Islenho de Almeida

Avaliação de Técnicas de Reconhecimento de Padrões na Classificação de Eventos Baseados em Vetores de Deslocamento

Dissertação apresentada ao Programa de Pós-Graduação em Informática Aplicada da Pontifícia Universidade Católica do Paraná como requisito parcial para obtenção do título de Mestre em Informática Aplicada.

Curitiba
2005

Islenho de Almeida

Avaliação de Técnicas de Reconhecimento de Padrões na Classificação de Eventos Baseados em Vetores de Deslocamento

Dissertação apresentada ao Programa de Pós-Graduação em Informática Aplicada da Pontifícia Universidade Católica do Paraná como requisito parcial para obtenção do título de Mestre em Informática Aplicada.

Área de Concentração: Ciência da Imagem

Orientador: Dr. Díbio Leandro Borges
Co-orientador: Dr. Alceu de Souza Britto Jr.

Curitiba
2005

Almeida, Islenho de
Avaliação de Técnicas de Reconhecimento de Padrões na Classificação de Eventos Baseados em Vetores de Deslocamento. Curitiba, 2005.

Dissertação - Pontifícia Universidade Católica do Paraná Programa de Pós-Graduação em Informática Aplicada.

1. Visão por Computador 2. Reconhecimento de Padrões 3. Classificação de Eventos I.Pontifícia Universidade Católica do Paraná. Centro de Ciências Exatas e Tecnologia. Programa de Pós-Graduação em Informática Aplicada II - t

À Deus, por me permitir saber que existo.

Agradecimentos

À minha tia Eveline, agora minha segunda mãe, pelo apoio financeiro, pelo carinho, pelo calor humano tão encorajador quando a saúde me faltou, pelos ensinamentos espirituais e de vida. Ao meu pai, por sempre, sempre acreditar em mim. À minha mãe, pela incansável dedicação para comigo. Ao meu sobrinho, William, por me dar tantas alegrias. À minha irmã, Tatiana, e ao meu cunhado Edson, pelo companheirismo em todas as situações. Aos meus amigos de mestrado e agora de vida, Fausto Vanin, Paulo Cavalin, Éderson Sgarbi, Francis Baranoski, Cristiane e William Ferreira, Fernanda Ramos, David Menoti, Kristian Capeline e Luiz Eiterer. Aos meus orientadores Prof. Díbio Borges e Prof. Alceu Britto pela oportunidade, pelos ensinamentos, pelo companheirismo, pela incondicional presteza em momentos difíceis e, principalmente, pela amizade construída. Aos meus amigos Cícero Wiecheteck, Rafael Wiecheteck, Guilherme, Lucas Madalozzo, Claudia Nekatschalow, Sintya Prestes, Elisa Rocha, Juliano Baniski, Fabiano Hasegawa, Marcelo Ferrasa, Arion de Campos e Alexandre Painka, pela amizade e companheirismo de longa data. À Gislaine Lima, pela insistência em me fazer uma pessoa melhor organizada e, principalmente, por confiar e acreditar em mim.

> "Solidários, seremos união. Separados uns dos outros, seremos pontos de vista." *Bezerra de Menezes.*

Sumário

Agradecimentos ii
Sumário iii
Lista de Figuras v
Lista de Tabelas viii
Lista de Símbolos ix
Lista de Abreviações x
Resumo xi
Abstract xii

Capítulo 1

Introdução 1
1.1 Objetivos 6
1.2 Justificativas 9
1.3 Contribuições 9
1.4 Organização 10

Capítulo 2

Fundamentação Teórica e Trabalhos Relacionados 11
2.1 Formação e concepção de vídeos 11
2.2 O padrão MPEG 12
2.3 Eventos em vídeo 15
2.4 Trabalhos relacionados 16
2.5 Discussão 18

Capítulo 3

Método Proposto 20
3.1 Especificações técnicas 20
3.2 Extrator de características 22
3.3 Medidas de divergência 23
3.4 Aglomeração 25
3.4.1 Algoritmo "K-Médias" 25
3.4.2 Algoritmo "ISODATA" 26
3.4.3 Algoritmo "Normalized Cut" 27

Capítulo 4

Experimentos 29
4.1 Protocolo de testes 29
4.2 Bases de dados 31
4.2.1 Vídeo "Movimentos" 31
4.2.2 Vídeo "Tênis" 34
4.2.3 Vídeo "Inria" 37
4.3 Experimento com o vídeo "Movimentos" 41
4.3.1 Resultados para 5 classes 42
4.3.2 Resultados para 2 classes 46

Capítulo 5

Conclusões e Trabalhos Futuros 52
Referências Bibliográficas 54

Apêndice A

Outros Experimentos 56
A.1 Experimento com o vídeo "Tênis" 56
A.1.1 Resultados para 3 classes 56
A.1.2 Resultados para 2 classes 57
A.2 Experimento com o vídeo "Inria" 59
A.2.1 Resultados para 4 classes 60
A.2.2 Resultados para 2 classes 61

Lista de Figuras

1.1 Quadros representativos de um evento rotulado como “andar” – Adaptado de [ZELNIK-MANOR and IRANI, 2001b] 3
1.2 Evento “andar” transcorrido durante cem quadros 4
1.3 Representação dos vetores de deslocamento para quadros 481 e 482 – Adaptado de [ZELNIK-MANOR and IRANI, 2001c] 5
1.4 Extração de vetores de deslocamento para a geração de valores de divergência 7
1.5 Aplicação de vetores de deslocamento em uma janela temporal de tamanho 2 para a geração de valores de divergência 8

2.1 Quadros explicativos sobre a concepção e utilização de vetores de deslocamento – Adaptado de [ZELNIK-MANOR and IRANI, 2001c] 14

3.1 Diagrama de blocos do extrator de características. Vetores de deslocamento são extraídos para coordenadas x e y para cada janela temporal. 21
3.2 Diagrama que demonstra as características extraídas para cada janela temporal (T). 22

4.1 Exemplo de quadros extraídos do vídeo “Movimentos” representando o evento “rolar” 32
4.2 Exemplo de quadros extraídos do vídeo “Movimentos” representando o evento “andar” 32
4.3 Exemplo de quadros extraídos do vídeo “Movimentos” representando o evento “acenar” 33
4.4 Exemplo de quadros extraídos do vídeo “Movimentos” representando o evento “correr” 33
4.5 Exemplo de quadros extraídos do vídeo “Tênis” representando o evento “batida” 34

4.6 Exemplo de quadros extraídos do vídeo “Tênis” representando o evento “salto” 35
4.7 Exemplo de quadros extraídos do vídeo “Tênis” representando o evento “cadência” 36
4.8 Exemplo de quadros extraídos do vídeo “Inria” representando o evento “diagonal” 38
4.9 Exemplo de quadros extraídos do vídeo “Inria” representando o evento “encontro” 39
4.10 Exemplo de quadros extraídos do vídeo “Inria” representando o evento “horizontal” 40
4.11 Exemplo de quadros extraídos do vídeo “Inria” representando o evento “vertical” 41
4.12 Resultados de Precisão para vídeo “Movimentos”, 5 classes, melhor caso, $\sigma = 0,40$. 42
4.13 Resultados de Revocação para vídeo “Movimentos”, 5 classes, melhor caso, $\sigma = 0,40$. 43
4.14 Resultados de Precisão para vídeo “Movimentos”, 5 classes, pior caso, $\sigma = 0,95$. 44
4.15 Resultados de Revocação para vídeo “Movimentos”, 5 classes, pior caso, $\sigma = 0,95$. 45
4.16 Resultados de Precisão para vídeo “Movimentos”, 2 classes, melhor caso, $\sigma = 0,55$. 47
4.17 Resultados de Revocação para vídeo “Movimentos”, 2 classes, melhor caso, $\sigma = 0,55$. 48
4.18 Resultados de Precisão para vídeo “Movimentos”, 2 classes, pior caso, $\sigma = 0,85$. 49
4.19 Resultados de Revocação para vídeo “Movimentos”, 2 classes, pior caso, $\sigma = 0,85$. 50

A.1 Resultados de Precisão para vídeo “Tênis”, 3 classes, melhor caso, $\sigma = 0,85$. 57
A.2 Resultados de Revocação para vídeo “Tênis”, 3 classes, melhor caso, $\sigma = 0,85$. 58
A.3 Resultados de Precisão para vídeo “Tênis”, 3 classes, pior caso, $\sigma = 0,10$. . 59
A.4 Resultados de Revocação para vídeo “Tênis”, 3 classes, pior caso, $\sigma = 0,10$. 60
A.5 Resultados de Precisão para vídeo “Tênis”, 2 classes, melhor caso, $\sigma = 0,50$. 61

A.6 Resultados de Revocação para vídeo “Tênis”, 2 classes, melhor caso, $\sigma = 0,50$. 62
A.7 Resultados de Precisão para vídeo “Tênis”, 2 classes, pior caso, $\sigma = 0,60$. . 63
A.8 Resultados de Revocação para vídeo “Tênis”, 2 classes, pior caso, $\sigma = 0,60$. 64
A.9 Resultados de Precisão para vídeo “Inria”, 4 classes, melhor caso, $\sigma = 0,85$. 65
A.10 Resultados de Revocação para vídeo “Inria”, 4 classes, melhor caso, $\sigma = 0,85$. 66
A.11 Resultados de Precisão para vídeo “Inria”, 4 classes, pior caso, $\sigma = 0,15$. . 67
A.12 Resultados de Revocação para vídeo “Inria”, 4 classes, pior caso, $\sigma = 0,15$. 68
A.13 Resultados de Precisão para vídeo “Inria”, 2 classes, melhor caso, $\sigma = 0,05$. 69
A.14 Resultados de Revocação para vídeo “Inria”, 2 classes, melhor caso, $\sigma = 0,05$. 70
A.15 Resultados de Precisão para vídeo “Inria”, 2 classes, pior caso, $\sigma = 0,10$. . 71
A.16 Resultados de Revocação para vídeo “Inria”, 2 classes, pior caso, $\sigma = 0,10$. 72

Lista de Tabelas

4.1 Especificações dos vídeos que compõem a base de dados 31
4.2 Resultados gerais dos vídeos que compõem a base de dados 47

Lista de Símbolos

Y	Luminância
U	Crominância
V	Crominância
VDx	Vetores de deslocamento em x
VDy	Vetores de deslocamento em y
H	Histograma
CH	Histograma Acumulado

Lista de Abreviações

MPEG	*Motion Picture Expert Group*
DVD	*Digital Versatile Disc*
HDTV	*High Definition Television*
IP	*Internet Protocol*
DCT	*Transformada de Cossenos Discreta*
VD	*Vetores de Deslocamento*
QPS	*Quadros por Segundo*
bits	*Conjunto de Bit – Dígito binário*
Mbytes	*Conjunto de 8.388.608 bits*
ISO	*International Organization for Standardization*
VCD	*Vídeo Digital em CD-ROM*
CODEC	*COdificador/DECodificador*
ISODATA	*Iterative Self-Organizing Data Analysis Techniques*
VDI	*Vetor de Deslocamento Independente (x,y)*
VDR	*Vetor de Deslocamento Resultante*

Resumo

Este trabalho descreve um método de avaliação de diferentes técnicas de reconhecimento de padrões para classificação de eventos em vídeos. Utilizam-se características diretas no espaço da imagem, baseadas em vetores de deslocamento, obtidas diretamente dos vídeos. A partir dessas características, diferentes medidas de divergência são aplicadas em subconjunto de dados no intuito de encontrar similaridades entre os mesmos. Um conjunto dessas medidas são fornecidas como entradas para diferentes classificadores. O resultado obtido desta classificação automática é comparado com resultados previamente rotulados através do processo *ground-truth*, gerando duas medidas de avaliação quantitativas, Precisão e Revocação, que serão utilizadas como parâmetros de avaliação final.

Palavras-chave: Visão por Computador, Reconhecimento de Padrões, Classificação de Eventos, Vetores de Deslocamento.

Abstract

This work describes an evaluation method of different pattern recognition techniques for classification of events based on videos. It uses direct features on the image space, based on motion vectors, acquired directly from the videos. From this features, different similarity measures are applied in a data subset in order to find similarities among themselves. A set of these measures is provided as an input for different classifiers. The result reached from this automatic classification is compared to prelabeled results throught a ground-truth process, producing two quantitative evaluation measures, Precision and Recall, which will be used as final evaluation parameters.

Keywords: Computer Vision, Patterns Recognition, Events Classification, Motion Vectors.

Capítulo 1

Introdução

A classificação de vídeos baseada em movimento faz parte dos projetos de pesquisa das comunidades de Reconhecimento de Padrões e Visão por Computador, pois a cada dia cresce a quantidade de informações gráficas (imagens), com movimento, armazenadas na forma de vídeo.

Armazenar tais informações de maneira crua, ou seja, sem nenhuma espécie de compactação temporal ou espacial demanda grandes áreas de armazenamento gerando dificuldades no seu transporte e exibição/difusão.

Com o intuito de prover uma forma mais eficiente neste processo de armazenamento/difusão, foi criado um comitê internacional demoninado MPEG (*Motion Picture Expert Group*) [1]. Este comitê cria padrões para compressão e difusão de áudio e vídeo. Dentre os vários padrões criados por este comitê, ressaltamos o padrão MPEG-2 que nasceu no início dos anos 90 sendo efetivamente lançado em 1994 [SIKORA, 1997].

Esta codificação tornou-se o padrão de fato na indústria devido à larga produção de componentes eletrônicos capazes de codificar e decodificar vídeo e áudio neste formato. É crescente o número de equipamentos eletrônicos que são concebidos para trabalharem com o padrão MPEG-2, reduzindo cada vez mais o seu custo.

Pode-se citar casos como os aparelhos de DVD (*Digital Versatile Disc*), a televisão de alta definição[2] – HDTV (*High Definition Television*) além de um grande número de câmeras. Dentre estas, existem as câmeras IP (*Internet Protocol*) (câmeras que capturam áudio e vídeo e retransmitem estas informações no padrão MPEG-2 diretamente para uma rede de computadores, ou de outros dispositivos, economizando banda e tempo de transmissão).

Assim, o padrão MPEG-2 torna-se um padrão interessante de ser explorado pois

[1] http://www.chiariglione.org/mpeg/
[2] http://www.hdtv.net

este padrão possui compressão espacial através da DCT (*Transformada de Cossenos Discreta*) e temporal através da utilização de VD (*Vetores de Deslocamento*).

Esta dissertação tem como tema central a avaliação de técnicas utilizadas para detectar a similaridade do conteúdo visual de um vídeo baseado em características de baixo nível. As características de baixo nível, utilizadas neste trabalho, são os vetores de deslocamento, que estão intrinsecamente codificados no padrão MPEG-2.

A maneira utilizada, neste trabalho, para medir a similaridade do conteúdo visual de um vídeo é agrupar as características de baixo nível, vetores de deslocamento, na forma de janelas temporais, a fim de que cada janela temporal venha a caracterizar um evento ocorrido no vídeo.

Um evento é um acontecimento temporal que pode ser periódico ou não-periódico além de possuir uma extensão temporal, ou seja, um mesmo evento pode ocorrer em diferentes tamanhos de janelas temporais.

Na Figura 1.1 pode-se observar o acompanhamento temporal, resumido nos quadros 250, 275, 300, 325 e 350 representados pelas Figuras 1.1(a) a 1.1(e), de um determinado evento que foi rotulado como evento "andar". Os quadros representados pelas Figuras 1.1(a), 1.1(c) e 1.1(e) foram combinados a fim de se obter a Figura 1.2 com o intuito de demonstrar o acontecimento do evento "andar", que transcorreu durante cem quadros, de forma resumida.

O evento "andar" representado pela Figura 1.2, levou 4 segundos (100 quadros a 25 QPS (*Quadros por Segundo*)) para percorrer uma determinada distância. Porém, o mesmo evento pode ser executado em mais ou menos de cem quadros. Isso caracteriza que eventos não são objetos fixos, eles possuem uma certa dinâmica em sua execução e esta dinâmica deve ser levada em consideração quando da sua modelagem.

Modelar um evento de forma minuciosa não é interessante quando se busca uma aplicação que possa ocorrer nos mais variados ambientes e nas mais variadas condições. Dessa forma, busca-se avaliar técnicas que tratem do problema de forma menos restritiva, perdendo em precisão, mas que podem ser aplicadas em ambientes e em condições genéricas.

A Figura 1.2 demonstra quadros extraídos de um vídeo codificado no padrão MPEG-2 com uma taxa de 25 QPS. Porém, esta figura tem por objetivo apenas tornar mais evidente a concepção de um evento simples. Como visto anteriormente, as técnicas avaliadas utilizam informações de baixo nível, os vetores de deslocamento.

Um exemplo destas informações, representadas através de setas, pode ser visto na Figura 1.3 onde demonstra-se os vetores de deslocamento para cada macrobloco (Figura 1.3(c)), obtidos dos quadros 481 (Figura 1.3(a)) e 482 (Figura 1.3(b)) do vídeo

(a) Quadro 250

(b) Quadro 275

(c) Quadro 300

(d) Quadro 325

(e) Quadro 350

Figura 1.1: Quadros representativos de um evento rotulado como “andar” – Adaptado de [ZELNIK-MANOR and IRANI, 2001b]

Figura 1.2: Evento “andar” transcorrido durante cem quadros

“Tênis”[ZELNIK-MANOR and IRANI, 2001c]

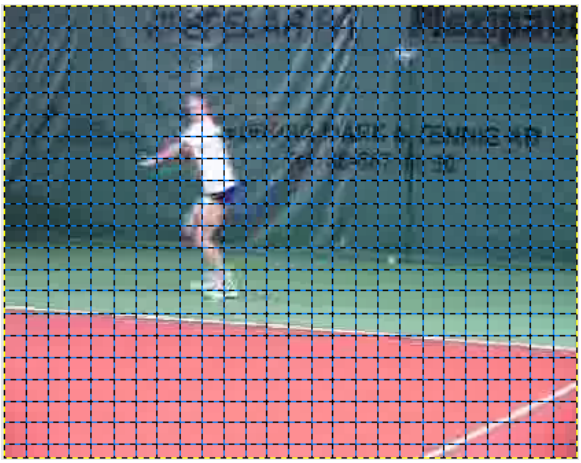
Definindo evento como um acontecimento temporal, pode-se representar suas características através dos vetores de deslocamento (Figura 1.3(c)). Um evento é um deslocamento ocasionado por um objeto que variou a taxa de luminância para uma determinada área de um quadro gerando uma representação na forma de vetores de deslocamento.

Um evento pode acontecer somente no plano horizontal (x), somente no plano vertical (y), ou ainda em ambos, no caso de um evento que ocorre de forma irregular ou diagonal.

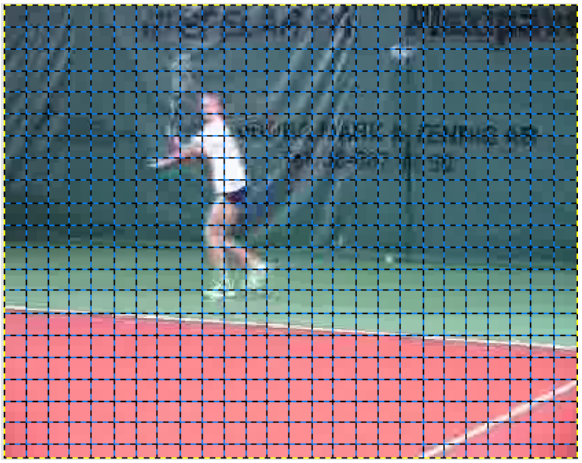
Baseado nesses critérios de deslocamento, busca-se modelar diferentes classes de eventos baseado na quantidade de movimentação no plano horizontal e vertical representada pelos vetores de deslocamento. Dessa forma, busca-se modelar eventos pela sua orientação e pela quantidade de movimentação/deslocamento exercida durante um determinado período, representado aqui na forma de uma janela temporal.

Considerando, dessa forma, evento como uma coleção temporal de vetores de deslocamento que exprimem, na sua magnitude, a quantidade de movimentação no plano horizontal e vertical, pode-se aplicar uma métrica que calcule entre duas janelas temporais o quanto uma diverge de outra, com o intuito de medir se ambas janelas denotam um evento similar ou divergente.

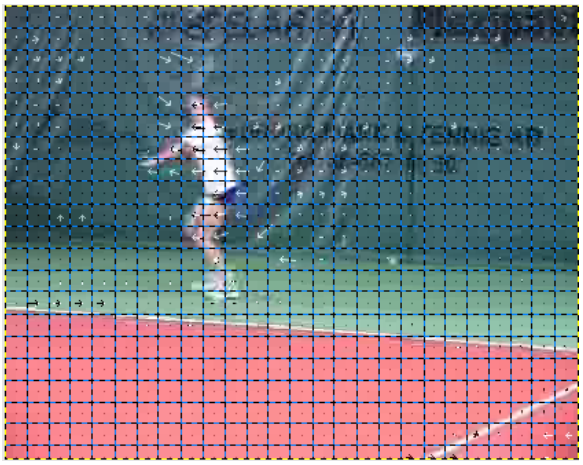
Estabelecida uma medida de divergência entre janelas temporais, pode-se reunir várias dessas medidas e classificá-las em grupos ou classes. Dessa maneira é possível inferir sobre a constância de determinados eventos que são realizados em um determinado ambiente.



(a) Quadro 481



(b) Quadro 482



(c) Vetores de deslocamento obtidos das Figuras 1.3(a) e 1.3(b)

Figura 1.3: Representação dos vetores de deslocamento para quadros 481 e 482 – Adaptado de [ZELNIK-MANOR and IRANI, 2001c]

1.1 Objetivos

Este trabalho visa realizar uma avaliação de diferentes técnicas, utilizando características de baixo nível, intrínsecas de um vídeo, para a classificação de eventos através da aplicação de medidas de divergência e de algoritmos de aglomeração.

A aplicação de medidas de divergência, fornece uma maneira estocástica de se obter valores de similaridade entre eventos.

Este processo é evidenciado pela Figura 1.4, onde demonstra-se o processo de concepção dos vetores de deslocamento a partir de dois quadros e pela Figura 1.5 onde demonstra-se a utilização de vetores de deslocamento para a geração de medidas de divergência.

