R.; WOODS, R. *Processamento Digital De Imagens* . [S.l.]: ADDISON WESLEY BRA, 2010. ISBN 9788576054016. Citado 3 vezes nas páginas15,29e30.

GUNES, H.; PICCARDI, M. Affect recognition from face and body: early fusion vs. late fusion. *Systems, Man and Cybernetics, 2005 IEEE International Conference on*, v. 4, p. 3437–3443 Vol. 4, 2005. ISSN 1062922X. Disponível em:<http://ieeexplore.ieee.org/stamp/stamp.jsp?tp={&}arnumber=1571>. Citado na página25.

GUNES, H. et al. Bodily expression for automatic affect recognition. *Advances in Emotion Recognition*, n. July, p. 1–34, 2013. Citado 3 vezes nas páginas25,26e50.

HUSSAIN, M. S.; MONKARESI, H.; CALVO, R. a. Combining Classifiers in Multimodal Affect Detection. *Proceedings of the Tenth Australasian Data Mining Conference (AusDM 2012)*, n. AusDM, p. 103–108, 2012. Citado na página35.

JÉGOU, H. et al. Aggregating local descriptors into a compact image representation. In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2010. p. 3304–3311. ISSN 1063-6919. Citado na página38.

LOBBAN, F. *Implementing clinical guidelines (or not?)*. [S.l.: s.n.], 2008. v. 81. 329–330 p. ISSN 1476-0835. ISBN 8764300080. Citado 4 vezes nas páginas15,31,32e33.

MAAOUI, C.; ABDAT, F.; PRUSKI, A. Physio-visual data fusion for emotion recognition. *Irbm*, v. 35, n. 3, p. 109–118, 2014. ISSN 19590318. Disponível em: <http://www.sciencedirect.com/science/article/pii/S195903181400044X>. Citado na página35.

MIHALCEA, R.; BANEA, C.; WIEBE, J. M. Learning multilingual subjective language via cross-lingual projections. 2007. Citado na página43.

MORENCY, L.-P.; MIHALCEA, R.; DOSHI, P. Towards multimodal sentiment analysis: Harvesting opinions from the web. ACM, New York, NY, USA, p. 169–176, 2011. Disponível em: <http://doi.acm.org/10.1145/2070481.2070509>. Citado 5 vezes nas páginas25,43,45,47e57.

ONEATA, D.; VERBEEK, J.; SCHMID, C. Action and event recognition with fisher vectors on a compact feature set. In: *The IEEE International Conference on Computer Vision (ICCV)*. [S.l.: s.n.], 2013. Citado na página27.

PALEARI, M.; CHELLALI, R.; HUET, B. Features for multimodal emotion recognition: An extensive study. In: *2010 IEEE Conference on Cybernetics and Intelligent Systems*. [S.l.: s.n.], 2010. p. 90–95. ISSN 2326-8123. Citado na página25.

PANNING, A. et al. Multimodal affect recognition in spontaneous HCI environment. *2012 IEEE International Conference on Signal Processing, Communication and Computing (ICSPCC 2012)*, p. 430–435, 2012. Disponível em:<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6335662>. Citado na página26.

PENG, X. et al. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. *CoRR*, abs/1405.4506, 2014. Disponível em: <http://arxiv.org/abs/1405.4506>. Citado 5 vezes nas páginas27,36,47,55e66.

PERRONNIN, F.; DANCE, C. Fisher kernels on visual vocabularies for image categorization. In: *2007 IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2007. p. 1–8. ISSN 1063-6919. Citado na página36.

PERRONNIN, F.; SÁNCHEZ, J.; MENSINK, T. Improving the fisher kernel for large-scale image classification. In:_____. *Computer Vision – ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV*.Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. p. 143–156. ISBN 978-3-642-15561-1. Disponível em:<http://dx.doi.org/10.1007/978-3-642-15561-1_11>. Citado na página36.

PLANET, S.; IRIONDO, I. Comparison between Decision-Level and Feature-Level Fusion of Acoustic and Linguistic Features for Spontaneous Emotion Recognition. *Information Systems and Technologies (CISTI), 2012 7th Iberian Conference on*, p. 1–6, 2012. Disponível em:<http://ieeexplore.ieee.org/xpl/login.jsp?tp={&}arnumber=6263129{&}url=http://ieeexplore.ieee.org/xpls/abs{_}all.jsp?arnumbe>. Citado na página40.

PORIA, S. et al. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*, v. 174, Part A, p. 50 – 59, 2016. ISSN 0925-2312. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0925231215011297>. Citado 4 vezes nas páginas25,45,47e61.

ROSAS, V. P.; MIHALCEA, R.; MORENCY, L. P. Multimodal sentiment analysis of spanish online videos. *IEEE Intelligent Systems*, v. 28, n. 3, p. 38–45, May 2013. ISSN 1541-1672. Citado 5 vezes nas páginas25,44,47,57e64.

SANTOS, E. M. D.; SABOURIN, R.; MAUPIN, P. A dynamic overproduce-and-choose strategy for the selection of classifier ensembles. *Pattern Recognition*, v. 41, n. 10, p. 2993–3009, 2008. ISSN 00313203. Disponível em:<http://linkinghub.elsevier.com/retrieve/pii/S003132030800126X>. Citado na página40.

TIAN, Y. et al. Hierarchical filtered motion for action recognition in crowded videos. *Systems, Man, and . . .* , v. 42, n. 3, p. 313–323, 2012. ISSN 1094-6977. Disponível em:<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5772028$\ delimiter"026E30F$nhttp://ieeexplore.ieee.org/xpls/abs{_}all.jsp?arnumber=5772>. Citado 2 vezes nas páginas33e52.

TRUCCO, E.; VERRI, A. *Introductory Techniques for 3-D Computer Vision*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1998. ISBN 0132611082. Citado na página29.

VEDALDI, A.; FULKERSON, B. *VLFeat: An Open and Portable Library of Computer Vision Algorithms*. 2008.<http://www.vlfeat.org/>. Citado na página54.

VIOLA, P.; JONES, M. Rapid object detection using a boosted cascade of simple features. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001. ISSN 1063-6919. Citado 3 vezes naspáginas15,31e32.

WAGNER, J. et al. Exploring Fusion Methods for Multimodal Emotion Recognition with Missing Data. *IEEE Transactions on Affective Computing*, v. 2, n. 4, p. 206–218, oct 2011. ISSN 1949-3045. Disponível em:<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5871582>. Citado na página40.

WITTEN, I. H.; FRANK, E.; HALL, M. A. *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd. ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011. ISBN 0123748569, 9780123748560. Citado 4 vezes nas páginas15,38,39e40.

WOOD, E. et al. Rendering of eyes for eye-shape registration and gaze estimation. *CoRR*, abs/1505.05916, 2015. Disponível em:<http://arxiv.org/abs/1505.05916>. Citado na página 44.

ZADEH, A. Micro-opinion sentiment intensity analysis and summarization in online videos. ACM, New York, NY, USA, p. 587–591, 2015. Disponível em:<http://doi.acm.org/10.1145/2818346.2823317>. Citado 3 vezes nas páginas25,45e47.

ZADEH, A. et al. MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *CoRR*, abs/1606.06259, 2016. Disponível em: <http://arxiv.org/abs/1606.06259>. Citado 7 vezes nas páginas25,43,45,47,57,58e63.