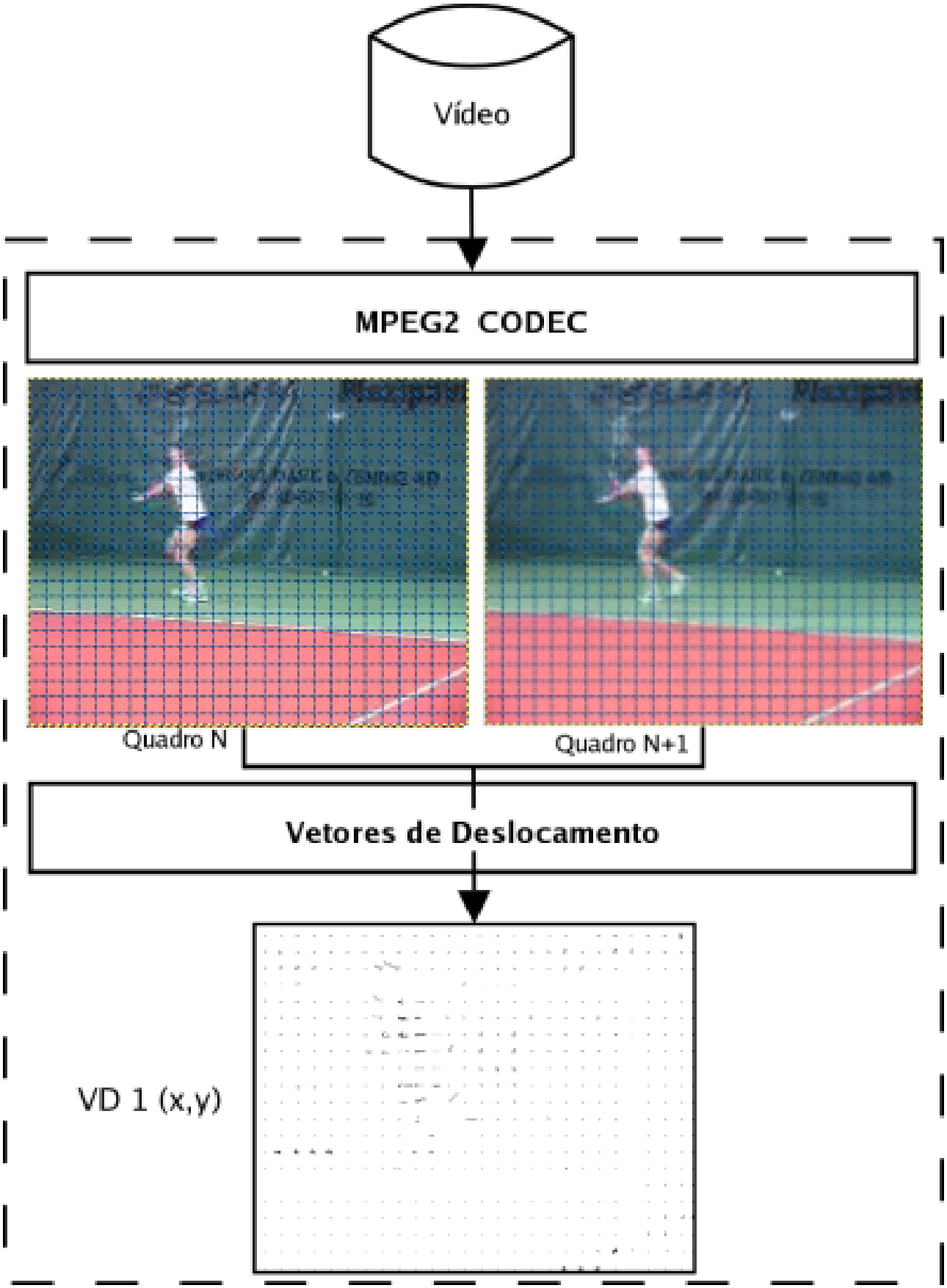


Figura 1.4: Extração de vetores de deslocamento para a geração de valores de divergência

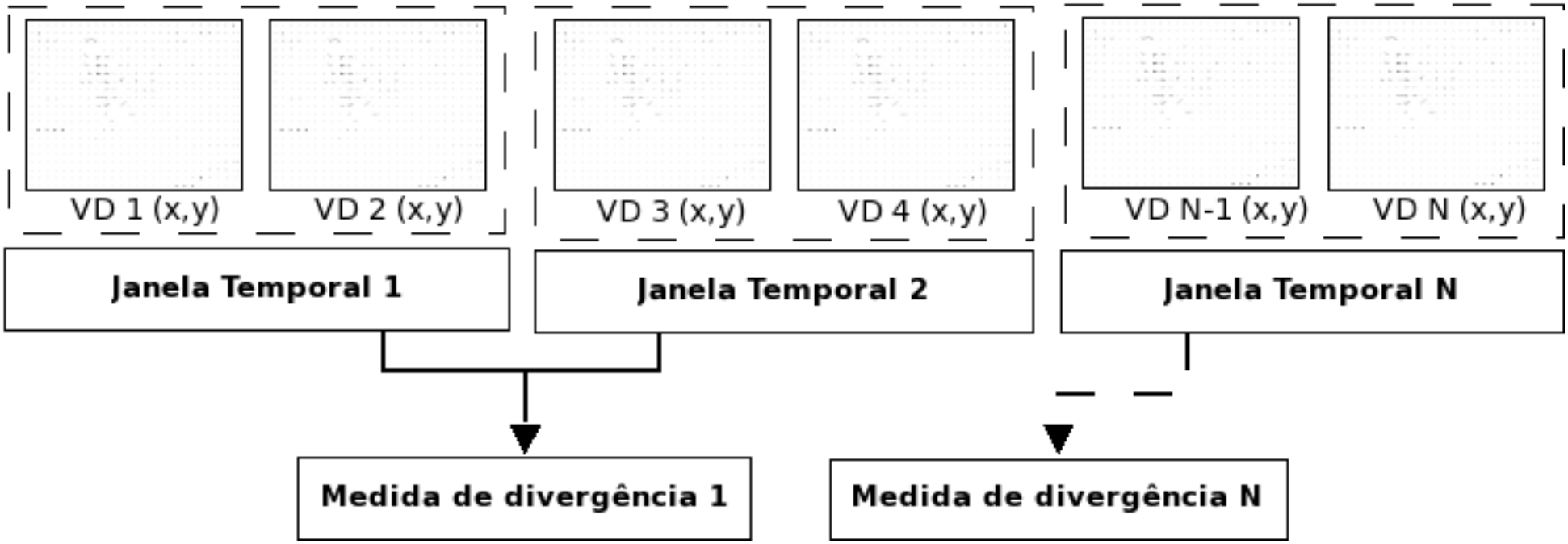


Figura 1.5: Aplicação de vetores de deslocamento em uma janela temporal de tamanho 2 para a geração de valores de divergência

Diferentes técnicas de agrupamento podem ser realizadas em cima de um grupo de tais medidas. Este processo, tem por objetivo, agrupar/classificar eventos representados por medidas de divergência.

Para atingir objetivo principal desta dissertação, devem ser cumpridas as seguintes etapas:

- Avaliar diferentes métricas de divergência utilizando vetores de deslocamento na forma de janelas temporais;
- Avaliar diferentes algoritmos de aglomeração para diferentes medidas de divergência;
- Empregar uma forma de avaliação quantitativa.

A aplicação de uma técnica de avaliação quantitativa é de fundamental importância. Deve-se relembrar que as medidas de divergência aplicadas nesse tipo de problema não são restritivas quanto como em outros modelos. Dessa forma, aliar uma avaliação qualitativa em problemas que utilizam modelos não-paramétricos gera um maior índice de incerteza na verificação dos resultados.

1.2 Justificativas

Combinar características de baixo nível, que estão intrinsecamente codificadas em um padrão de fato, com métricas de divergência aplicadas de maneira a evidenciar a diferenciação de eventos une o lado tecnológico com tópicos recentes de pesquisa nas comunidades de Reconhecimento de Padrões e Visão por Computador.

Este trabalho visa analisar a melhor forma de se estabelecer uma correlação entre ambos. Dessa maneira, pode-se chegar a sistemas que serão capazes de supervisionar ambientes com a melhor taxa de precisão possível.

Deve-se lembrar que mesmo o lado tecnológico nada mais é que a consolidação de procedimentos científicos. Saber extrair o melhor do conhecimento científico já existente e combiná-lo, da melhor forma, com conhecimentos científicos em desenvolvimento é um processo inerente que a evolução da ciência requer.

1.3 Contribuições

A maior contribuição deste trabalho é o processo de análise realizado na combinação de diferentes modelos estatísticos, utilizando características de baixo nível, codificadas diretamente em um padrão de fato da indústria, obtidas de vídeos.

Outras contribuições são a análise de forma quantitativa de resultados e a conseqüente avaliação de possibilidade para sistemas que sejam executados em tempo real.

1.4 Organização

Esta dissertação está organizada em cincos capítulos mais um apêndice. Além deste capítulo, no Capítulo 2 é dada uma fundamentação teórica com o intuito de tornar alguns termos mais claros juntamente com uma revisão sobre os trabalhos relacionados. O método proposto de avalição das técnicas é apresentado no Capítulo 3. Os experimentos realizados neste trabalho no sentido de validar o método proposto são apresentados no Capítulo 4 e por fim no Capítulo 5 é feito um apanhado geral no sentido de dar-se uma conclusão.

No Apêndice A são discutidos os demais experimentos realizados mas que não foram inclusos diretamente no Capítulo 4 por este somente conter apenas um estudo de caso, realizado com o Vídeo “Movimentos”.

Capítulo 2

Fundamentação Teórica e Trabalhos Relacionados

Este capítulo tem por objetivo evidenciar alguns termos e procedimentos que servem de base para o trabalho. A formação de um vídeo e algumas de suas formas de concepção serão os assuntos principais deste capítulo.

2.1 Formação e concepção de vídeos

Pode-se definir vídeo como uma coleção temporal de imagens, ou quadros que é o termo mais utilizado nesta área.

Cada quadro quantifica uma determinada intensidade de luminosidade e são representados, computacionalmente, na forma mais trivial através de matrizes. Cada elemento desta matriz é denominado "pixel". Quadros que possuem apenas um canal de medida de intensidade representam, no máximo, 8bits (*Conjunto de Bit – Dígito binário*) de informação por "pixel". Este tipo de quadro representa, no máximo, 256 níveis/tons de cinza.

Vídeos em cores são concebidos, de forma geral, com três canais de intensidade luminosa, a saber: vermelho, verde e azul. Este sistema de representação de cor em canais Vermelho, Verde e Azul é conhecido como sistema RGB em alusão ao termo em inglês (*Red*, *Green* e *Blue*). Estes canais, quando combinados, fornecem a representação de várias cores (24 bits).

A união simples de vários quadros forma um vídeo não compactado. Existe ainda a possibilidade de unir mais, ou menos, quadros, para um mesmo segundo de vídeo, para serem exibidos posteriormente. Esta medida caracteriza uma taxa de exibição que é medida em QPS.

Como pode-se notar, para um vídeo simples com quadros de 320×288 pixels, com

duração de 1 minuto, em cores, com uma taxa de exibição de 25 QPS (para fornecer a impressão de movimento natural) temos o seguinte cálculo para representação da quantidade de informação em Mbytes (*Conjunto de 8.388.608 bits*):

$$Mbytes = 320 \times 288 \times 3 \times 60 \times 25 = 395,5 \tag{2.1}$$

Este exemplo demonstra a necessidade de se criar formas de reduzir a representação de informação em vídeos no sentido de haver menor demanda por espaço de armazenamento e de transmissão.

Muitas técnicas de compactação e representação de vídeos vem surgindo. Algumas dessas técnicas buscam reduzir a representação da informação através de procedimentos de compactação. Estes procedimentos de compactação podem ser realizados de forma temporal e espacial.

A compactação espacial diz respeito ao procedimento de se representar, da melhor forma, as informações contidas em cada quadro utilizando um menor número de dados ou ainda utilizando outro espaço representacional. Já a compactação temporal vem como um elemento auxiliar na tentativa de armazenar somente elementos que mudem sua representação entre quadros.

2.2 O padrão MPEG

O MPEG é um grupo que trabalha para gerar especificações para a ISO (*International Organization for Standardization*). A sua especificação para compactação de vídeos é comumente chamada por MPEG a qual será tratada aqui de forma suscinta visando facilitar o seu macro-entendimento por se tratar de um padrão complexo, se analisado em sua plenitude.

O grupo MPEG conta atualmente com dois padrões finalizados e bastante populares, o MPEG-1 e MPEG-2. Estes dois padrões são bastante similares, eles se baseiam na utilização de transformadas de cossenos para a compressão espacial e na compensação de deslocamento baseada em blocos para compressão temporal.

A maior distinção entre os dois padrões é relacionada à resolução espacial dos quadros codificados (352×240 "pixels", 30 QPS para MPEG-1 e 720×480 "pixels", 30 QPS para MPEG-2) e sua aplicação principal (VCD (*Vídeo Digital em CD-ROM*) para MPEG-1, TV Digital e HDTV para MPEG-2).

Para realizar compressão espacial, o padrão MPEG-2 primeiramente converte a imagem para um novo espaço de cor denominado YUV (Y, U, V) e realiza uma reamos-

tragem dos componentes de crominância (U e V) pela metade.

Este procedimento é realizado devido ao fato de que informações de brilho (luminância – Y) são, fisiologicamente, mais relevantes para a visão humana quando comparado com informações de crominância (U e V)[SIKORA, 1997].

As Equações 2.2, 2.3 e 2.4 são utilizadas para derivar o espaço de cor YUV do espaço de cor RGB[Group, 2003].

$$Y = 0,299R + 0,587G + 0,114B \tag{2.2}$$

$$U = 0,492(B - Y) = -0,147R - 0,289G + 0,436B \tag{2.3}$$

$$V = 0,877(R - Y) = 0,615R - 0,515G - 0,100B \tag{2.4}$$

Composto o novo espaço de cor, os quadros são divididos em macroblocos de tamanho 16×16 pixels. Duas camadas tratam da compressão espacial (um macrobloco de 16×16 pixels para Y e outro reamostrado pela metade (8×8) para U e V). Cada macrobloco destes será submetido à transformada de cossenos e coeficientes DCT (definido pela Equação 2.5) serão armazenados para compor o novo quadro codificado.

$$t(i,j) = c(i,j) \sum_{n=1}^{N-1} \sum_{m=0}^{N-1} s(m,n) \cos \frac{\pi(2m+1)i}{2N} \cos \frac{\pi(2n+1)j}{2N} \tag{2.5}$$

onde:

- $t(i,j)$ representa os valores tranformados;
- $s(m,n)$ representa os valores a serem transformados;
- N representa o tamanho do macrobloco
- $c(i,j)$ é dado por $c(i,j) = c(i)c(j)$;
- $c(k)$ é dado por $c(0) = \sqrt{1/N}, c(k) = \sqrt{2/N}$

A Equação 2.6 representa a transformada inversa (decodificação) obedecendo os mesmos parâmetros discutidos na Equação 2.5.

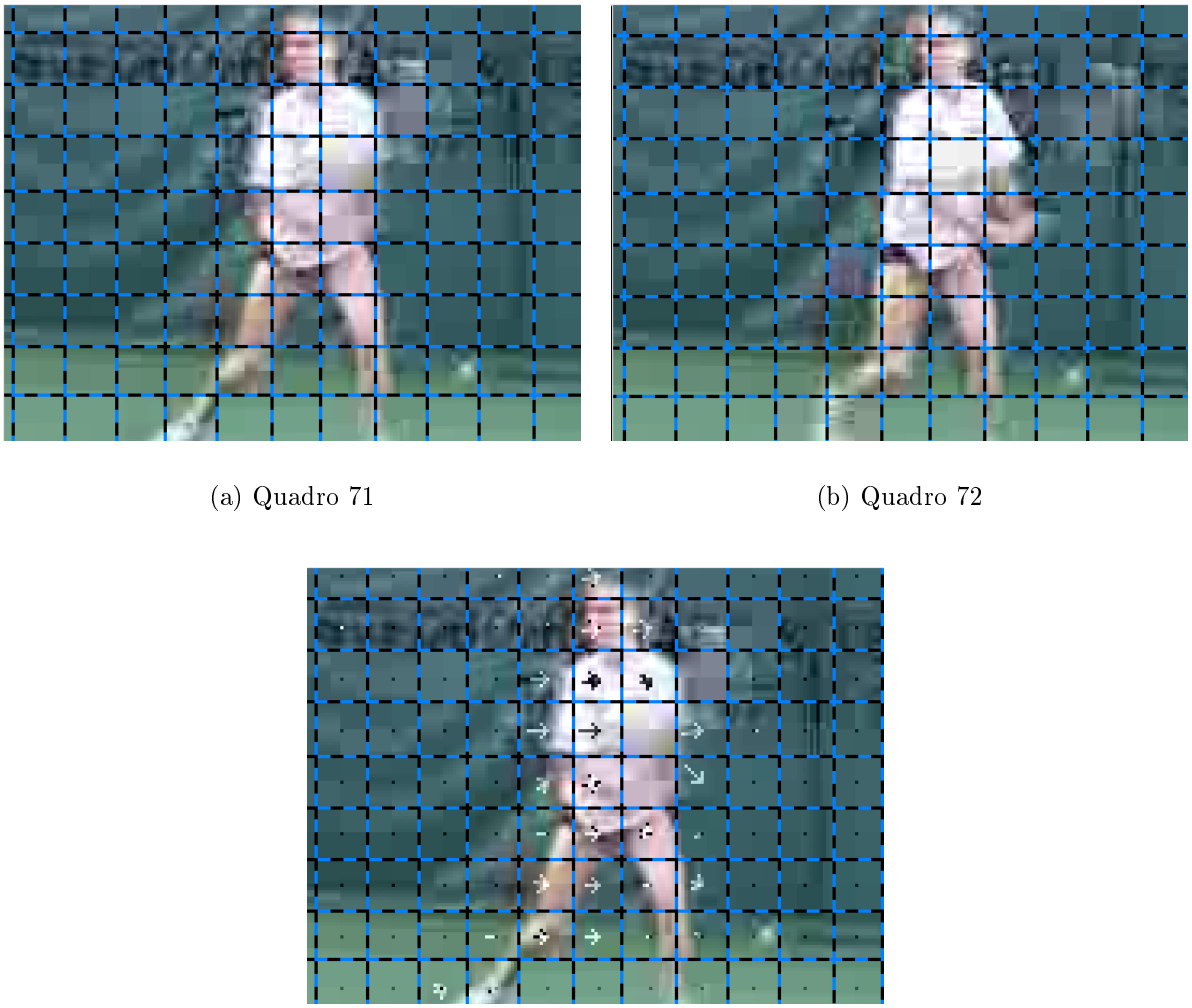
$$s(m,n) = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} c(i,j)t(i,j) \cos \frac{\pi(2m+1)i}{2N} \cos \frac{\pi(2n+1)j}{2N} \tag{2.6}$$

De forma similar ao que ocorre com a divisão espacial dos componentes YUV em macroblocos, ocorre para os vetores de deslocamento que são responsáveis pela compensação de movimento provendo a compressão temporal do padrão MPEG-2.

Cada vetor de deslocamento representa a disparidade entre macroblocos correspondentes (correnpondentes em x e em y) a cada dois quadros codificados. Como cada macrobloco pode assumir valores positivos ou negativos, assumir um ou outro dependerá do macrobloco anterior correspondente.

Um vetor de deslocamento positivo ou negativo indica o sentido do deslocamento (cima-baixou ou esquerda-direita) e sua magnitude expressa a quantidade de discrepância (deslocamento) detectada entre macroblocos correspondentes entre um quadro e outro.

Na Figura 2.1(a) tem-se o quadro com os macroblocos que servem de referência. Na Figura 2.1(b) tem-se o quadro com os macroblocos que foram concebidos a partir dos vetores de deslocamento representados na Figura 2.1(c).



(a) Quadro 71

(b) Quadro 72

(c) Vetores de deslocamento

Figura 2.1: Quadros explicativos sobre a concepção e utilização de vetores de deslocamento – Adaptado de [ZELNIK-MANOR and IRANI, 2001c]

Para cada macrobloco do componente de luminância, são codificados dois vetores

de deslocamento (x e y). Assim, os vetores de deslocamento são relativos a cada macrobloco codificado e produzem uma forma de acompanhar, temporalmente, os deslocamentos de cada macrobloco.

2.3 Eventos em vídeo

Eventos são objetos temporais que usualmente se extendem durante dezenas ou centenas de quadros. Um evento pode compreender objetos temporais quanto espaciais. Objetos espaciais são caracterizados por possuírem várias escalas espaciais (dentro do escopo do quadro ou imagem). De forma similar, objetos temporais são caracterizados por várias escalas temporais (dentro do escopo do tempo de duração).

Entretanto, existem grandes diferenças entre objetos espaciais e temporais. Devido a natureza perspectiva da projeção na dimensão espacial, um objeto espacial pode aparecer em diversas escalas espaciais em diferentes quadros.

De outra forma, um objeto temporal é sempre caracterizado pela mesma escala temporal, ou seja, tem a mesma duração no tempo, em toda as seqüências. Por exemplo, um simples passo de uma pessoa andando, visto de diferentes câmeras com a mesma taxa de QPS, se extenderá sobre o mesmo número de quadros em ambas as câmeras[ZELNIK-MANOR and IRANI, 2001a].

Levando-se em consideração tais características de que um evento possui, é uma forma coerente a tentativa de caracterizar difenrentes tipos de eventos através de sua escala temporal. Pois este é uma forma imparcial para análise de eventos quando se define um intervalo de análise temporal, tratado neste trabalho como “janela temporal”.

2.4 Trabalhos relacionados

Nesta seção serão apresentados alguns trabalhos relacionados à extração de características de deslocamento, além do emprego de medidas de divergência e aglomeração, que são os elementos principais no tema desta dissertação.

Em [ZELNIK-MANOR and IRANI, 2001a] é apresentada uma abordagem para classificação de eventos temporais que utiliza uma medida de divergência (D^2) baseada no teste estatístico qui-quadrado. Para o estabelecimento desta medida são consideradas janelas temporais, de tamanho pré-definido, onde se aplica um gradiente entre dois quadros da janela para se extrair informações de deslocamento.

Em cada gradiente é realizado um processo de normalização para preservar propriedades de orientação e tornar mínima a influência de propriedades fotométricas fornecidas pela magnitude do gradiente. Este processo é realizado em diferentes níveis de resolução, através da utilização de uma pirâmide gaussiana, para que, no resultado final da medida de divergência, somente pontos que tendem a ser consistentes possuam maior relevância no resultado final.

Obtendo-se pontos de divergência entre janelas temporais através do teste qui-quadrado os autores realizam um processo de agrupamento, utilizando um número fixo de classes, construindo uma matriz de similaridades (M) calculado da seguinte maneira:

$$M(i,j) = M(j,i) = e^{\frac{-D_{ij}^2}{\sigma}} \tag{2.7}$$

onde D_{ij} representa a divergência entre o evento i e j.

O parâmetro σ é um fator de escala das medidas de divergência. Um valor menor de σ tende a aproximar medidas de divergência com menor diferença de valores enquanto, um valor maior de σ tende a unir medidas de divergência com maior diferença de valores.

O algoritmo utilizado neste processo é o algoritmo "Normalized Cut" descrito em [SHI and MALIK, 2000]. A matriz de similaridades (M) representada pela Equação 2.7, é correspondente a matriz de adjacências (W) conforme referenciada no trabalho de [SHI and MALIK, 2000].

O modelo proposto pode ainda ser refinado quando se conhece, *a priori*, alguns dos eventos que estão sendo modelados. Este refinamento é realizado através de uma adaptação no teste de divergência tornando-o ponderado, através da associação de pesos, quando existe correlação entre a informação conhecida *a priori* e o evento sendo modelado no momento.

Shi e Malik apresentam em [SHI and MALIK, 2000] uma abordagem baseada em grafos para o problema de aglomeração em processos de segmentação de imagens. A

abordagem utiliza-se de um grafo (G) ponderado e não-direcionado. Cada nó do grafo (V) é um ponto do espaço de características. Uma aresta (E) é criada entre cada par de nós para a formação do grafo. Assim, é construída uma matriz de adjacências (W) que denota medidas de similaridade entre nós.

O problema de segmentação aplicado à teoria de grafos resume-se em encontrar uma solução discreta para uma função que particione o grafo em dois conjuntos distintos A e B. Esta função de corte pode ser denotada matematicamente como:

$$NCUT(A,B) = \frac{cut(A,B)}{assoc(A,V)} + \frac{cut(A,B)}{assoc(B,V)} \tag{2.8}$$

onde:

- $cut(A,B) = \sum_{u \in A, v \in B} (W(u,v))$
- $assoc(A,V) = \sum_{j} \sum_{i \in A} W(i,j)$
- $assoc(B,V) = \sum_{j} \sum_{i \in B} W(i,j)$

O particionamento do grafo é aquele que minimize o valor da Equação 2.8. Os autores sugerem que este problema pode ser formulado como um problema de autovetores. Para isto, é demonstrado que o autovetor associado ao segundo menor autovalor que resolve a Equação 2.9 pode ser utilizando na bipartição de um grafo.

$$D^{-\frac{1}{2}}(D-W)D^{-\frac{1}{2}}z = \lambda z \tag{2.9}$$

onde $D(i,i) = \sum_{j} W(i,j)$.

Quando da bipartição do grafo, se o mesmo procedimento for efetuado em cada sub-grafo, de forma recursiva, pode-se obter um número K de sub-grafos onde, cada sub-grafo, representa um aglomerado. Associando valores inteiros, iguais, para cada nó de um mesmo sub-grafo, obtêm-se uma solução discreta para o problema de aglomeração.

No trabalho de [CHOWDWDHURY and CHELLAPPA, 2003] os autores tratam de uma abordagem para reconhecimento de atividades baseado na trajetória de padrões repetitivos tratanto qualquer desvio significante de tal padrão de atividade.

Uma matriz de medidas 2D, que representa as características dos objetos monitorados no ambiente, é fatorada em uma composição 3D através da decomposição de valores singulares (SVD) definindo um espaço de formas 3D representando os movimentos comuns para o ambiente estudado.

Obtendo os dados de cada movimento sendo analisado e plotando-o neste espaço de formas, é possível observar o desvio dos movimentos comuns, através de uma trajetória média, realizando o reconhecimento como um movimento conhecido ou não.

Em [RAMANAN and FORSYTH, 2003] é apresentado um mecanismo para detecção e rastreamento de pessoas baseado em modelos de aparência. São utilizados nove segmentos do corpo humano sendo: abdômen, braços, ante-braços, pernas (parte superior e inferior).

Uma abordagem baseada em modelos de aparência pode trazer como benefício a maior facilidade em se recuperar a identificação de pessoas em momentos de oclusão, haja visto que a aparência das pessoas não se modifica de forma expressiva quadro a quadro.

O método proposto primeiramente realiza uma etapa de aprendizado dos modelos de aparência, onde os possíveis candidatos em um determinado quadro são marcados. Esta marcação ou modelagem é realizada através da segmentação por cilindros (ajustados conforme a parte do corpo que se está buscando segmentar). Com isso, cria-se um vetor de características (utilizando um histograma normalizado no espaço $L * u * v$[1]) para cada segmento candidato.

As características de cada segmento são aplicadas em um classificador. Os grupos que possuem um modelo muito esparso ou que não respeitam determinadas convenções cinemáticas (movimento por exemplo) são descartados.

2.5 Discussão

O trabalho [ZELNIK-MANOR and IRANI, 2001a] apresenta uma abordagem não-supervisionada para o problema de detecção de eventos através da utilização de informações de gradiente aplicado entre duas imagens. Apresenta como pontos fortes, uma métrica de divergência temporal baseada em um teste de divergência estatístico juntamente com uma função de normalização para minimizar a influência de propriedades fotométricas.

Porém, a utilização do gradiente como forma de obtenção de valores de deslocamento de objetos entre duas imagens e sua análise em várias escalas através de pirâmides gaussianas torna o tempo de processamento e armazenamento das informações em memória elevado em relação a outras técnicas.

A abordagem de aglomeração baseada em grafos de [SHI and MALIK, 2000] demonstra ser uma abordagem eficaz, por considerar tanto informações de similaridade entre os diferentes grupos (global) assim como intra-grupos (local).

[1] Espaço de cor que separa componentes de luminância e crominância

Na prática observou-se que este método tem como ponto fraco questões relacionadas ao tempo de execução. Dependendo da quantidade de medidas que venham a formar o grafo, o cálculo de autovetores pode se tornar bastante moroso.

Além disso, o processo de criação da matriz de adjacências (W) é computacionalmente menos eficiente quando comparado com outras técnicas de aglomeração não baseadas em grafo. Isto deve-se ao fato de ser necessário estabelecer arestas (E) entre todos os nós (V) do grafo.

Já em [CHOWDWDHURY and CHELLAPPA, 2003], o método proposto traz o inconveniente do processo de aprendizado a ser realizado. Dependendo da dinâmica das trajetórias, o processo de aprendizado pode não ser suficiente necessitando da combinação com técnicas não supervisionadas para o reconhecimento de trajetórias. Porém, para ambientes bastante restritos quanto a execução de eventos em trajetórias são bem definidas, o método demonstra ser eficaz.

No próximo capítulo, será discutido uma forma de se avaliar eventos temporais através de diferentes métricas de divergência, utilizando características de vetores de deslocamento, combinadas com diferentes algoritmos de aglomeração para isolar e agrupar eventos dentro de uma seqüência de vídeo.

Capítulo 3

Método Proposto

A proposta deste trabalho de dissertação é avaliar diferentes métricas de divergência e aglomeração para a classificação de eventos utilizando informações de vetores de deslocamento obtidos de um vídeo codificado no padrão MPEG-2.

Para isso, serão avaliadas cinco métricas estatísticas que utilizarão informações de vetores de deslocamento, de forma janelada, para gerar uma medida de divergência entre eventos.

Como pode ser visto na Seção 2.2 do Capítulo 2, em um vídeo codificado no padrão MPEG-2, existem vetores de deslocamento, em x e y, referente a cada macrobloco. Dessa maneira, obtêm-se duas matrizes (VDx – Vetores de deslocamento em x e VDy – Vetores de deslocamento em y) de $\frac{1}{16}$ do tamanho original dos quadros codificados.

A Figura 3.1 retrata bem este caso. Um vídeo é submetido a um CODEC (*COdificador/DECodificador*) MPEG-2 que tem por finalidade interpretar o arquivo codificado e disponibilizar coeficientes DCT, vetores de deslocamento e outros elementos do padrão de forma que seja possível manipulá-los.

Após isso, é realizado um procedimento de extração/leitura (dentre todas as informações disponibilizadas pelo CODEC) somente dos vetores de deslocamento, que é o elemento onde a parte inicial da metodologia, as métricas de divergência ($Div.1$, $Div.2$, ... , $Div.N$), irão atuar.

3.1 Especificações técnicas

Todos os testes foram feitos em um sistema implementado em linguagem C/C++, utilizando o compilador gcc e g++ da GNU[1]. Foram utilizadas as bibliotecas LtiLib[2] (bi-

[1] http://gcc.gnu.org
[2] http://ltilib.sourceforge.net

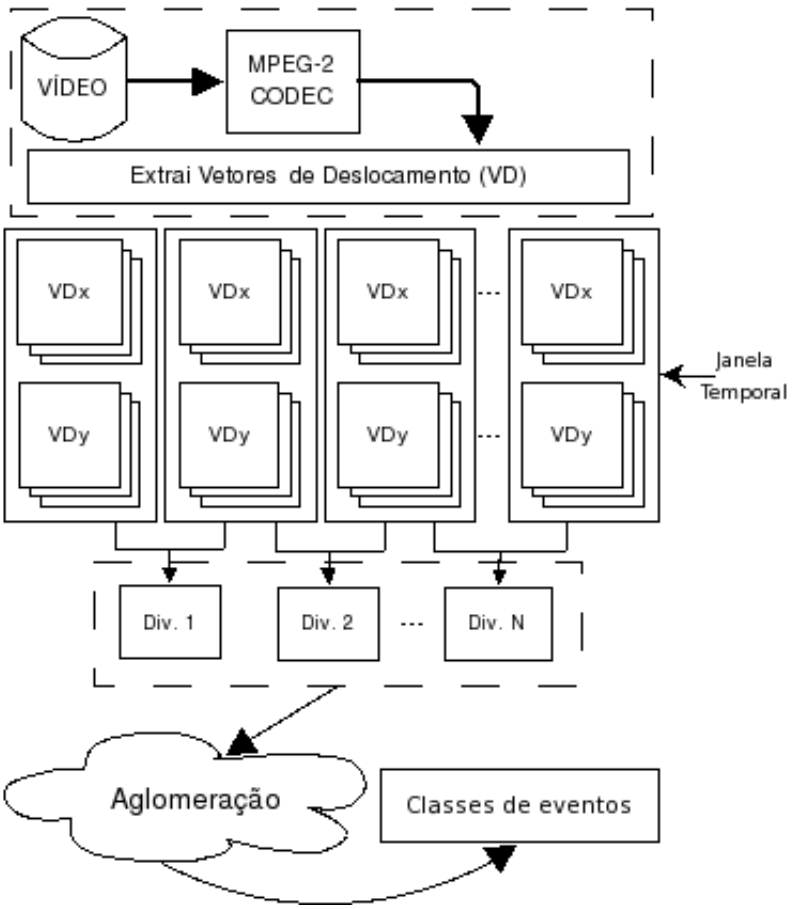


Figura 3.1: Diagrama de blocos do extrator de características. Vetores de deslocamento são extraídos para coordenadas x e y para cada janela temporal.

blioteca com rotinas para álgebra linear, processamento de imagens e outros) e FFMPEG[3] (biblioteca utilizada para leitura e interpretação dos vídeos). Para a edição de vídeos foi utilizado o utilitário Avidemux2[4] e para a edição de imagens, o utilitário GIMP[5].

Os experimentos foram executados em um microcomputador Intel Celeron de 1.1 GHz com 256Mb de memória RAM tendo como sistema operacional uma distribuição GNU/Linux. O tempo médio de processamento, para um vídeo com duração de 6 minutos e 24 segundos codificado com 25 QPS, foi de 2 minutos e 50 segundos para a extração dos vetores de movimento, utilizando uma janela de 64 quadros, e realizando o processo de aglomeração baseado em grafo ("NCUT").

[3]http://ffmpeg.sourceforge.net
[4]http://fixounet.free.fr/avidemux/
[5]http://www.gimp.org

3.2 Extrator de características

Na Figura 3.2, pode-se observar, com maiores detalhes, como as características de baixo nível são extraídas, para cada janela temporal, a partir dos vetores de deslocamento. O conjunto final de características, de cada janela, é a soma da magnitude de deslocamento obtido em x e y. Assim, as características tornam-se invariantes ao sentido (direta-esquerda – x ou cima-baixo – y) de ação.

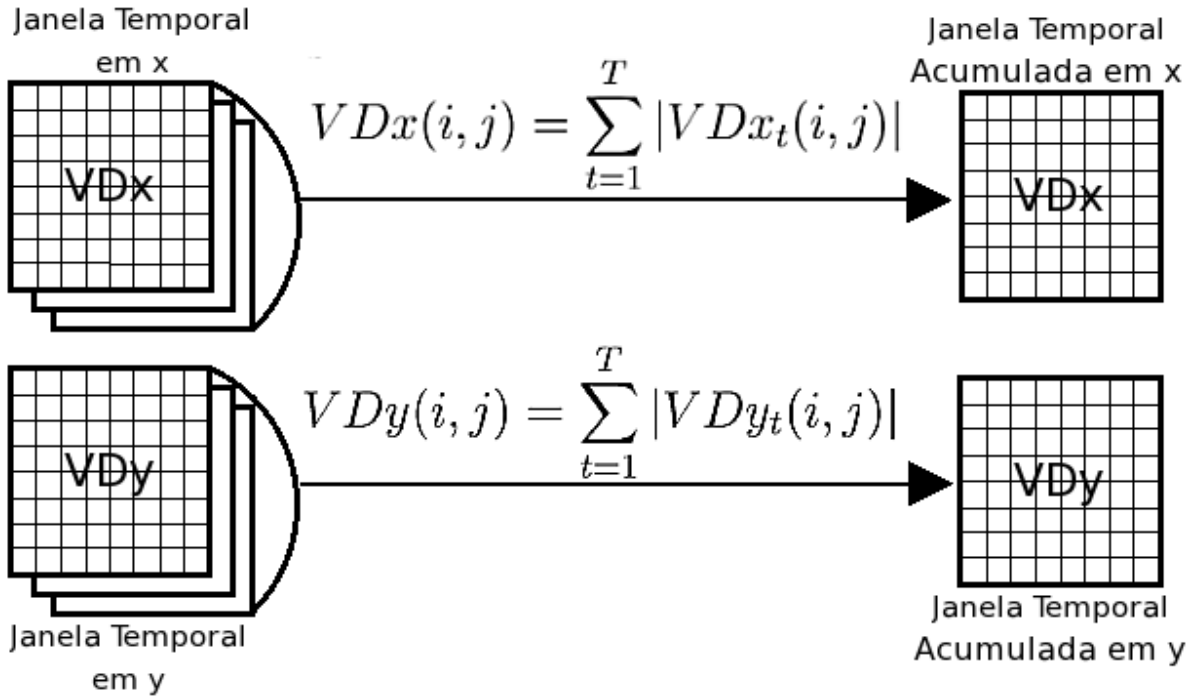


Figura 3.2: Diagrama que demonstra as características extraídas para cada janela temporal (T).

Neste ponto, já se possui as características necessárias para a aplicação de métricas de divergência. Todas as métricas tratadas neste trabalho baseiam-se na análise das características extraídas através de histogramas. A utilização de histogramas em tal processo é uma maneira bastante empregada para obter/calcular similaridades [ANTANI et al., 2002].

A avaliação dos histogramas em questão se dá através de um teste de comparação. Alguns desses testes empregados em histogramas, fazem parte desta metodologia como método de avaliação e descritos neste capítulo.

A utilização de histogramas baseados na freqüência relativa de suas classes fornecem resultados similares para eventos com grande correlação em extensão temporal ou espacial (escala) na sua execução[ZELNIK-MANOR and IRANI, 2001a].

Como as características de deslocamento são fornecidas de maneira independente (em x e em y), quando da extração do vídeo codificado no padrão MPEG-2, buscou-se

avaliar as técnicas de divergência utilizando desta informação de deslocamento de duas maneiras:

- Utilizando os vetores de deslocamento de forma disjunta (VDx e VDy) (Aplicando a métrica de divergência de forma independente para VDx e VDy);
- Utilizando o vetor resultante de x e y (VDR), obtido através da Equação 3.1.

$$VDR(i,j) = \sqrt{|VDx(i,j)^2| + |VDy(i,j)^2|} \tag{3.1}$$

Assim, caso utilizem-se os vetores de deslocamento de forma disjunta, serão calculados dois histogramas (H) para cada janela temporal (quatro para aplicar em uma medida de divergência, por ser necessário dois eventos). Já no outro caso, utilizando o vetor de deslocamento resultante (Equação 3.1), será necessário um histograma apenas para cada evento (dois para aplicar medida de divergência).

Com um conjunto de medidas de divergência, aplica-se em um algoritmo de aglomeração para encontrar grupos de eventos. Porém, para uma medida de divergência compor o conjunto de características que serão aplicadas nos algoritmos de aglomeração, ela deve ser submetida a Equação 3.2:

$$D_{final} = e^{\frac{-D_{ij}}{\sigma}} \tag{3.2}$$

onde i e j representam as características dos eventos e D representa a divergência entre i e j.

A utilização da exponencial tem por finalidade especificar que quanto menor a divergência entre dois eventos, mais correlacionados estarão tais eventos no algoritmo de aglomeração. Isto é exatamente o que se busca com o processo de aglomeração, manter em grupos eventos altamente correlacionados e separar em classes os eventos menos correlacionados.

Assim, o resultado da Equação 3.2 será o valor efetivo que irá compor o conjunto de características final, submetido aos algoritmos de aglomeração avaliados neste trabalho.

3.3 Medidas de divergência

Neste trabalho serão avalidas cinco medidas de divergência, todas elas baseadas na utilização de histogramas gerados a partir de vetores de deslocamento. Os valores das classes dos histogramas utilizados foram tratados na forma de freqüência relativa pelos motivos já descritos na Seção 3.2 deste mesmo capítulo.

Resalta-se que, as características fornecidas pelos vetores de deslocamento podem ser tratadas de maneira independente (em VDx e VDy) ou utilizando-se o vetor resultante (VDR) através da Equação 3.1. Assim cada medida de divergência pode ser aplicada de duas maneiras sobre as características fornecidas pelos vetores de deslocamento.

As medidas de divergência utilizadas neste trabalho e que serviram de base para a geração de resultados através de experimentos foram as seguintes:

- Qui-quadrado (D_1): representado pela Equação 3.3;
- Diferença de histogramas (D_2): o mais simples, representado pela Equação 3.4, esta medida retorna o somatório das diferenças em valores absolutos;
- Interseção de histogramas (D_3): representado pela Equação 3.5;
- Kolmogorov-Smirnov (D_4): representado pela Equação 3.6;
- Kuiper (D_5): representado pela Equação 3.7.

$$D_1 = \sum_{i=1}^{n} \frac{(H_1(i) - H_2(i))^2}{(H_1(i) + H_2(i))} \tag{3.3}$$

$$D_2 = \sum_{i=1}^{n} |H_1(i) - H_2(i)| \tag{3.4}$$

$$D_3 = \frac{\sum_{i=1}^{n} min(H_1(i), H_2(i))}{\sum_{i=1}^{n} H_2(i)} \tag{3.5}$$

$$D_4 = \max_i |CH_2(i) - CH_1(i)| \tag{3.6}$$

$$D_5 = \max_i |CH_2(i) - CH_1(i)| + \max_i |CH_1(i) - CH_2(i)| \tag{3.7}$$

onde H_1, H_2, CH_1 e CH_2 são, respectivamente, os histogramas dos eventos observado, esperado, observado acumulado e esperado acumulado.

Valores baixos de Dx, onde x indica o número do teste de divergência, indicam baixa divergência (alta correspondência/similaridade entre os eventos).

Obtendo-se uma métrica de divergência para eventos em vídeo, conclui-se a primeira etapa da metodologia. A segunda parte fica dedicada à aplicação e análise de algoritmos de aglomeração que é relatada na Seção 3.4 neste mesmo capítulo.

3.4 Aglomeração

Aglomeração é a classificação não-supervisionada de padrões (observações, dados ou vetores de características) em grupos. O problema de agrupamento foi tratado em muitos contextos e por muitos pesquisadores. Isso demonstra seu grande apelo e utilidade como um dos passos na análise de dados. Algumas das principais aplicações de algoritmos de agrupamento são: segmentação de imagens, reconhecimento de objetos e recuperação de informação[JAIN et al., 1999].

Uma abordagem dita aglomerativa inicia com cada padrão em um grupo distinto sendo que a cada iteração do algoritmo, grupos definidos como próximos são combinados até que um critério de parada seja satisfeito. Pode-se incluir nesta abordagem os algoritmos "K-Médias" e "ISODATA".

Já a abordagem divisória, inicia com todos os padrões em um único grupo e inicia-se um processo de separação até que um critério de parada seja satisfeito. Enquadra-se nesta abordagem o algoritmo "Normalized Cut".

A aplicação de um processo de aglomeração neste trabalho tem como objetivo principal isolar grupos de eventos em uma seqüência de vídeo. Outro objetivo é analisar, dentro de um grupo de algoritmos conhecidos na área de Reconhecimento de Padrões e Visão por Computador, se existe alguma correlação ótima entre uma determinada métrica de divergência com um determinado algoritmo de aglomeração para classificação de eventos baseado em característcas fornecidas por vetores de deslocamento.

3.4.1 Algoritmo "K-Médias"

O algoritmo "K-Médias" é um dos algoritmos de aglomeração mais utilizados para a realização de agrupamentos em conjunto de características. É também chamado de algoritmo das "médias móveis" porque em cada iteração são recalculados os centros dos agrupamentos.

O algoritmo consiste em, dado um número K de centros desejáveis, onde estes K centros são escolhidos arbitrariamente na inicialização, calcular os melhores K centros no conjunto de características, de forma que os centros fiquem distribuídos de forma homogênea em seus grupos tornando-se elementos centrais para cada grupo em relação ao conjunto de dados.

Este algoritmo pode ser resumido nos seguinte passos:

1. **Inicialização**: consiste em inicializar, arbitrariamente, os centros do K grupos desejados;

2. **Associação e atualização dos centros**: neste passo, cada característica é associada ao grupo mais próximo e se recalculam os centros a partir das médias das características pertencentes a cada grupo;

3. **Convergência**: no passo anterior algumas características podem trocar de agrupamento e, por conseqüência, os centros destes serão alterados. Se isto ocorrer, deve-se repetir o passo 2 até que os centros se tornem estáveis não se modificando mais. Quando não houver mais modificações considera-se que foi encontrado um bom particionamento e finaliza-se a aglomeração.

Este algoritmo é bastante difundido e utilizado por sua simplicidade. Um inconveniente é que o resultado final depende do valor de K. Além disso, existe um problema relacionado com a inicialização dos centros. Como os centros podem ser escolhidos aleatoriamente, se estes não são bem selecionados de maneira que estejam bem distribuídos sobre o espaço de características, o processo de convergência do algoritmo tende a ser menos eficiente.

3.4.2 Algoritmo "ISODATA"

O algoritmo ISODATA (*Iterative Self-Organizing Data Analysis Techniques*) é um algoritmo até certo ponto similar ao algoritmo "K-Médias". Isto por que o algoritmo "ISODATA" processa, de forma repetitiva, as características que se busca aglomerar e, a cada iteração, associa-se uma dada característica com o grupo mais próximo rearranjando os centros de cada grupo.

Porém este algoritmo, diferentemente do algoritmo "K-Médias", não necessita conhecer previamente o número K de padrões/grupos que se busca classificar. O algoritmo "ISODATA" incorpora procedimentos que buscam eliminar agrupamentos pouco numerosos, mesclar agrupamentos próximos e dividir agrupamentos dispersos para inferir, a partir das características fornecidas, um número de grupos que seja ideal para o conjunto de dados.

O algoritmo "ISODATA" pode ser resumido nos seguinte passos:

Sendo f_i, $i = 1, \ldots, N$, o vetor de características, então:

1. **Inicialização**: inicializar o número de grupos desejados inicialmente, g, onde $g \leq N$;

2. **Associação de centros**: gerar arbitrariamente g conjuntos de características médias, $\mu_j, j = 1, \ldots, g$;

3. **Classificação**: classificar todos f_i associando com a média μ_j mais próxima;

4. **Recalcular médias**: recalcular as médias de cada grupo ($\mu = (\frac{1}{k_j}\sum_j g_j)$ — k_j, é número de elementos do grupo j.);

5. **Se**: algum μ_j mudou no passo 4, voltar para o passo 3;

6. **Calcular**: $M_j = min|\mu_{j_1} - \mu_{j_2}|$, para $j_1 \neq j_2$

7. **Se**: $M < L$, onde L é um limiar, então existem $g - 1$ grupos finais. Parar.

8. **Senão**: armazene os valores anteriores de $k_j, j = 1, g$

9. **Incrementar** g e voltar para o passo 2.

Este algoritmo tem como características marcante, e a maior vantagem, de não ser necessário conhecer previamente a provável distribuição de grupos (K) nas características fornecidas ao classificador.

3.4.3 Algoritmo "Normalized Cut"

Este algoritmo, baseado na idéia de tratar as características de forma interligada formando um grafo ponderado, faz parte de uma categoria de técnicas de classificação denominada "Classificação Espectral". Este termo refere-se ao fato do algoritmo não utilizar as características da mesma maneira, no mesmo espaço, que são fornecidas como entrada durante o processo de classificação efetivo.

A classificação das características em grupos é realizada, efetivamente, com base nos autovetores resultantes de um processo de transformação linear. Esta transformação é uma forma de aproximar uma solução para o processo de encontrar nós (valores) em grafos, cujo espaço de busca é exponencial, podendo assim, ser um processo muito demorado podendo não haver solução em tempo hábil.

Dessa forma, este processo de transformação linear fornece uma maneira de se resolver o problema de separação/corte em grafos em tempo polinomial.

O algoritmo para esta técnica pode ser resumido pelos seguinte passos:

Dada uma matriz de similaridades W:

1. **Calcular**: matriz degrau $D_{ii} = \sum_j W(i,j)$

2. **Tranformação**: obter a transformação linear de W, $A = D^{-\frac{1}{2}}(D - W)D^{-\frac{1}{2}}$;

3. **Obter**: o autovetor associado ao segundo menor autovalor de A;

4. **Dividir**: o grafo representado por W através da aplicação de um limiar contra o autovetor obtido no passo 3;

5. **Se**: o número de grupos desejado for 2, parar. Senão, voltar ao passo 1 e aplicar o método, em cada sub-grafo, de forma recursiva.

Esta técnica de aglomeração, na mesma forma que o algoritmo "K-Médias", necessita conhecer previamente o valor de grupos K que se busca encontrar. Este número irá definir o número de recursões a serem efetuadas pelo algoritmo.

Em [FORSYTH and PONCE, 2003], o autores comentam que um procedimento que demonstra funcionar na prática e que pode tornar mais simples a elaboração do algoritmo é, ao invés de tratá-lo como um procedimento recursivo, utilizar os autovalores subseqüentes (terceiro, quarto, etc — associado ao terceiro, quarto, etc menor autovalor) conforme o número de grupos (K) que se busca.

De acordo com [ANTANI et al., 2002] as métricas de divergência propostas neste trabalho, como uma forma de se medir disparidades entre eventos, fazem parte de um conjunto de medidas bem conhecidas e utilizadas para medir similaridades.

Os algoritmos de aglomeração "K-Médias" e "ISODATA" fazem parte de uma categoria denominada "aglomerativa" e se tratam de aglomeradores bastante difundidos na área de Reconhecimento de Padrões. Sua complexidade é $O(n)$ onde n representa o número de padrões (dados)[JAIN et al., 1999].

Já o algoritmo "Normalized Cut" faz parte de uma categoria denominada "divisória" onde um conjunto de padrões, inicialmente interligados, são divididos sucessivamente. Este algoritmo demonstra ser eficaz para o problema de bipartição de grafos em tempo polinomial.

A partir disso, propõe-se combinar as diferentes medidas de divergências com os diferentes algoritmos de aglomeração juntamente com a utilização de características de forma independe (VDI (*Vetor de Deslocamento Independente (x,y)*)) e resultante (VDR (*Vetor de Deslocamento Resultante*)) a fim de se procurar a(s) melhores e pior(es) correlações para um conjunto de três vídeos que representam a base de testes empregada.

O próximo capítulo enfatiza o procedimento de experimentos realizados a partir do proposto método a fim de mensurar as diferentes combinações entre medidas de divergência, algoritmos de aglomeração e variação dos diferentes parâmetros comentados até então.

Capítulo 4

Experimentos

Os experimentos foram realizados através de uma combinação, para cada vídeo que compõe a base, entre medidas de divergência e algoritmos de aglomeração, ambos discutidos no Capítulo 3. O protoloco obedecido para a realização e combinação das diferentes técnicas, os resultados obtidos além das bases de dados utilizadas são discutidos neste capítulo.

4.1 Protocolo de testes

Como as características de movimento obtidas através dos vetores de deslocamento podem ser utilizadas de forma independente (VDx e em VDy) ou combinados (VDR), gerando um vetor de deslocamento resultante, tem-se uma combinação das duas formas de tratamento das características de movimento com cinco medidas de divergência e três algoritmos de aglomeração.

Além desse fato, foi executado, para cada combinação de características $\times$ medida de divergência $\times$ aglomeração, uma variação de σ que compõe a divergência final entre dois eventos representada pela Equação 3.2.

Esta variação compreende o intervalo $[0,05; \ldots; 1,0]$, com incremento de $0,05$; totalizando assim 20 variações de σ.

Para realizar o processo de verificação de resultados, foi utilizado o processo *ground-truth* que pode ser traduzido, em português, como abordagem de referência ou ainda abordagem verdadeira.

Este processo consiste em comparar resultados considerados ideais, que podem ser obtidos através de rotulagem visual/manual dos vídeos, separando-o em classes de eventos, com os resultados obtidos pela combinação: tipo de característica (VDI/VDR) $\times$ medida de divergência $\times$ algoritmo de aglomeração.

Para cada base (vídeo) foi realizado um processo de rotulagem visual, observando a ocorrência de todos os eventos e separando-os em classes. A quantidade de classes, pré-estabelecidas, para a aplicação dos testes, obedeceu o seguinte critério:

1. Observação real do número de classes;
2. Observação em duas classes, no item 1, associando os eventos mais similares (visualmente).

Este critério de observação do número de classes em 2 e K classes foi realizado no sentido de se poder analisar o comportamento dos algoritmos de aglomeração, na classificação real do número de eventos e na classificação dos eventos que menos se assemelham em relação a todos os eventos tratados (por este motivo tem-se 2 classes).

Este procedimento pode ser empregado, ainda, na tentativa de relacionar/separar eventos convencionais de não-convencionais (duas classes).

Os resultados são medidos de forma quantitativa através de duas medidas: Precisão e Revocação [llas GARGI et al., 2000]. Define-se λ como Precisão e φ como Revocação, respectivamente, através das Equações 4.1 e 4.2:

$$\lambda = \frac{DT}{DT + DE} \tag{4.1}$$

$$\varphi = \frac{DT}{DT + AF} \tag{4.2}$$

onde DT é a quantidade de rotulações corretas para uma dada classe, DE é a quantidade de detecções errôneas e AF representa os alarmes falsos.

A medida de Revocação, indica a quantidade de acertos para uma determinada classe. Já a medida de Precisão indica o escopo em que a medida de Revocação está atuando. O ideal é se ter alta Precisão e alta Revocação.

Uma medida de baixa Revocação e alta Precisão indica que o sistema acertou pouco, mas houve pouca confusão na classe em questão. Já o inverso, alta Revocação e baixa Precisão indica que o sistema acertou bastante mas que o espaço de acerto foi maior do que o ideal.

Os resultados discutidos neste trabalho e evidenciados através de gráficos, serão relacionados às melhores combinações e piores valores de σ na Equação 3.2 (utilizado para compor a divergência final que irá valer como entrada para um determinado algoritmo de aglomeração).

Os resultados serão observados através de gráficos representando medidas de Precisão e Revocação, para um dado valor de σ, e para uma combinação de VDR/VDI, medidas de divergência e algoritmos de aglomeração.

4.2 Bases de dados

O processo de seleção do vídeos que compõem a base de dados buscou levar em consideração a maior variabilidade de tipos de eventos, assim como sua periodicidade e extensão temporal. Foi utilizado, dessa maneira, três vídeos cada um contendo diversos tipos de eventos.

Tabela 4.1: Especificações dos vídeos que compõem a base de dados

Vídeo	"Movimentos"	"Tênis"	"Inria"
QPS	25	30	25
Resolução	190×480	330×418	288×384
Duração	6m 24s	0m 16s	2m 46s
Quadros	9601	487	4168
VD	12×30	21×27	18×24
Janela	64	12	64
Eventos	5	3	4

4.2.1 Vídeo "Movimentos"

A base de dados representada pelo vídeo "Movimentos" contém basicamente eventos que são executados em sentido longitudinal e alguns poucos de forma diagonal. Este vídeo foi, conforme a Tabela 4.1, rotulado em 5 classes de eventos ("andar", "acenar", "correr", "rolar" e "mover no lugar").

Pelo fato de se tratarem de eventos com maior extensão temporal, o tamanho da janela temporal que extrai as características de movimento foi definida em 64 quadros (quadros estes referentes a vetores de deslocamento e não imagem propriamente dita).

Nas Figuras 4.1 a 4.4, são demonstrados alguns quadros contidos na base. Não foi demonstrado o evento "mover no lugar" por ser de difícil representação em um ambiente estático como este.

Pode-se notar que os eventos ocorrem sempre com apenas uma pessoa em cena não existindo, dessa forma, a possibilidade de dois eventos (iguais ou mesmo diferentes) ocorrendo ao mesmo tempo.

(a) Quadro 3679 — “rolar”

(b) Quadro 3684 — “rolar”

(c) Quadro 3689 — “rolar”

Figura 4.1: Exemplo de quadros extraídos do vídeo “Movimentos” representando o evento “rolar”

(a) Quadro 5779 — “andar”

(b) Quadro 5784 — “andar”

(c) Quadro 5789 — “andar”

Figura 4.2: Exemplo de quadros extraídos do vídeo “Movimentos” representando o evento “andar”

(a) Quadro 6739 — “acenar”

(b) Quadro 6744 — “acenar”

(c) Quadro 6749 — “acenar”

Figura 4.3: Exemplo de quadros extraídos do vídeo “Movimentos” representando o evento “acenar”

(a) Quadro 9551 — “correr”

(b) Quadro 9556 — “correr”

(c) Quadro 9561 — “correr”

Figura 4.4: Exemplo de quadros extraídos do vídeo “Movimentos” representando o evento “correr”

4.2.2 Vídeo "Tênis"

Esta base de dados, representa eventos executados com menor extensão temporal. A dinâmica com que os eventos são executados difere grandemente da dinâmica das outras duas bases ("Movimentos" e "Inria").

Este fato, dos eventos possuírem menor extensão temporal, caracterizando-os como eventos de maior dinâmica que das demais bases, fez com que o tamanho da janela temporal para extração de característas tenha sido definido com um valor menor, como pode ser conferido na Tabela 4.1.

O número de eventos caracterizado para esta base foram três: "batida", "cadência" e "salto". Uma demonstração de como estes eventos são executados é ilustrado pelas Figuras 4.5 a 4.7.

(a) Quadro 150 — "batida"

(b) Quadro 155 — "batida"

(c) Quadro 160 — "batida"

Figura 4.5: Exemplo de quadros extraídos do vídeo "Tênis" representando o evento "batida"

(a) Quadro 233 — “salto”

(b) Quadro 238 — “salto”

(c) Quadro 243 — “salto”

Figura 4.6: Exemplo de quadros extraídos do vídeo “Tênis” representando o evento “salto”

(a) Quadro 309 — “cadência”

(b) Quadro 314 — “cadência”

(c) Quadro 319 — “cadência”

Figura 4.7: Exemplo de quadros extraídos do vídeo “Tênis” representando o evento “cadência”

4.2.3 Vídeo "Inria"

Este vídeo representa um ambiente monitorado por uma câmera com lente angular. Este tipo de lente consegue retratar o ambiente com maior amplitude do campo visual. Os eventos ocorridos nesta base foram divididos em 4 classes, a saber:

- "diagonal": são eventos que ocorrem de forma diagonal. Podem, ou não, compreender o ambiente inteiro, ou seja, o evento pode ocorrer de uma ponta à outra no ambiente ou apenas parcialmente.
- "encontro": é um tipo de evento que ocorre da execução de dois outros (dois eventos diagonais).
- "horizontal": evento que ocorre de forma horizontal, podendo ser parcial ou completo.
- "vertical": evento que ocorre de forma vertical, podendo ser parcial ou completo.

Nas Figuras 4.8 a 4.11 tem-se quadros demonstrativos das classes de eventos para o vídeo em questão.

(a) Quadro 296 — “diagonal”

(b) Quadro 301 — “diagonal”

(c) Quadro 306 — “diagonal”

Figura 4.8: Exemplo de quadros extraídos do vídeo “Inria” representando o evento “diagonal”

(a) Quadro 1999 — “encontro”

(b) Quadro 2004 — “encontro”

(c) Quadro 2009 — “encontro”

Figura 4.9: Exemplo de quadros extraídos do vídeo “Inria” representando o evento “encontro”

(a) Quadro 2796 — “horizontal”

(b) Quadro 2801 — “horizontal”

(c) Quadro 2806 — “horizontal”

Figura 4.10: Exemplo de quadros extraídos do vídeo “Inria” representando o evento “horizontal”

(a) Quadro 3038 — “vertical”

(b) Quadro 3043 — “vertical”

(c) Quadro 3048 — “vertical”

Figura 4.11: Exemplo de quadros extraídos do vídeo “Inria” representando o evento “vertical”

4.3 Experimento com o vídeo “Movimentos”

Este vídeo possui 5 classes de eventos conforme descrito na Seção 4.2. Os procedimentos para avaliação foram executados primeiro utilizando o processo *ground-truth* com 5 classes e comparando os resultados obtidos do processo de aglomeração com a rotulagem manual.

Os resultados para este vídeo são demonstrados para o melhor e pior valor de σ para a medida de Precisão e a corresponde taxa de Revocação para o mesmo valor de σ.

Na Subseção 4.3.1 tem-se a análise desses resultados levando-se em consideração 5 classes e na Subseção 4.3.2 a análise referente ao processo *ground-truth* para 2 classes.

4.3.1 Resultados para 5 classes

Analisando o gráfico representado pela Figura 4.12, demonstra-se que o algoritmo de aglomeração "ISODATA", possui uma taxa de Precisão bastante inferior, quando utilizando características dos vetores de deslocamento de forma resultante (VDR), para as medidas de divergência D_1 e D_4 , quando comparado com os outros algoritmos de aglomeração. Quando da utilização do algoritmo "ISODATA" com vetores de deslocamento independentes (VDI) somente as medida de divergência D_4 e D_5 demonstraram-se prover resultados próximos à média.

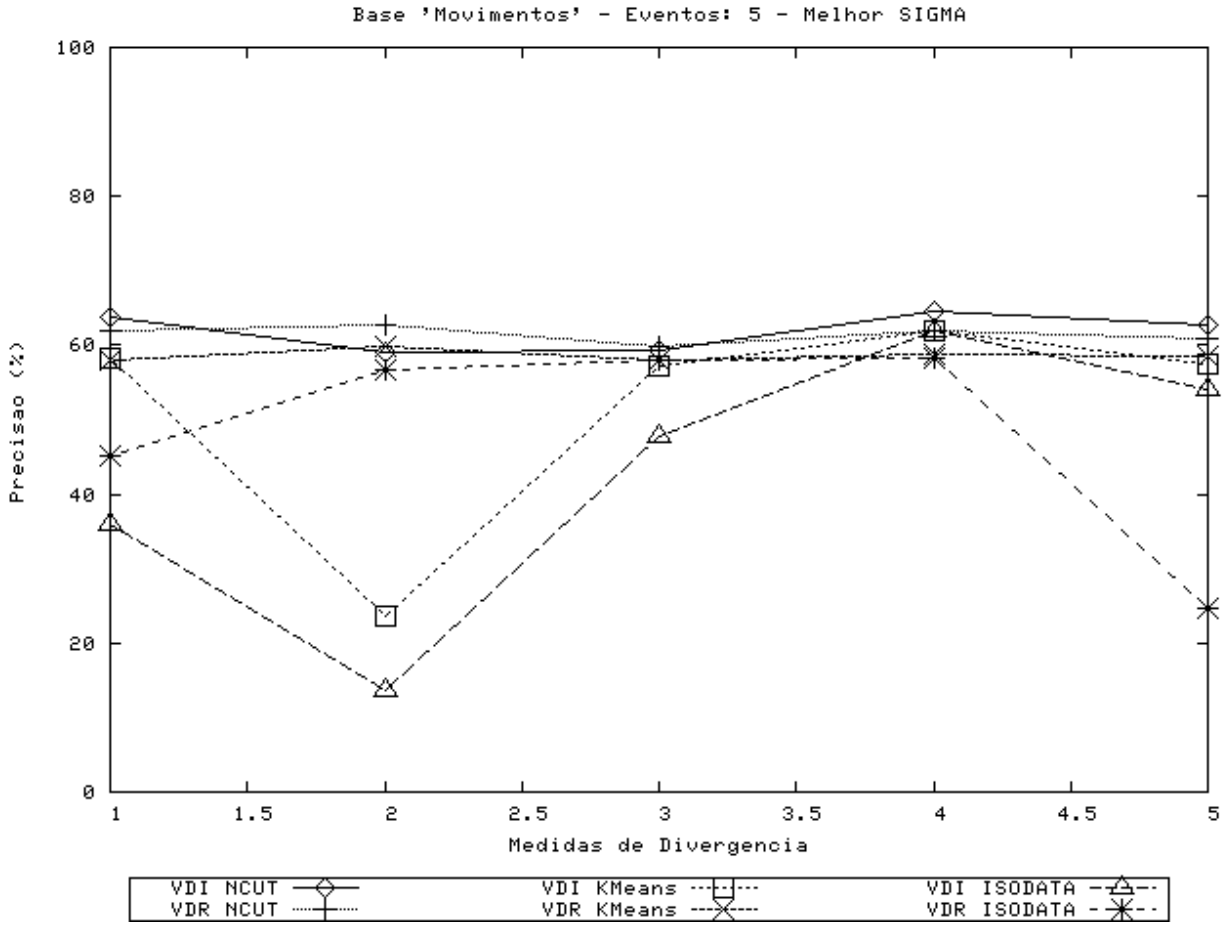


Figura 4.12: Resultados de Precisão para vídeo "Movimentos", 5 classes, melhor caso, $\sigma = 0,40$.

Quando associa-se a gráfico de Revocação, demonstrado pela Figura 4.13, pode-se chegar a uma melhor conclusão a respeito dos métodos.

O algoritmo de aglomeração "Normalized Cut" demonstra possuir a melhor taxa de Revocação média (entre as medidas de divergência), exceto para as medidas de divergência D_2 e D_3 como pode ser visto na Figura 4.13. Assim, o algoritmo "Normalized Cut"

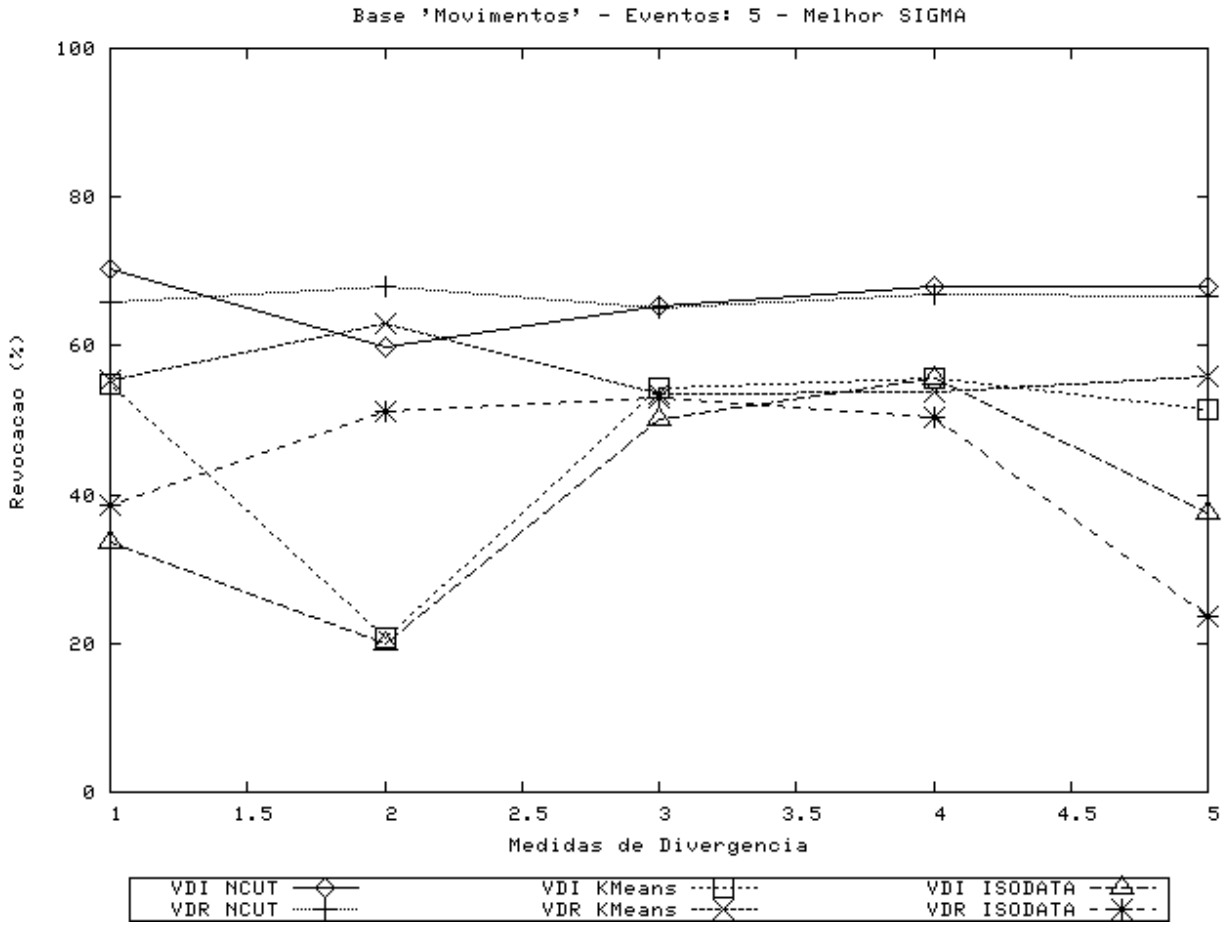


Figura 4.13: Resultados de Revocação para vídeo "Movimentos", 5 classes, melhor caso, $\sigma = 0,40$.

demonstra possuir a melhor taxa de acertos (Revocação) e com elevada taxa de Precisão de forma geral.

Nas Figuras 4.14 e 4.15 fica claro a discrepância do algoritmo "ISODATA" em relação aos demais aglomeradores. Este fato leva a acreditar que o algoritmo "ISODATA" está agrupando o conjunto de características em um número de classes diferente de 5 classes.

Isto pode ser um indício que o conjunto de dados possui eventos de pequena representatividade que não foram rotulados, ou ainda, o conjunto de dados foi rotulado em um número maior de classes do que realmente aparenta possuir.

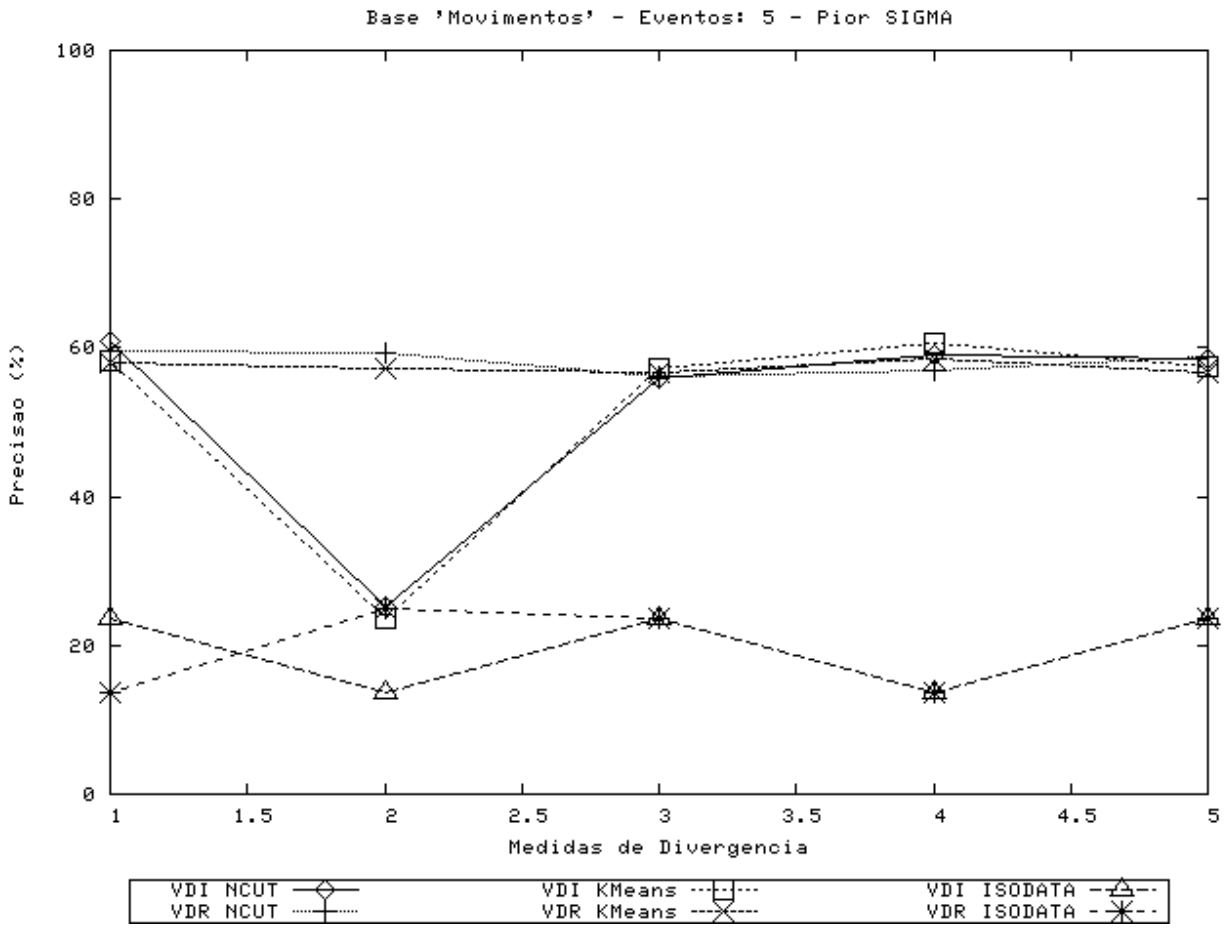


Figura 4.14: Resultados de Precisão para vídeo "Movimentos", 5 classes, pior caso, $\sigma = 0,95$.

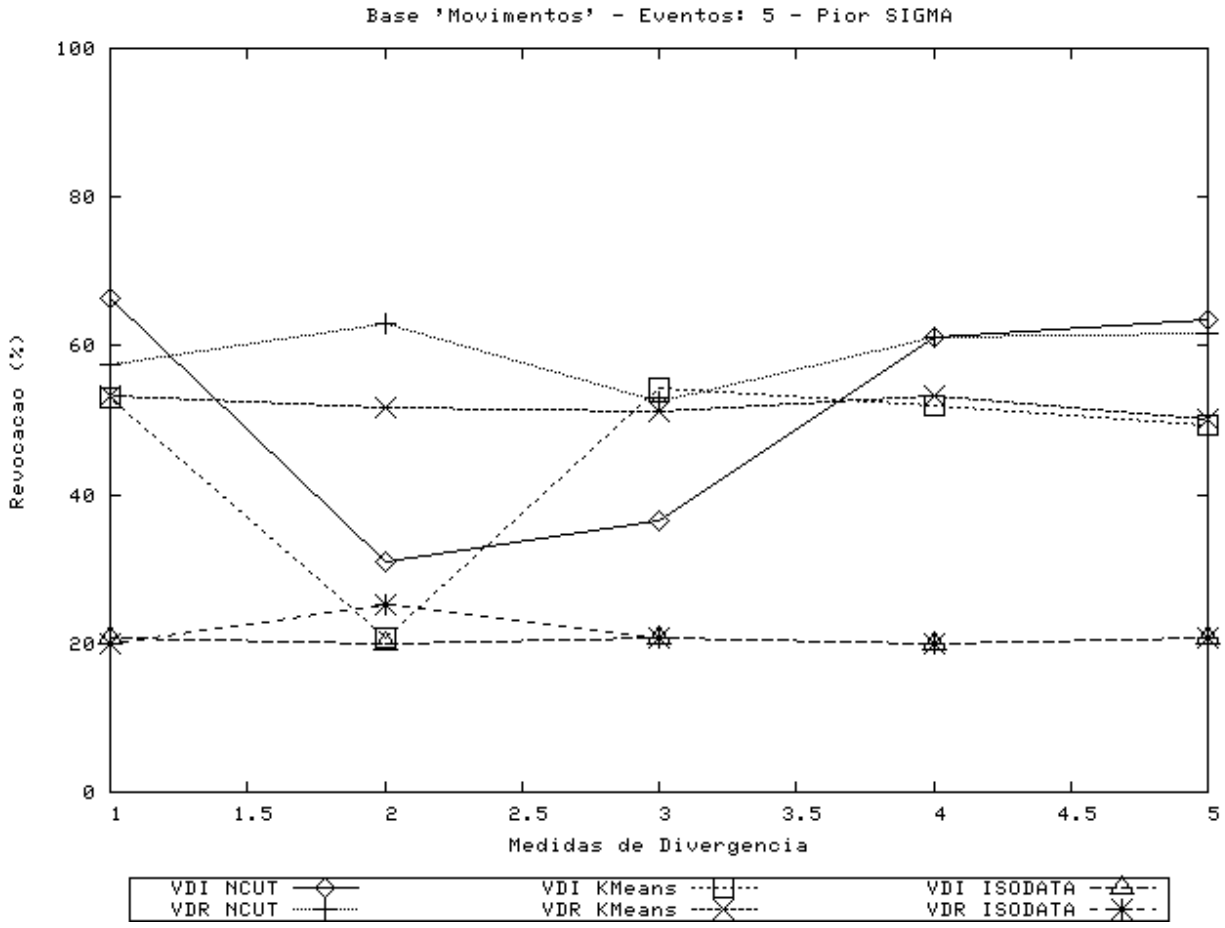


Figura 4.15: Resultados de Revocação para vídeo "Movimentos", 5 classes, pior caso, $\sigma = 0,95$.

4.3.2 Resultados para 2 classes

Para a análise com 2 classes, existe uma tendência natural do algoritmo "ISODATA" ser menos eficaz que os outros métodos de aglomeração pelo provável motivo do algoritmo "ISODATA" sempre retornar um maior número de classes, pois como visto anteriormente, o vídeo "Movimentos" que representa a base de dados em questão foi rotulado, visualmente, com 5 classes, ou seja, existem 5 grupos de padrões reais no vídeo.

Para realizar os testes com 2 classes, foi realizado uma equiparação de eventos similares. O evento "andar" foi definido, agora, como sendo da mesma classe "correr" juntamente com o evento "rolar". Já o evento "acenar" foi rotulado como pertencente à classe "mover no lugar", tem-se dessa forma, uma aproximação de eventos em duas classes.

Quando da análise das Figuras 4.16 e 4.17, têm-se a informação de um comportamento mediano do algoritmo "ISODATA". Porém, este representa o melhor caso, quando da análise das Figuras 4.18 e 4.19 é possível concluir que para um melhor caso o algoritmo "ISODATA" trouxe uma aproximação mediana para 2 classes. Porém, no pior caso, o seu valor de Revocação fica distoante em relação aos demais algoritmos de aglomeração. Fato este causado pela representatividade dos dados ser, efetivamente, maior que duas classes.

Pode-se analisar também, através dos gráficos representados pelas Figuras 4.16 a 4.19 que o algoritmo "Normalized Cut" possui uma taxa de Precisão mediana, porém destaca-se nas medidas de Revocação, exceto pela medida de divergência D_2 que demonstra-se, para este caso, não a melhor medida para se encontrar similaridades entre eventos.

De forma geral, o algoritmo "Normalized Cut" foi o melhor para classificar eventos baseados nas diferentes medidas de divergência, exceto pela medida de divergência D_2 que demonstrou menores taxas de Revocação, para os tipos de eventos representados pelo vídeo "Movimentos".

A questão de se utilizar vetores de deslocamento independes ou resultante demonstrou, para o algoritmo "Normalized Cut", realizar maior influência nas medidas de divergência D_1 e D_2, fato este evidenciado, principalmente, pela Figura 4.13.

Na Tabela 4.2, é dado um resumo geral para o melhor caso dos resultados obtidos através dos experimentos discutidos neste capítulo juntamente com o Apêndice A.1.

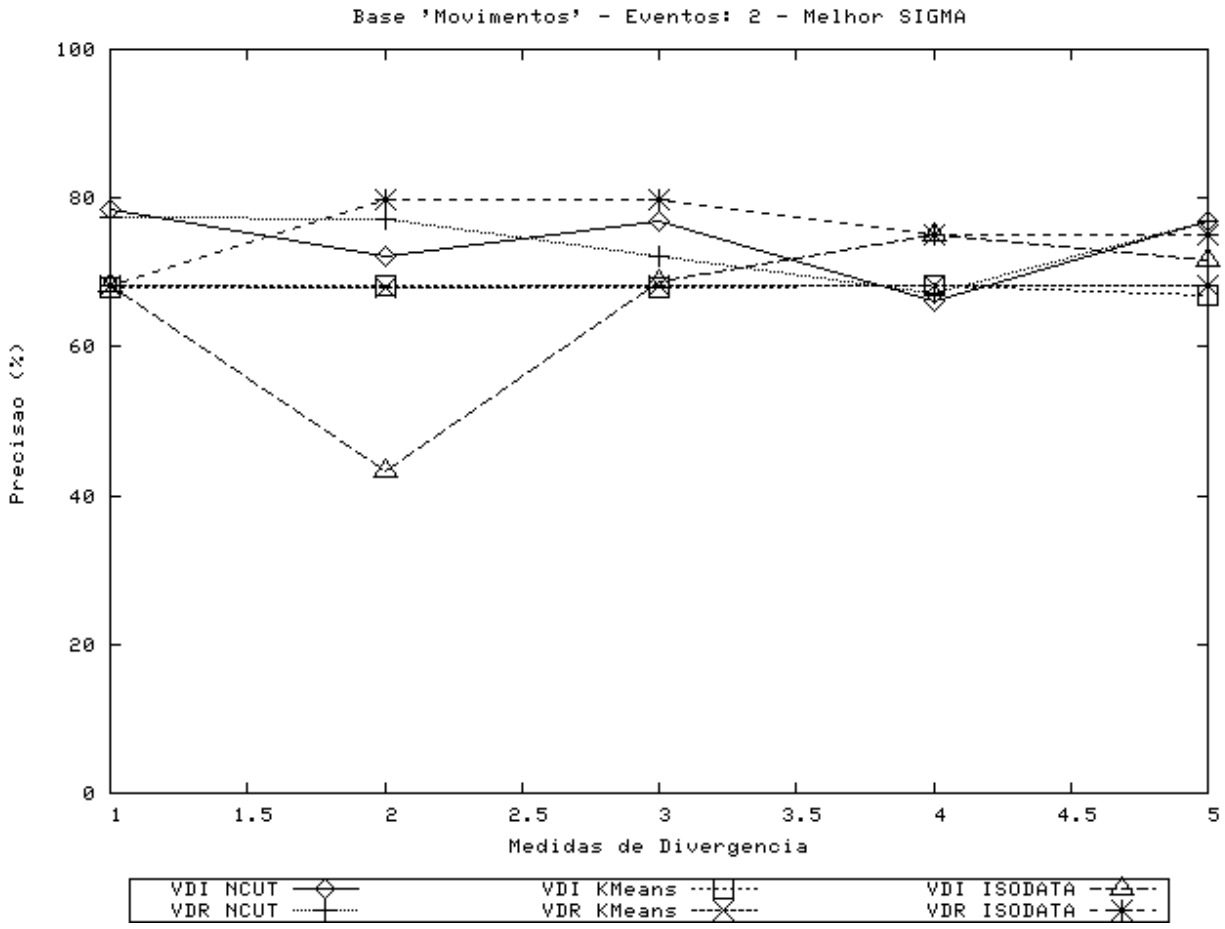


Figura 4.16: Resultados de Precisão para vídeo "Movimentos", 2 classes, melhor caso, $\sigma = 0,55$.

Tabela 4.2: Resultados gerais dos vídeos que compõem a base de dados

Vídeo	"Movimentos"'	"Tênis"	"Inria"
k	5	3	4
Precisão – k classes	64,48%	88,05%	84,91%
Revocação – k classes	70,27%	70,33%	73,18%
Precisão – 2 classes	79,87%	86,46%	86,44%
Revocação – 2 classes	77,84%	70,16%	73,89%

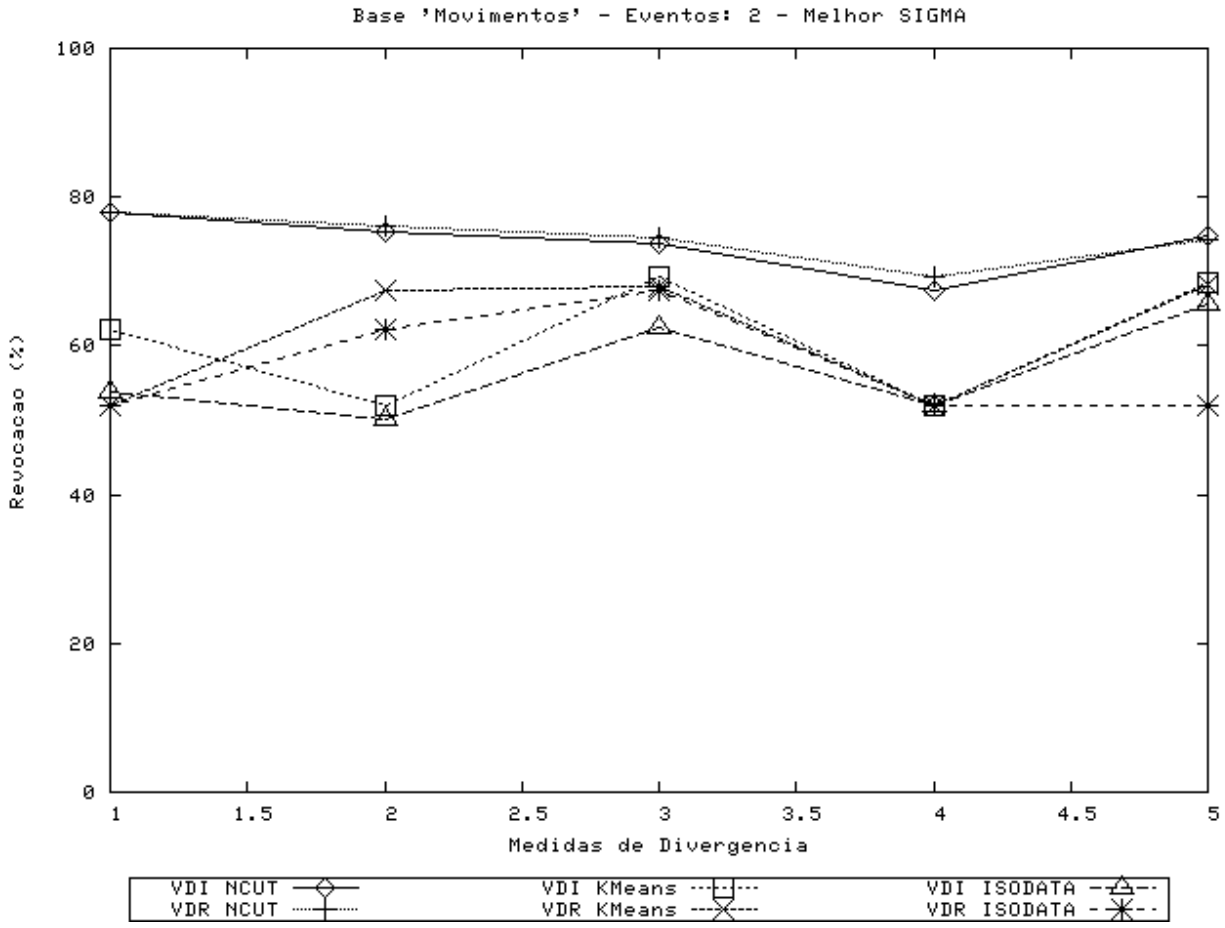


Figura 4.17: Resultados de Revocação para vídeo "Movimentos", 2 classes, melhor caso, $\sigma = 0,55$.

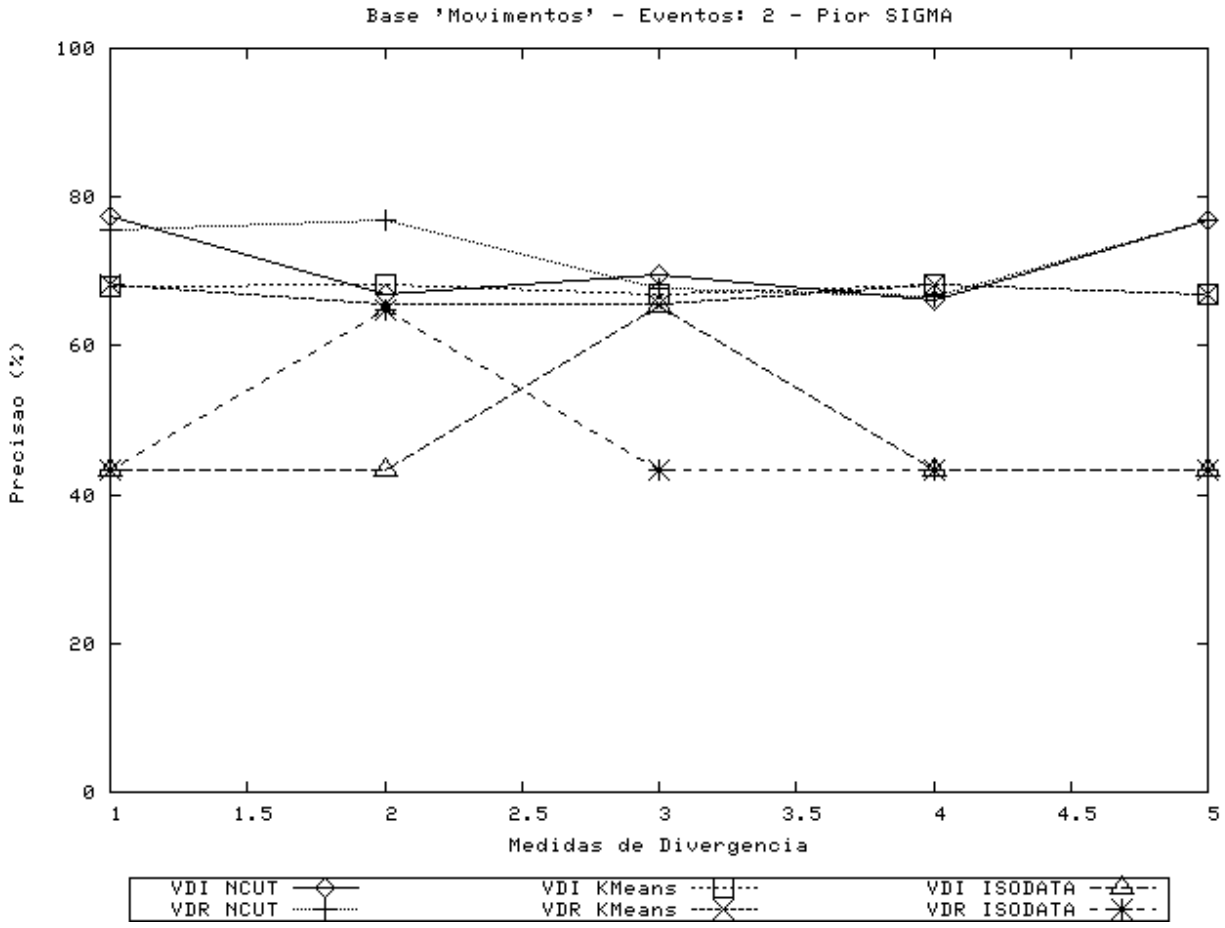


Figura 4.18: Resultados de Precisão para vídeo "Movimentos", 2 classes, pior caso, $\sigma = 0,85$.

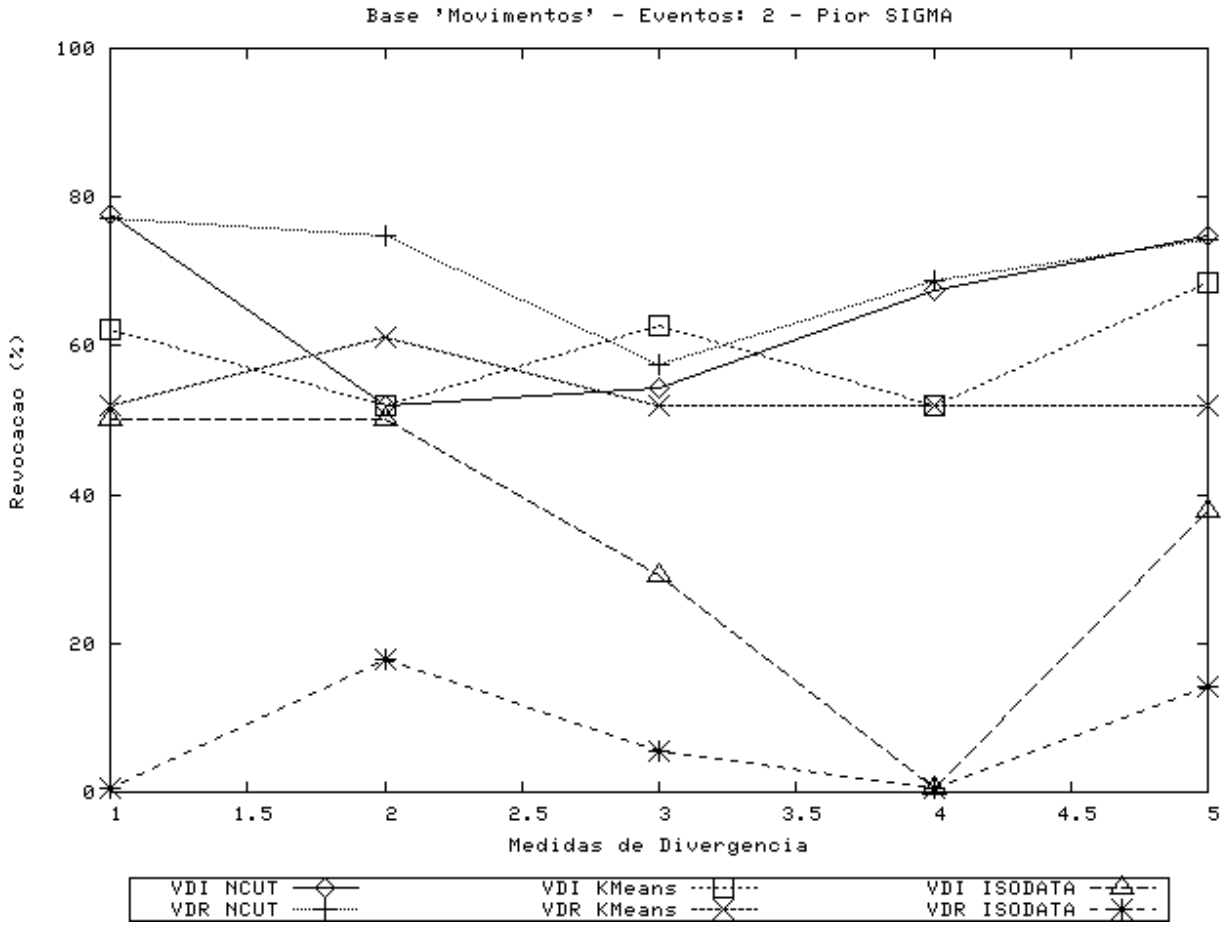


Figura 4.19: Resultados de Revocação para vídeo "Movimentos", 2 classes, pior caso, $\sigma = 0,85$.

No Apêndice A.1 pode-se ter, de forma mais detalhada, outras análises dos resultados dos experimentos realizados nas bases evidenciadas neste capítulo.

Com base em tais informações juntamente com as análises referidas no Apêndice A.1 pode-se realizar algumas conclusões que são o tópico do próximo capítulo.

Capítulo 5

Conclusões e Trabalhos Futuros

Este trabalho descreveu diferentes medidas de divergência e de algoritmos de aglomeração que foram aplicados em um problema de classificação de eventos. As características de deslocamento utilizadas como subsídio para as diferentes medidas de divergência foram extraídas diretamente de vídeos.

Estas características foram arranjadas de duas formas, vetores de deslocamento independe e resultante. Pelos experimentos realizados, a utilização de vetores de deslocamento resultante (VDR) não traz muitos benefícios pelo custo computacional que o processo de cálculo de resultantes demanda.

Um parâmetro, σ foi avaliado em vinte combinações. Observando os experimentos, pode-se dizer que σ é um fator de escala estreitamente relacionado com as informações de eventos específicos, ou seja, é difícil poder pre-estabelecer um valor de σ para um determinado ambiente.

Dentre as medidas de divergência utilizadas, as medidas baseadas nos testes de "Qui-quadrado", "Kolmogorov-Smirnov" e "Kuiper" (as duas últimas baseadas em histogramas acumulados) se destacam como métricas de avaliação de divergências entre eventos.

Analisando os algoritmos de aglomeração, pode-se citar o algoritmo "Normalized Cut" como uma técnica eficiente para o processo de classificação de eventos. Existe uma pequena ressalva, no processo de concepção do grafo (utilizado pelo algoritmo "Normalized Cut") que demanda um esforço computacional extra quando comparado com o algoritmo simplista do "K-Médias".

Como extensão desse estudo, ou trabalhos futuros, pode-se citar uma forma de se inferir sobre o valor de σ de acordo com a distribuição de dados (características) que serão entradas para um algoritmo de aglomeração.

Além disso, uma métrica de avaliação do tamanho da janela temporal, que define

um evento, a fim de se estabelecer janelas de tamanho dinâmico o que pode ser útil e possivelmente fornecer resultados interessantes para ambiente onde ocorrem eventos com uma dinâmica ou variabilidade temporal em maiores ou menores limites.

Conclusões a respeito da eficiência das medidas de divergência e aglomerados utilizando caracteristicas de baixo nível (VD).

Trabalhos futuros: método para ajustar o tamanho da janela temporal de acordo com o ambiente através dos VD.

Referências Bibliográficas

[ANTANI et al., 2002] ANTANI, S., KASTURI, R., and JAIN, R. (2002). A survey on the use of pattern reconition methods for abstraction, indexing and retrieval of images and video. *Patter Recognition*, 35:945–965. Published by Elsevier Science Ltd.

[CHOWDWDHURY and CHELLAPPA, 2003] CHOWDWDHURY, A. K. R. and CHELLAPPA, R. (2003). A factorization approach for activity recognition. *IEEE*.

[COLLINS et al., 1999] COLLINS, R. T., Lipton, A. J., and Kanade, T. (1999). A system for video surveillance and monitoring. Robotics Institute, Carnegie Mellon University.

[FISHER, 2004] FISHER, R. (2003 – 2004). Caviar: Context aware vision using image-based active recognition: Inria (1^{st} set). Internet, Web site. http://homepages.inf.ed.ac.uk/rbf/CAVIAR.

[FORSYTH and PONCE, 2003] FORSYTH, D. and PONCE, J. (2003). *Computer Vision: A Modern Approach*. Prentice Hall.

[Group, 2003] Group, M. P. E. (2003). Moving picture experts group (mpeg) — official site. Internet, Web site. http://www.chiariglione.org/mpeg/index.htm.

[JAIN et al., 1999] JAIN, A. K., MURTY, M. N., and FLYNN, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323.

[JAIN et al., 1995] JAIN, R., Kasturi, R., and Schunck, B. G. (1995). *Machine Vision*, chapter 14. McGraw-Hill.

[KIM and HWANG, 2002] KIM, C. and HWANG, J.-N. (2002). Object-based video abstraction for video surveillance systems. *IEEE Transactions on Circuits and Systems for Video Technology*, 12(12):1128–1138.

[KOPRINSKA and Carrato, 2003] KOPRINSKA, I. and Carrato, S. (2003). Temporal video segmentation: A survey. Technical report, Institute for Information Technologies.

[llas GARGI et al., 2000] llas GARGI, KASTURI, R., and STRAYER, S. H. (2000). Performance characterization of video-shot-change detection methods. *IEEE Trans. Circuits Syst. Video Techn.*, pages 1–13.

[RAMANAN and FORSYTH, 2003] RAMANAN, D. and FORSYTH, D. A. (2003). Finding and tracking people from bottom up. *IEEE*.

[SHI and MALIK, 2000] SHI, J. and MALIK, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*.

[SIKORA, 1997] SIKORA, T. (1997). The mpeg-1 and mpeg-2 digital video coding standards. *IEEE Signal Processing Magazine*.

[WEISS, 1999] WEISS, Y. (1999). Segmentation using eigenvectors: A unifying view. *International Conference on Computer Vision*, pages 975–982.

[ZELNIK-MANOR and IRANI, 2001a] ZELNIK-MANOR, L. and IRANI, M. (2001a). Event-based analysis of video. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[ZELNIK-MANOR and IRANI, 2001b] ZELNIK-MANOR, L. and IRANI, M. (2001b). Human activities. Internet, Web site. http://www.wisdom.weizmann.ac.il/∼vision/VideoAnalysis/Demos/EventDetection/EventDetection.html.

[ZELNIK-MANOR and IRANI, 2001c] ZELNIK-MANOR, L. and IRANI, M. (2001c). Tennis player. Internet, Web site. http://www.wisdom.weizmann.ac.il/∼vision/VideoAnalysis/Demos/EventDetection/EventDetection.html.

[]

Apêndice A

Outros Experimentos

A.1 Experimento com o vídeo "Tênis"

A base de dados representada pelo vídeo "Tênis" possui 3 classes de eventos conforme descrito na Seção 4.2. Os procedimentos para avaliação foram executados primeiro utilizando o processo *ground-truth* com 3 classes e comparando os resultados obtidos do processo de aglomeração com a rotulagem manual.

Os resultados para este vídeo são demonstrados para o melhor e pior valor de σ para a medida de Precisão e a corresponde taxa de Revocação para o mesmo valor de σ.

Na Subseção A.1.1 tem-se a análise desses resultados levando-se em consideração 3 classes e na Subseção A.1.2 a análise referente ao processo *ground-truth* para 2 classes.

A.1.1 Resultados para 3 classes

Observando os gráficos representados pelas Figuras A.1 e A.2, nota-se que o algoritmo "Normalized Cut" obteve a melhor combinação das taxas de Precisão e Revocação para todas as métricas de divergência, exceto a medida de divergência D_4 (Figura A.2).

A utilização de vetores de deslocamento independentes ou resultatante não trouxe um ganho significativo para uma determinada medida de divergência ou algoritmo de aglomeração.

O parâmetro σ demonstra ser altamente correlacionado com a base de dados, ou ainda, com a dinâmica dos eventos tratados. A diferença entre as medidas de σ para o melhor e pior caso quando comparadas com as mesmas medidas dos testes realizados com o vídeo "Movimentos" fornece esta hipótese.

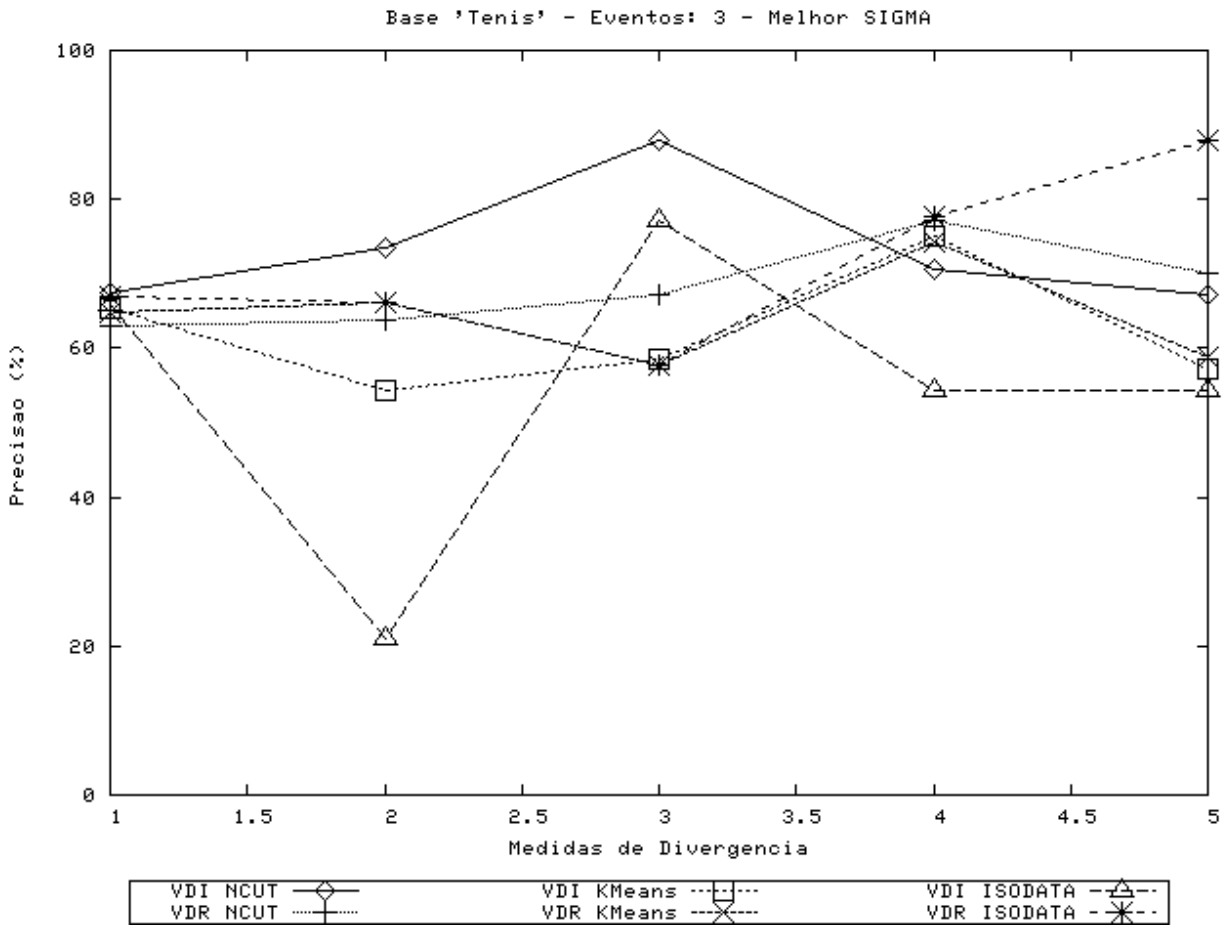


Figura A.1: Resultados de Precisão para vídeo "Tênis", 3 classes, melhor caso, $\sigma = 0,85$.

A.1.2 Resultados para 2 classes

Para a análise com 2 classes do vídeo "Tênis", foram agrupados, na rotulagem manual, os eventos que mais se assemelham, assim, o evento "cadência" foi rotulado como sendo da mesma classe que o evento "batida". O evento "salto" foi o único evento que irá compor a outra classe.

Deve-se relembrar que, existem nos padrões fornecidos aos algoritmos de aglomeração 3 eventos reais, que foram rotulados inicialmente de forma visual, e que agora foram combinados para se obter 2 classes de eventos apenas.

Através da análise dos gráficos representados pelas Figuras A.5 e A.6, nota-se, que o algoritmo de aglormeração "Normalized Cut" se sobressai na relação das medidas de Precisão e Revocação, para a medida de divergência D_2, usando vetores de deslocamento independentes.

Analisando as Figuras A.5 e A.6 pode-se observar que existe uma menor influência por parte da utilização de vetores de deslocamento de forma independente e resultante

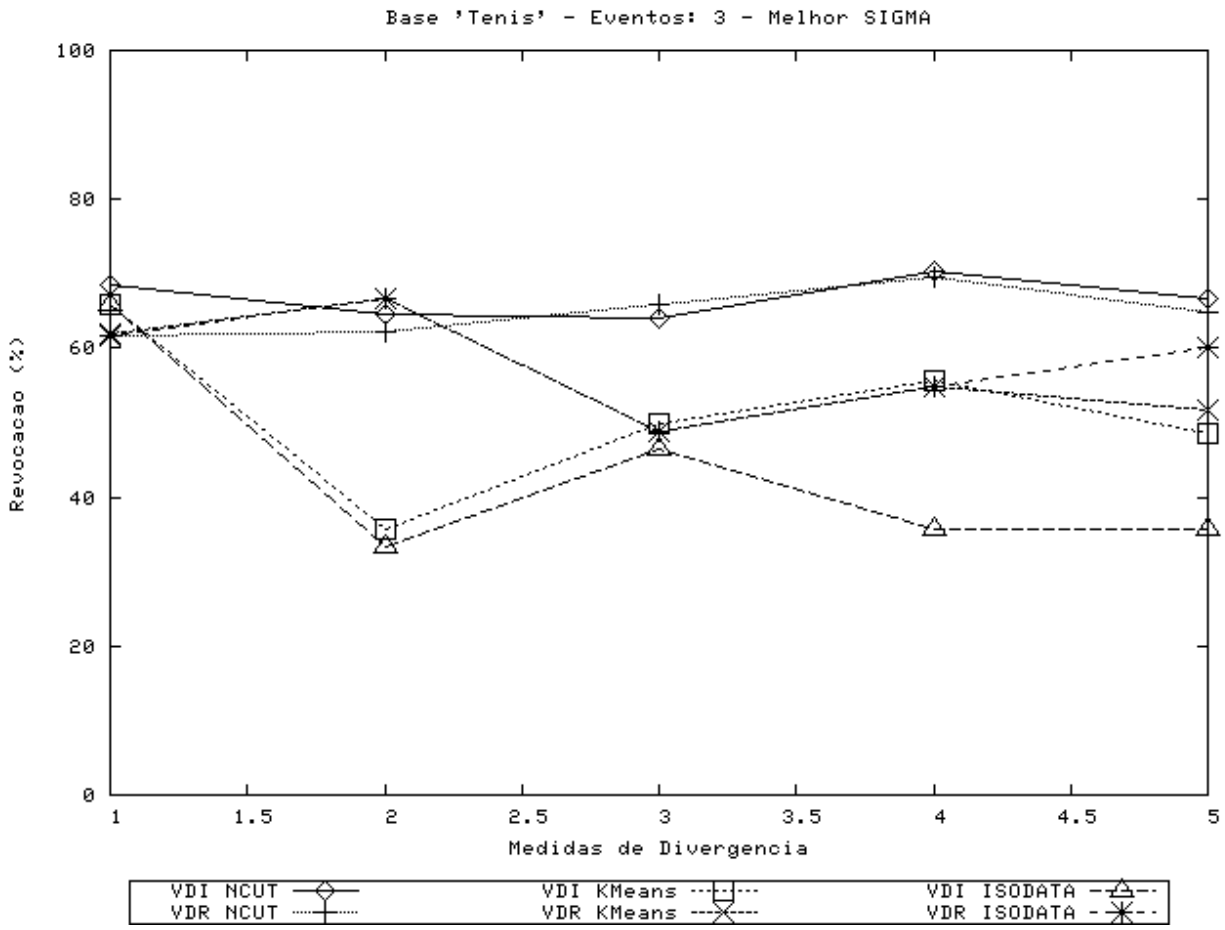


Figura A.2: Resultados de Revocação para vídeo "Tênis", 3 classes, melhor caso, $\sigma = 0,85$.

nas medidas de Precisão do que nas medidas de Revocação. Porém, é muito dependente da medida de distância e do valor de σ utilizado.

No pior caso, representado pelas Figuras A.7 e A.8, o algoritmo "Normalized Cut" mantêm-se dentro de um comportamento mediano enquanto que o algoritmo "ISODATA" sofre maiores perdas com as taxas de Precisão e Revocação.

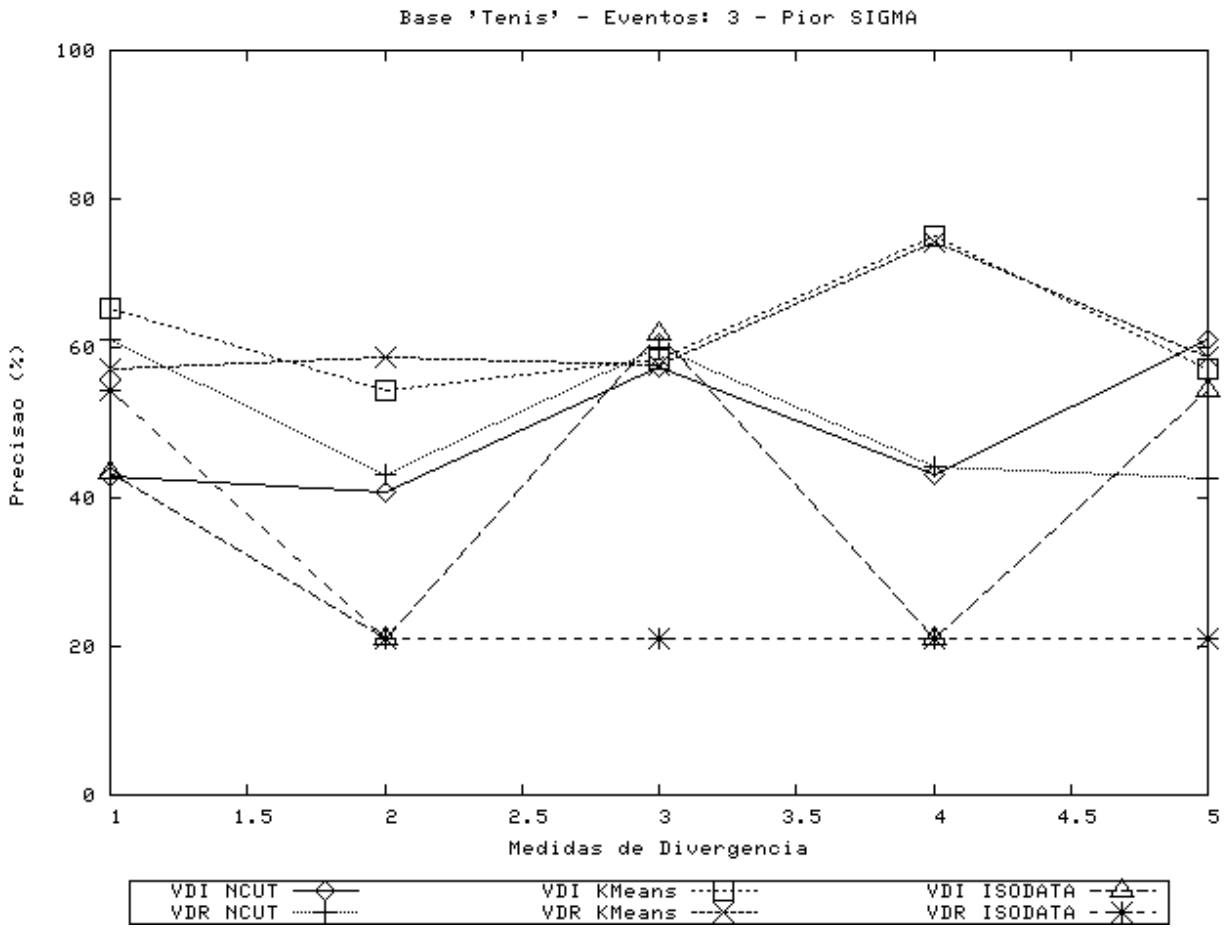


Figura A.3: Resultados de Precisão para vídeo "Tênis", 3 classes, pior caso, $\sigma = 0,10$.

A.2 Experimento com o vídeo "Inria"

A base de dados representada pelo vídeo "Inria" possui 4 classes de eventos conforme descrito na Seção 4.2. Os procedimentos para avaliação foram executados primeiro utilizando o processo *ground-truth* com 4 classes, comparando os resultados obtidos do processo de aglomeração com a rotulagem manual e depois para 2 classes, aproximando classes semelhantes obtidas da rotulagem manual

Os resultados para este vídeo são demonstrados para o melhor e pior valor de σ para a medida de Precisão e a correspondente taxa de Revocação para o mesmo valor de σ.

Na Subseção A.2.1 tem-se a análise desses resultados levando-se em consideração 4 classes e na Subseção A.2.2 a análise referente ao processo *ground-truth* para 2 classes.

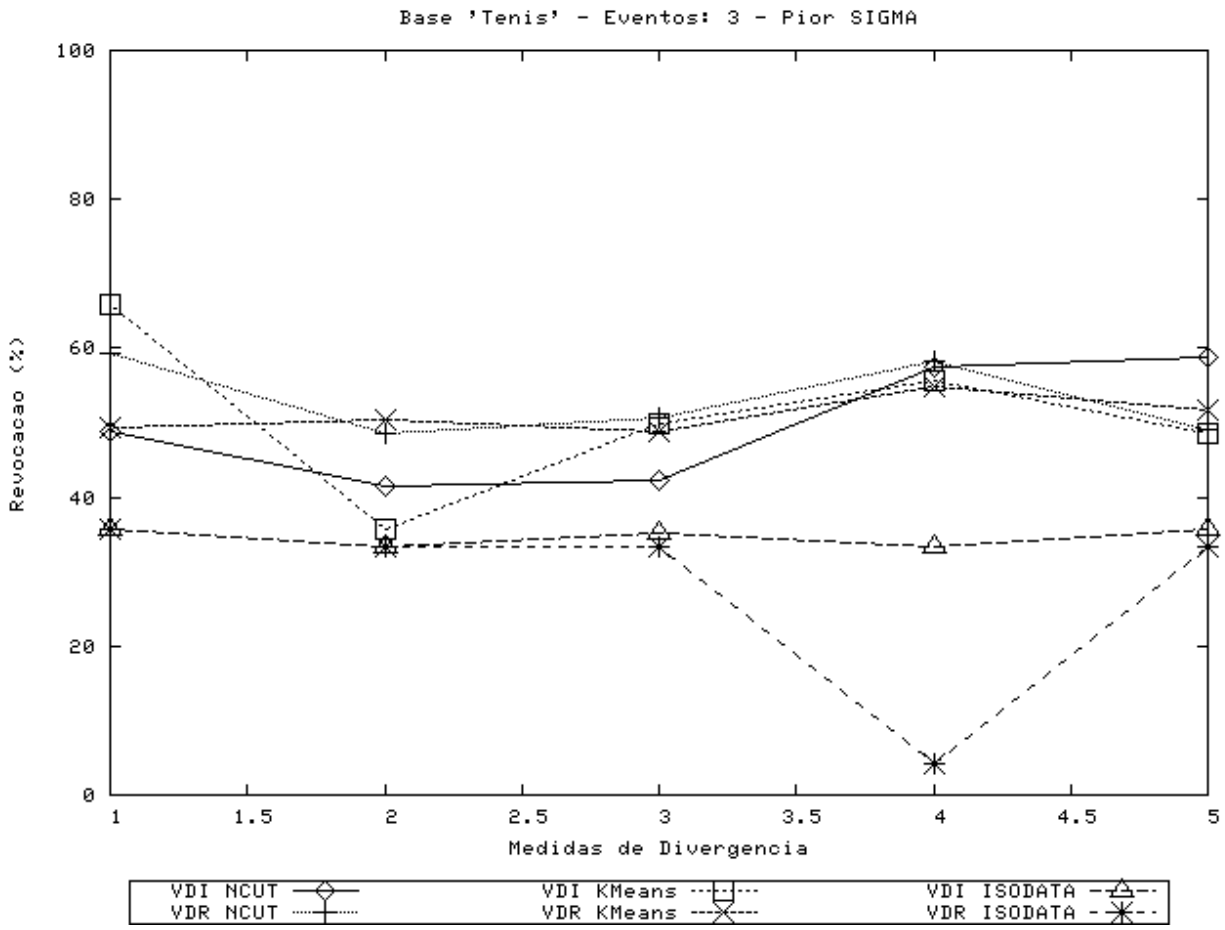


Figura A.4: Resultados de Revocação para vídeo "Tênis", 3 classes, pior caso, $\sigma = 0,10$.

A.2.1 Resultados para 4 classes

Analisando as Figuras A.9 e A.10 o algoritmo "Normalized Cut" demonstra uma taxa de acertos (Revocação) constante e superior aos outros algoritmos de aglomeração em qualquer medida de divergência.

Para a medida de Precisão, o algoritmo "K-Médias" demonstrou obter maiores taxas de Precisão, de forma média, não diferenciando, significativamente, a utilização de vetores de deslocamento de forma independente (VDI) ou de forma a gerar um vetores resultante (VDR).

Para o pior caso, Figuras A.11 e A.12, o algoritmo "K-Médias" possui alta Precisão porém, estas característica não se mantém para a medida de Revocação.

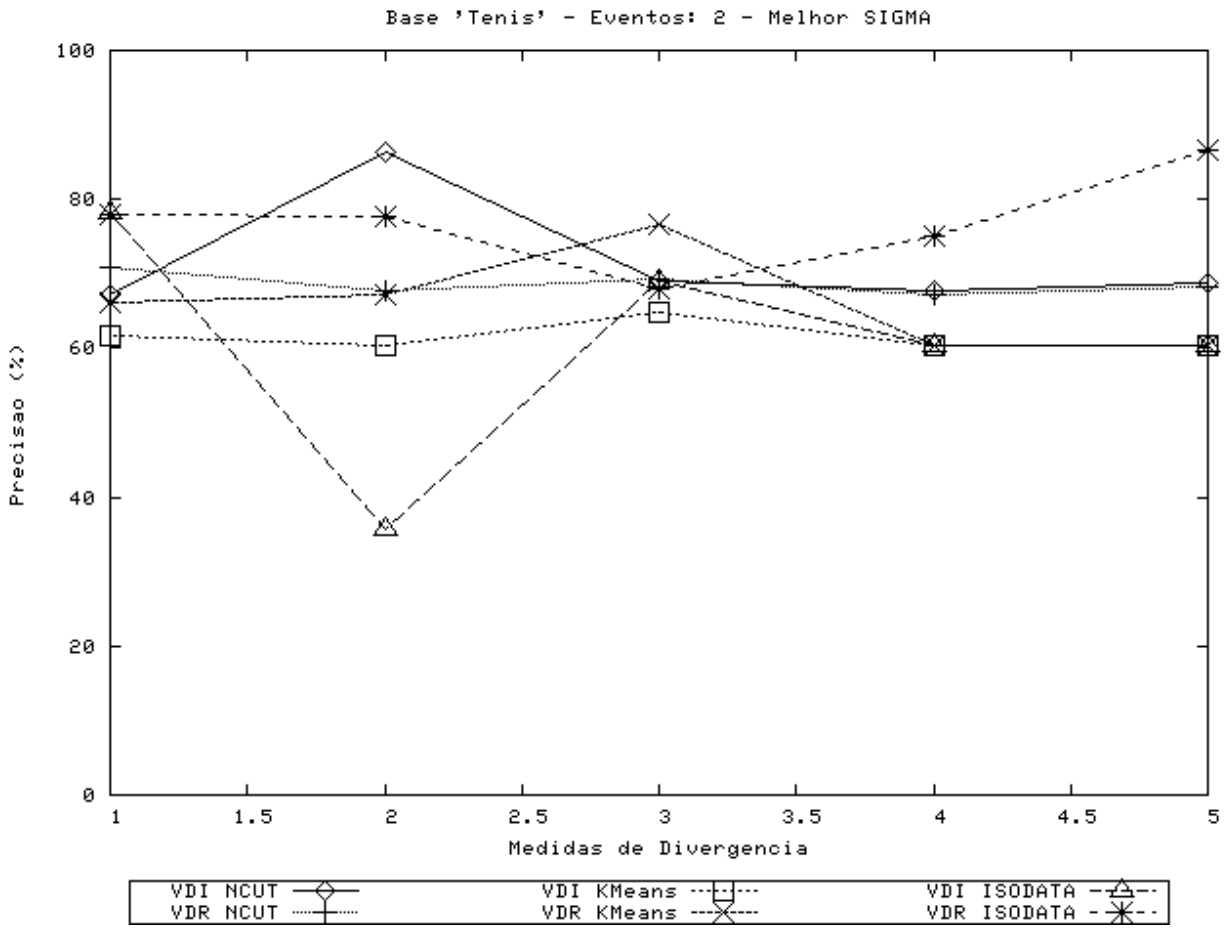


Figura A.5: Resultados de Precisão para vídeo "Tênis", 2 classes, melhor caso, $\sigma = 0,50$.

A.2.2 Resultados para 2 classes

Para a análise com 2 classes do vídeo "Inria", foram agrupados, na rotulagem manual, os eventos "vertical" e "horizontal" em uma mesma classe e em outra classe os eventos "encontro" e "diagonal".

Observando os gráficos representados pelaa Figuras A.13 e A.14 nota-se que houve um ponto ótimo, para a medida de divergência D_2 onde se obteve um alto valor de Precisão e de Revocação para vetores de deslocamento independentes (VDI).

Nas Figuras A.15 e A.16 demonstra que o algoritmo de aglomeração "Normalized Cut" teve melhor taxa de Revocação para as medidas de divergência D_4 e D_5, para VDI e VDR, e para as medidas de divergência D_1 e D_2 para VDR, mas com uma taxa de Precisão inferior ao do algoritmo "K-Médias".

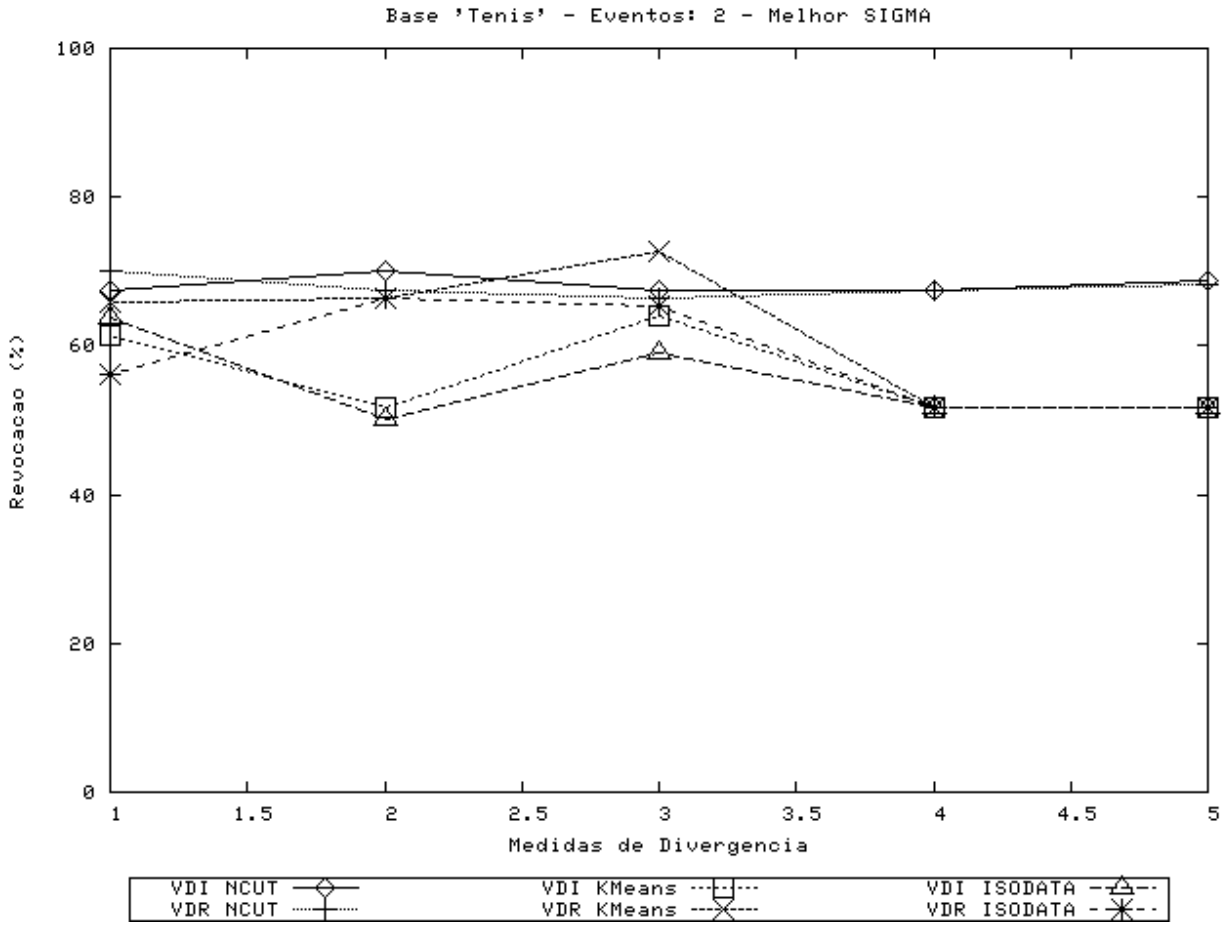


Figura A.6: Resultados de Revocação para vídeo "Tênis", 2 classes, melhor caso, $\sigma = 0,50$.

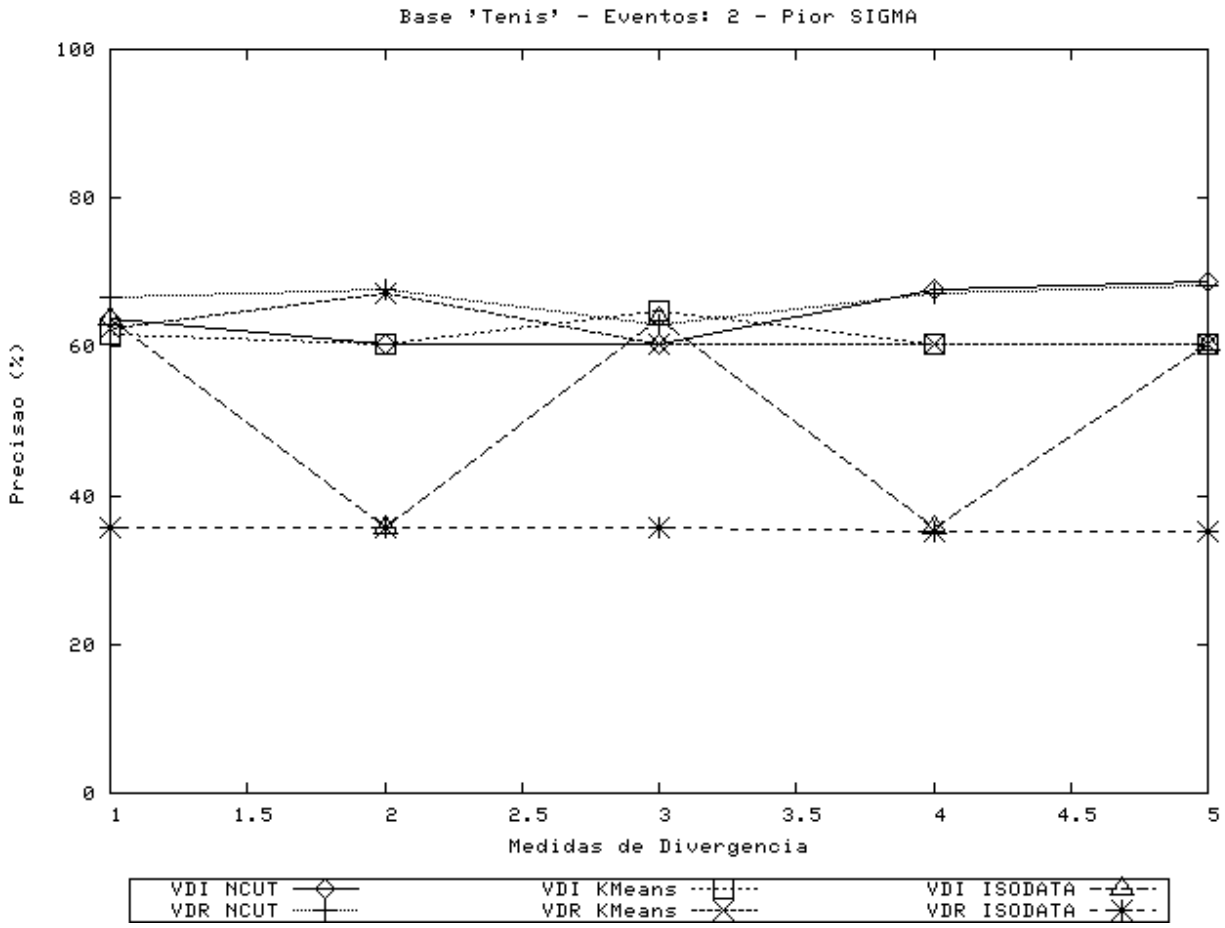


Figura A.7: Resultados de Precisão para vídeo "Tênis", 2 classes, pior caso, $\sigma = 0,60$.

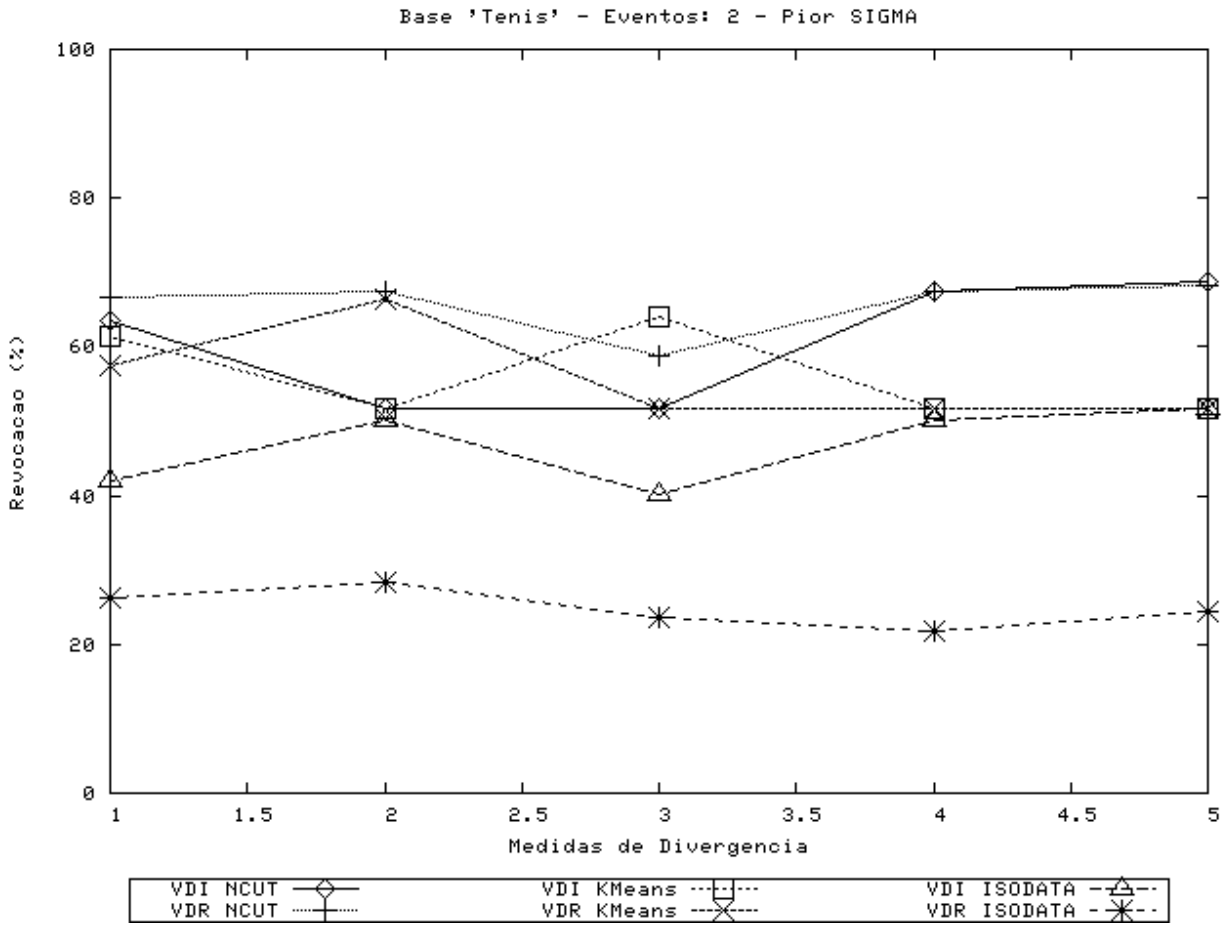


Figura A.8: Resultados de Revocação para vídeo "Tênis", 2 classes, pior caso, $\sigma = 0,60$.

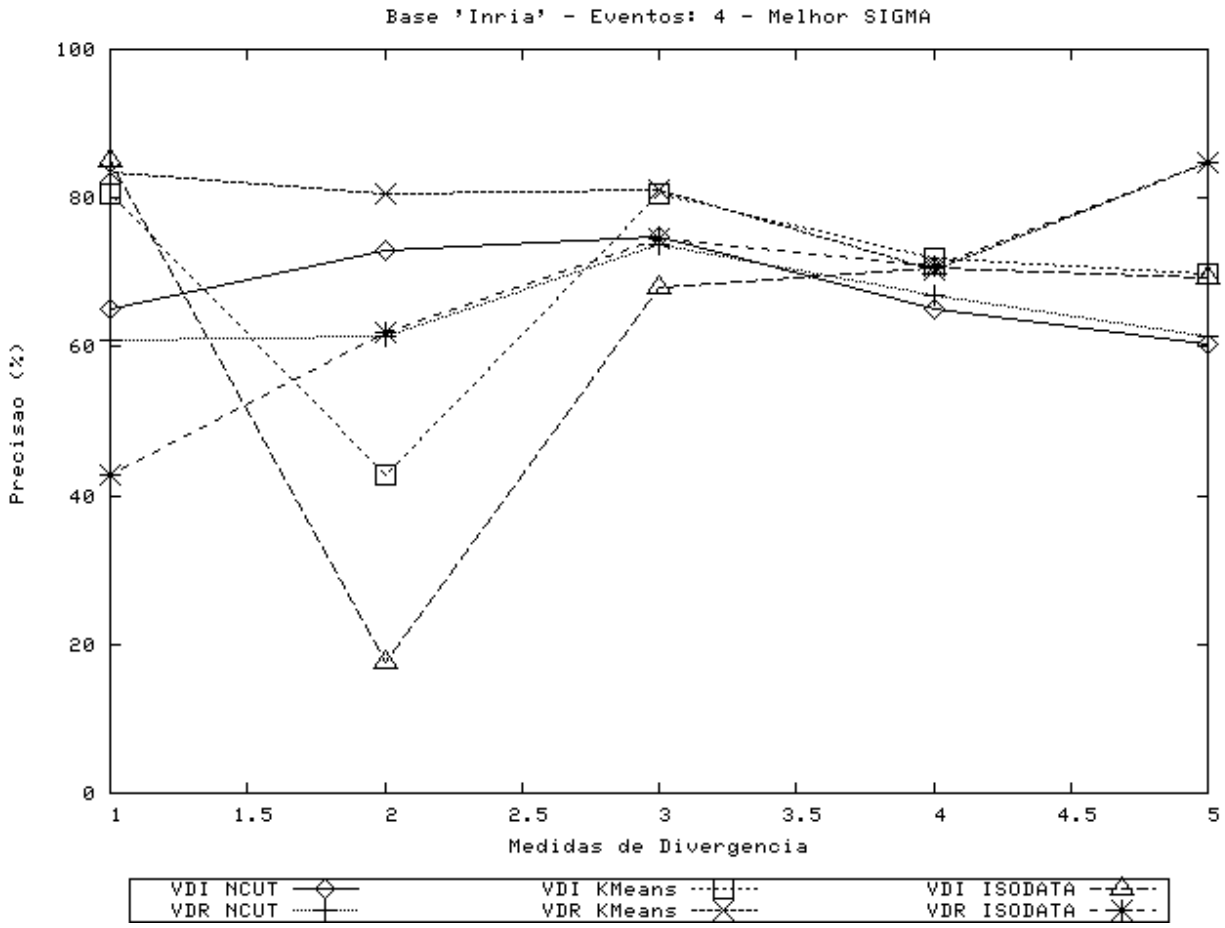


Figura A.9: Resultados de Precisão para vídeo "Inria", 4 classes, melhor caso, $\sigma = 0,85$.

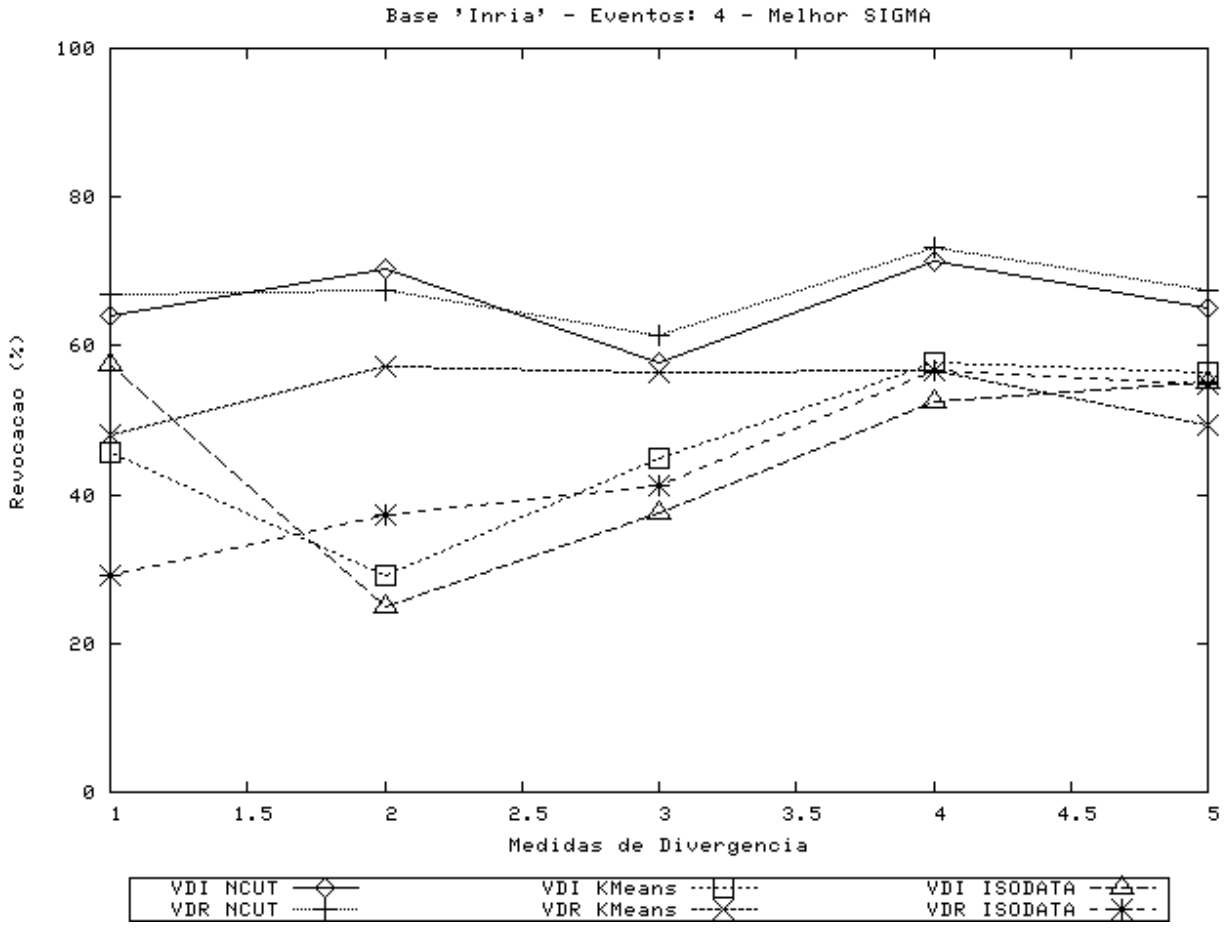


Figura A.10: Resultados de Revocação para vídeo "Inria", 4 classes, melhor caso, $\sigma = 0,85$.

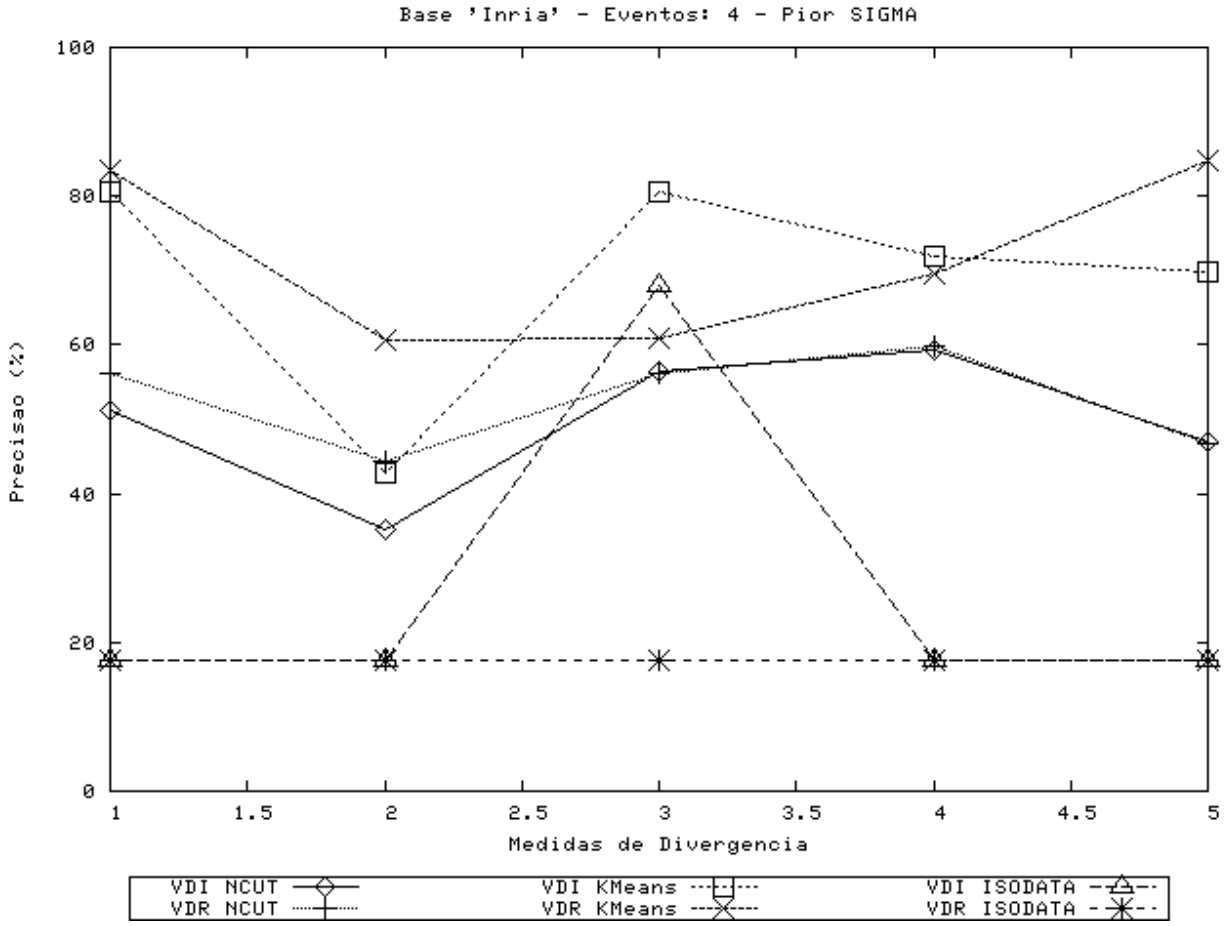


Figura A.11: Resultados de Precisão para vídeo "Inria", 4 classes, pior caso, $\sigma = 0,15$.

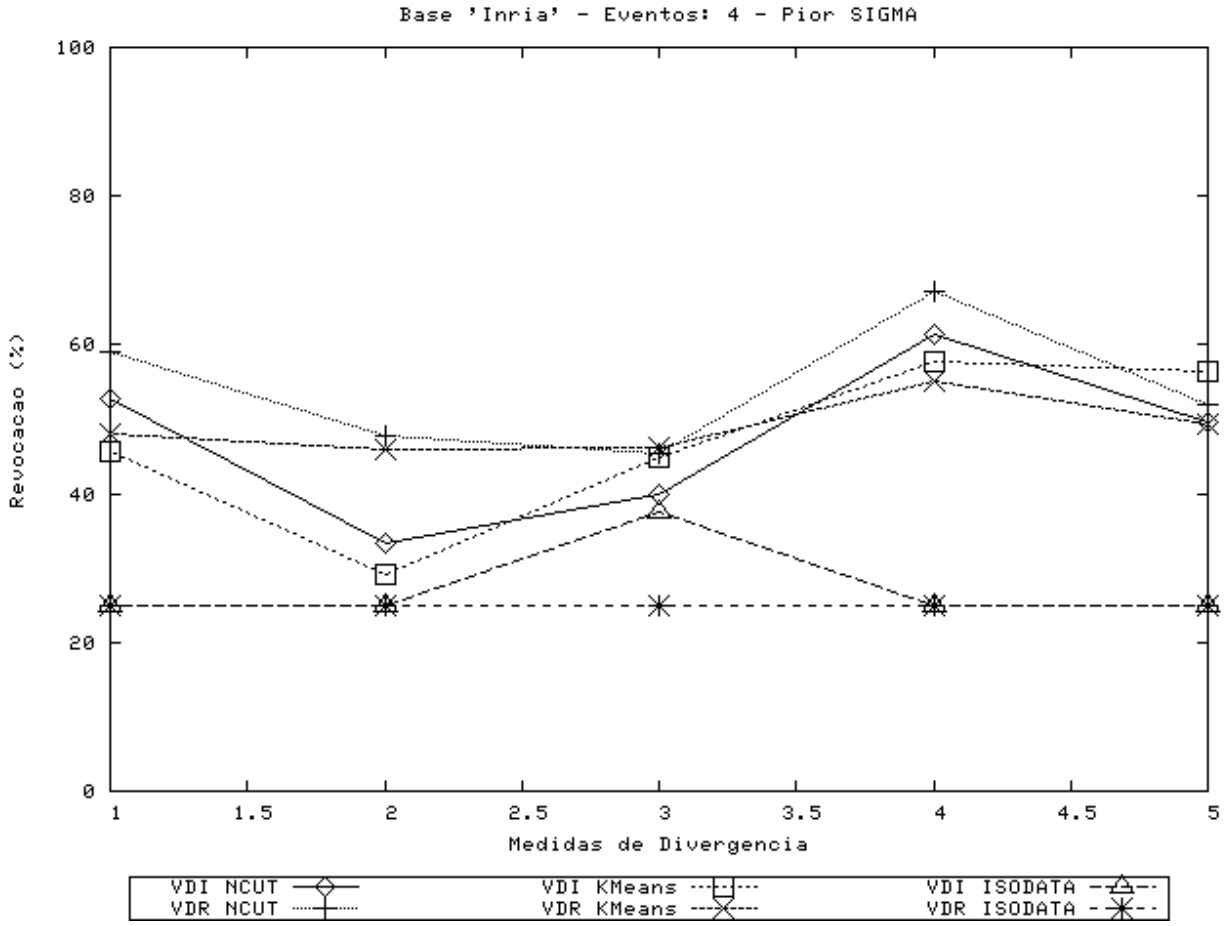


Figura A.12: Resultados de Revocação para vídeo "Inria", 4 classes, pior caso, $\sigma = 0,15$.

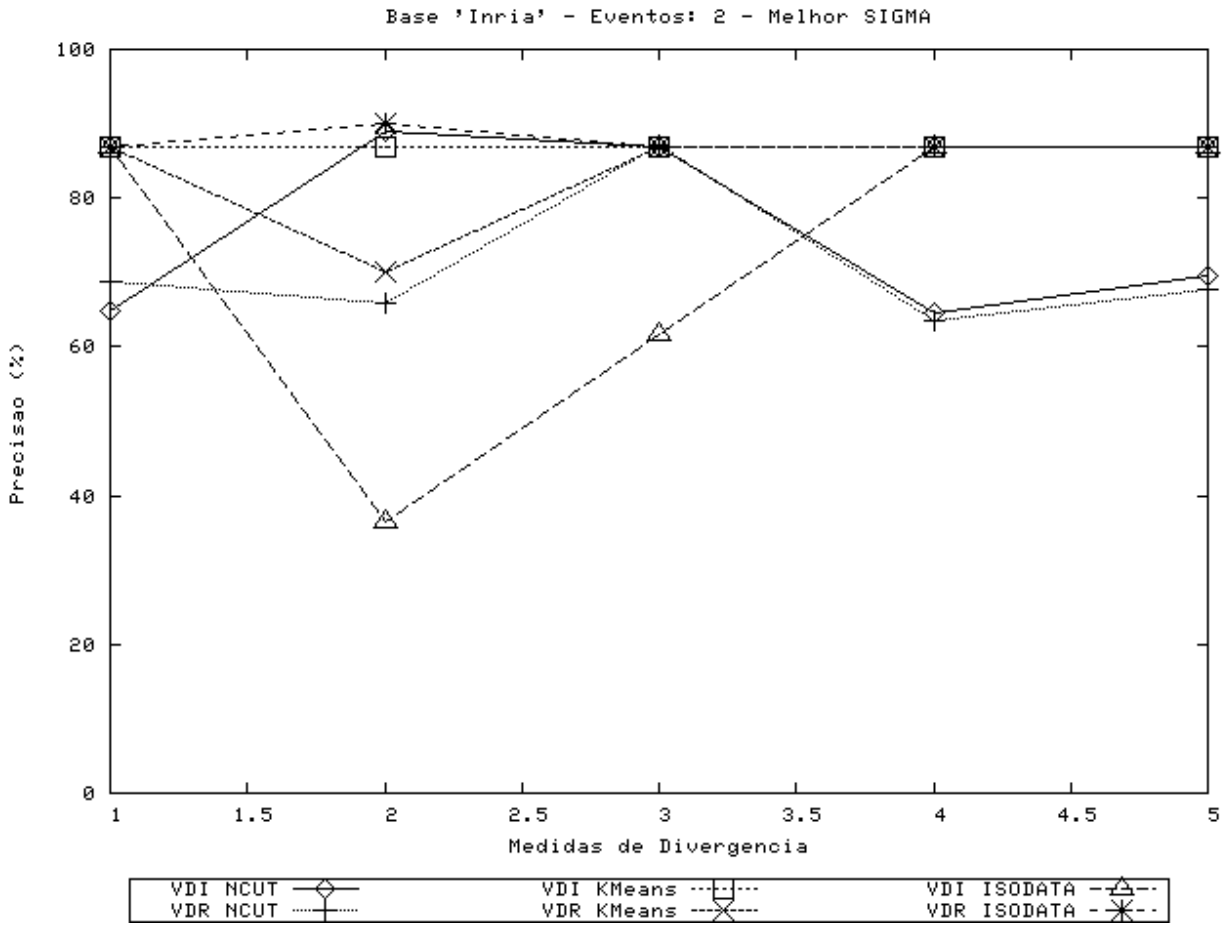


Figura A.13: Resultados de Precisão para vídeo "Inria", 2 classes, melhor caso, $\sigma = 0,05$.

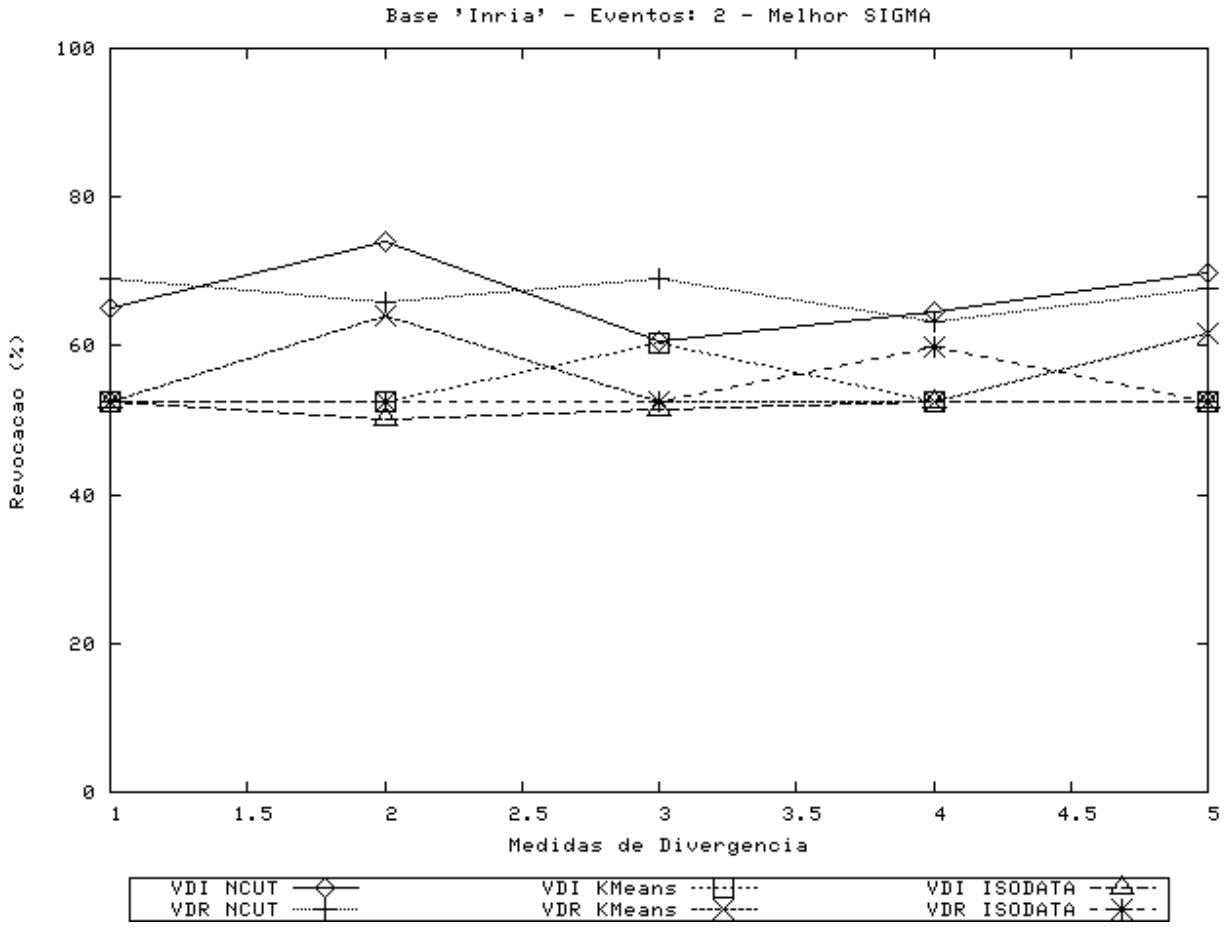


Figura A.14: Resultados de Revocação para vídeo "Inria", 2 classes, melhor caso, $\sigma = 0,05$.

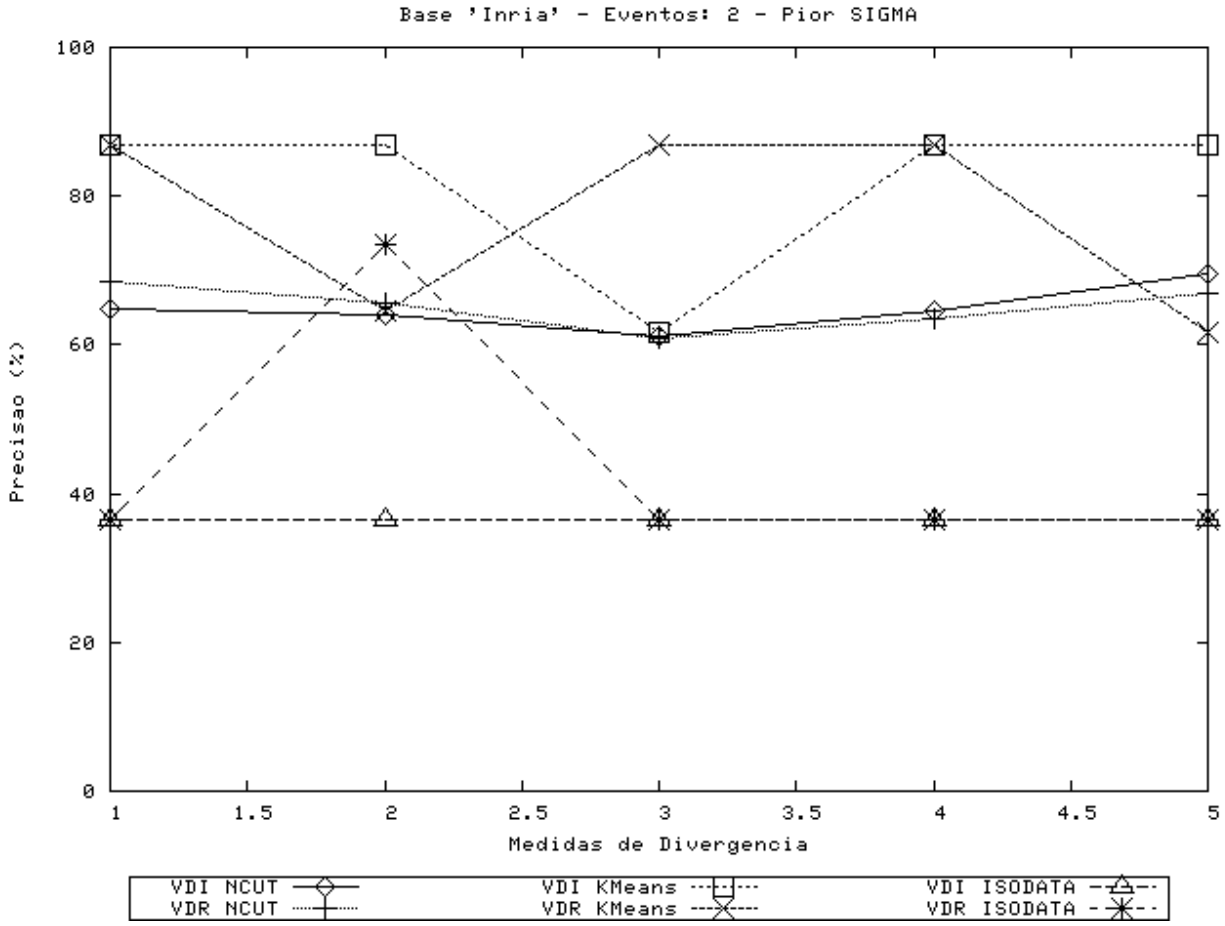


Figura A.15: Resultados de Precisão para vídeo "Inria", 2 classes, pior caso, $\sigma = 0,10$.

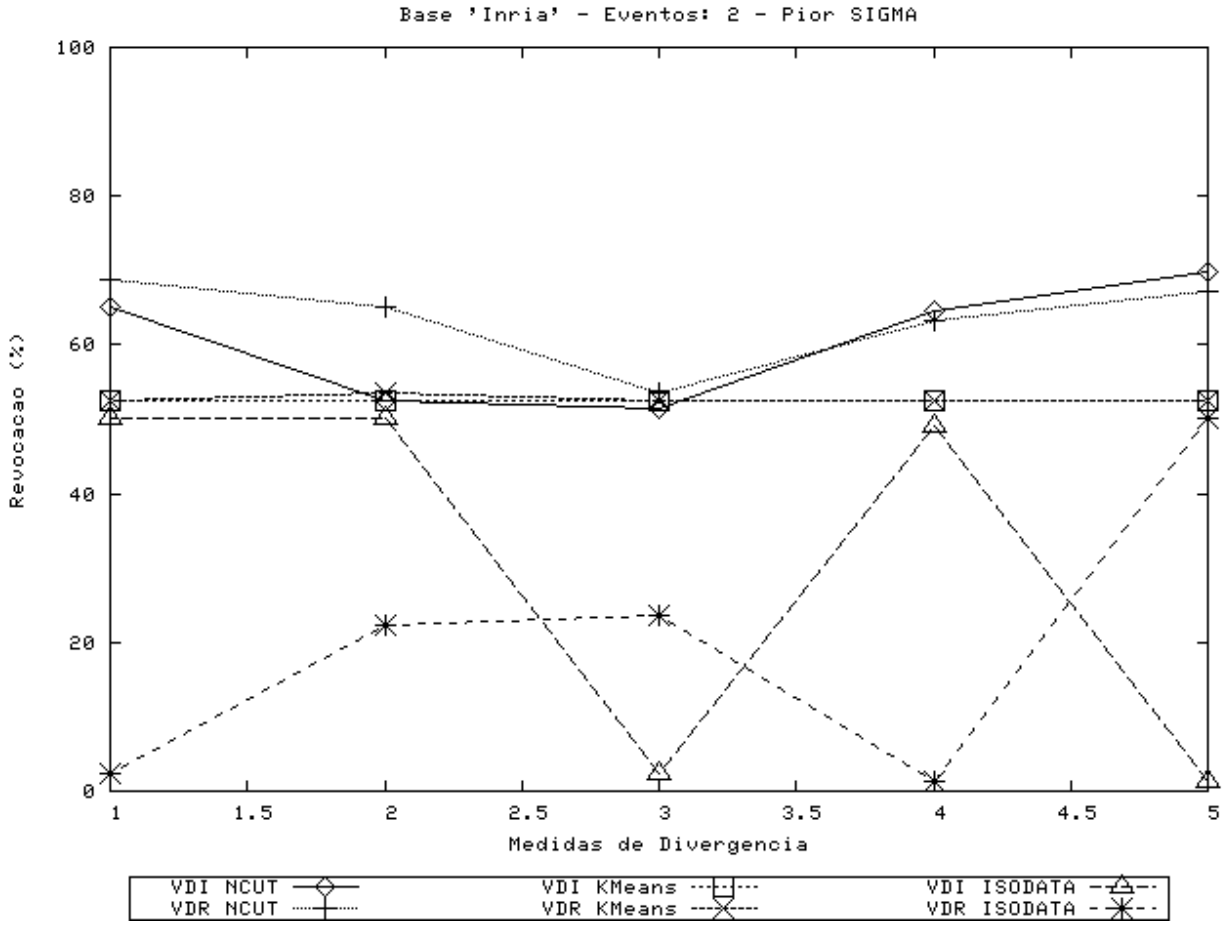


Figura A.16: Resultados de Revocação para vídeo "Inria", 2 classes, pior caso, $\sigma = 0,10$.